Data

Data Science Example

You receive data from a medical researcher concerning a project that you are eager to work on.

Data Science Example

You receive data from a medical researcher concerning a project that you are eager to work on.

The data consists of a 1000 lines table with five columns:

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	165	24.0	0	427.6

...

The aim is to predict the last field given the others.

Data Science Example

You receive data from a medical researcher concerning a project that you are eager to work on.

The data consists of a 1000 lines table with five columns:

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	165	24.0	0	427.6
		...		

The aim is to predict the last field given the others.

The medical researcher does not elaborate further on the data, but they seem to be pretty easy to work with, right?

Data Science Example

You receive data from a medical researcher concerning a project that you are eager to work on.

The data consists of a 1000 lines table with five columns:

012	232	33.5	0	10.7
020	121	16.9	2	210.1
027	165	24.0	0	427.6

...

The aim is to predict the last field given the others.

The medical researcher does not elaborate further on the data, but they seem to be pretty easy to work with, right?

After a few days, you have trained a model that predicts numbers resembling the ones in the table.

You contact the medical researcher and discuss the results.

Model Discussion

Researcher: So, you got the data for all the patients?

Model Discussion

Researcher: So, you got the data for all the patients?
Data Miner: Yes. I haven't had much time for analysis, but I do have a few interesting results.

Model Discussion

Researcher: So, you got the data for all the patients?
Data Miner: Yes. I haven't had much time for analysis, but I do have a few interesting results.
Researcher: Amazing. There were so many data issues with this set of patients that I couldn't do much.

Model Discussion

Researcher: So, you got the data for all the patients?
Data Miner: Yes. I haven't had much time for analysis, but I do have a few interesting results.
Researcher: Amazing. There were so many data issues with this set of patients that I couldn't do much.
Data Miner: Oh? I didn't hear about any possible problems.

Model Discussion

Researcher: So, you got the data for all the patients?
Data Miner: Yes. I haven't had much time for analysis, but I do have a few interesting results.
Researcher: Amazing. There were so many data issues with this set of patients that I couldn't do much.
Data Miner: Oh? I didn't hear about any possible problems.
Researcher: Well, first, there is field 5, the variable we want to predict. It's common knowledge among people who analyze this type of data that results are better if you work with the log of the values, but I didn't discover this until later. Was it mentioned to you?

Model Discussion

Researcher: So, you got the data for all the patients?
Data Miner: Yes. I haven't had much time for analysis, but I do have a few interesting results.
Researcher: Amazing. There were so many data issues with this set of patients that I couldn't do much.
Data Miner: Oh? I didn't hear about any possible problems.
Researcher: Well, first, there is field 5, the variable we want to predict. It's common knowledge among people who analyze this type of data that results are better if you work with the log of the values, but I didn't discover this until later. Was it mentioned to you?
Data Miner: No.

Model Dicsuccion

Researcher: But surely you heard about what happened to field 4? It's supposed to be measured on a scale from 1 to 10, with 0 indicating a missing value, but because of a data entry error, all 10's were changed into 0's. Unfortunately, since some of the patients have missing values for this field, it's impossible to say whether a 0 in this field is a real 0 or a 10. Quite a few of the records have that problem.

Model Dicsuccion

Researcher: But surely you heard about what happened to field 4? It's supposed to be measured on a scale from 1 to 10, with 0 indicating a missing value, but because of a data entry error, all 10's were changed into 0's. Unfortunately, since some of the patients have missing values for this field, it's impossible to say whether a 0 in this field is a real 0 or a 10. Quite a few of the records have that problem.
Data Miner: Interesting. Were there any other problems?

Model Dicsuccion

Researcher: But surely you heard about what happened to field 4? It's supposed to be measured on a scale from 1 to 10, with 0 indicating a missing value, but because of a data entry error, all 10's were changed into 0's. Unfortunately, since some of the patients have missing values for this field, it's impossible to say whether a 0 in this field is a real 0 or a 10. Quite a few of the records have that problem.
Data Miner: Interesting. Were there any other problems?
Researcher: Yes, fields 2 and 3 are basically the same, but I assume that you probably noticed that.

Model Dicsuccion

Researcher: But surely you heard about what happened to field 4? It's supposed to be measured on a scale from 1 to 10, with 0 indicating a missing value, but because of a data entry error, all 10's were changed into 0's. Unfortunately, since some of the patients have missing values for this field, it's impossible to say whether a 0 in this field is a real 0 or a 10. Quite a few of the records have that problem.
Data Miner: Interesting. Were there any other problems?
Researcher: Yes, fields 2 and 3 are basically the same, but I assume that you probably noticed that.
Data Miner: Yes, but these fields were only weak predictors of field 5.

Model Discussion

Researcher: Anyway, given all those problems, I'm surprised you were able to accomplish anything.
Data Miner: True, but my results are really quite good. Field 1 is a very strong predictor of field 5. I'm surprised that this wasn't noticed before.
Researcher: What? Field 1 is just an identification number.
Data Miner: Nonetheless, my results speak for themselves.
Researcher: Oh, no! I just remembered. We assigned ID numbers after we sorted the records based on field 5. There is a strong connection, but it isn't very sensible. Sorry.

Model Discussion

Researcher: Anyway, given all those problems, I'm surprised you were able to accomplish anything.
Data Miner: True, but my results are really quite good. Field 1 is a very strong predictor of field 5. I'm surprised that this wasn't noticed before.
Researcher: What? Field 1 is just an identification number.
Data Miner: Nonetheless, my results speak for themselves.
Researcher: Oh, no! I just remembered. We assigned ID numbers after we sorted the records based on field 5. There is a strong connection, but it isn't very sensible. Sorry.

OK, what's the point?

You have to

Understand the task you want to solve and the data!

Data Objects

Data objects represent entities we work with (e.g., classify them).

For example, in cancer prediction, the data objects are patients. In fruit classification, the data objects are individual fruits.

Data Objects

Data objects represent entities we work with (e.g., classify them).

For example, in cancer prediction, the data objects are patients. In fruit classification, the data objects are individual fruits.

Data objects are described by *attributes* (or *features* or *variables*).

For example, the age, weight, genetic profile, and other patient characteristics. Or the width and height of a fruit.

Attributes vs Features vs Variables

The name differs from field to field.

Attributes vs Features vs Variables

The name differs from field to field.

So, the following names are usually used as synonyms:

- *Attributes* - used mostly by database and data mining experts.
- *Features* - used mostly by machine learning experts.
- *Variables* - used mostly by statisticians.

Attributes vs Features vs Variables

The name differs from field to field.

So, the following names are usually used as synonyms:

- *Attributes* - used mostly by database and data mining experts.
- *Features* - used mostly by machine learning experts.
- *Variables* - used mostly by statisticians.

One may make some distinctions

- *Attributes* represent information about the object without any additional assumptions.
- *Features* assume that their values are somewhat characteristic of the object.
- *Variables* assume that there is some process behind them (typically a random process in the case of statistics).

Data Types - Categorical Attributes

Categorical attributes (nominal attributes) are symbols or names of things.

- Each value represents some kind of category, code, or state.
- Values are not ordered and should not be used quantitatively (in computer science, the values are known as enumerations).

Data Types - Categorical Attributes

Categorical attributes (nominal attributes) are symbols or names of things.

- Each value represents some kind of category, code, or state.
- Values are not ordered and should not be used quantitatively (in computer science, the values are known as enumerations).
- **Examples:**

$$\text{hair\_color} \in \{\text{black}, \text{brown}, \text{blond}, \text{red}, \text{auburn}, \text{gray}, \text{white}\}$$

$$\text{marital\_status} \in \{\text{single}, \text{married}, \text{divorced}, \text{widowed}\}$$

$$\text{customer\_ID} \in \{0, 1, 2, \ldots\}$$

Even though the last one is usually expressed using numbers, it should not be used quantitatively.

Data Types - Categorical Attributes

Categorical attributes (nominal attributes) are symbols or names of things.

- Each value represents some kind of category, code, or state.
- Values are not ordered and should not be used quantitatively (in computer science, the values are known as enumerations).
- **Examples:**

$$\text{hair\_color} \in \{\text{black}, \text{brown}, \text{blond}, \text{red}, \text{auburn}, \text{gray}, \text{white}\}$$

$$\text{marital\_status} \in \{\text{single}, \text{married}, \text{divorced}, \text{widowed}\}$$

$$\text{customer\_ID} \in \{0, 1, 2, \ldots\}$$

Even though the last one is usually expressed using numbers, it should not be used quantitatively.

Binary attributes are categorical attributes with only two values.

DataTypes - Ordinal Attributes

Ordinal attribute is an attribute with values that have a meaningful order or ranking among them.

DataTypes - Ordinal Attributes

Ordinal attribute is an attribute with values that have a meaningful order or ranking among them.

Examples:

$$\text{drink\_size} \in \{\text{small}, \text{medium}, \text{large}\}$$

$$\text{grades} \in \{\text{A}, \text{B}, \text{C}, \text{D}, \text{E}, \text{F}\}$$

It can also be obtained by discretizing numeric quantities into series of intervals.

Ordinal attributes do not allow arithmetic operations.

DataTypes - Ordinal Attributes

Ordinal attribute is an attribute with values that have a meaningful order or ranking among them.

Examples:

$$\text{drink\_size} \in \{\text{small}, \text{medium}, \text{large}\}$$

$$\text{grades} \in \{\text{A}, \text{B}, \text{C}, \text{D}, \text{E}, \text{F}\}$$

It can also be obtained by discretizing numeric quantities into series of intervals.

Ordinal attributes do not allow arithmetic operations.

Categorical and ordinal attributes are called *qualitative* attributes.

Next, we look at numeric, i.e., *quantitative* attributes.

Data Types - Numeric Attributes

Numeric attributes are quantities represented by numbers.

Data Types - Numeric Attributes

Numeric attributes are quantities represented by numbers.

Distinguish two types: *Interval-scale* and *ratio-scale*.

	INTERVAL SCALE	**RATIO SCALE**
Measurement interval	Equal intervals between consecutive points.	Equal intervals with the presence of a true zero.
Absolute zero	Lacks a true zero point.	Possesses a true zero point.
Statistical analysis	Limited to addition and subtraction	Allows for meaningful multiplication and division.
Meaningful ratios	Ratios are not meaningful due to the lack of zero.	Ratios are meaningful due to the presence of zero.
Examples	IQ scores, Celsius temperature, NPS data, etc.	Height, weight, income, etc.

Discrete vs Continuous Attributes

Often, two kinds of numeric attributes are distinguished:

Discrete vs Continuous Attributes

Often, two kinds of numeric attributes are distinguished:

- *Discrete*

  A finite or countably infinite range of values, i.e., integers may represent the values.

  Some (but not all) authors count the qualitative (categorical, ordinal) attributes among the discrete attributes.

Discrete vs Continuous Attributes

Often, two kinds of numeric attributes are distinguished:

- *Discrete*

  A finite or countably infinite range of values, i.e., integers may represent the values.

  Some (but not all) authors count the qualitative (categorical, ordinal) attributes among the discrete attributes.

- *Continuous*

  An uncountably infinite range of values, typically an interval.

  There are several more or less formal definitions of continuous attributes in the literature. For example:

  - All non-discrete variables.
  - Have an infinite number of values between any two values.
  - Their values are measured (??).

Deeper characteristics of data (statistical properties, etc.) will be examined at tutorials.

Classifier Evaluation

Classifier

Assume binary classification into two classes $\{0, 1\}$.

Classifier

Assume binary classification into two classes $\{0, 1\}$.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in \{0, 1\}$ for all k.

Classifier

Assume binary classification into two classes $\{0, 1\}$.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \dots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in \{0, 1\}$ for all k.

Consider a sequence of predictions generated by a classifier:

$$h_1, \dots, h_p \in \{0, 1\}$$

Here each h_k has been predicted for the k-the example $(\vec{x}_k, c_k)$.

Classifier

Assume binary classification into two classes $\{0, 1\}$.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in \{0, 1\}$ for all k.

Consider a sequence of predictions generated by a classifier:

$$h_1, \ldots, h_p \in \{0, 1\}$$

Here each h_k has been predicted for the k-the example $(\vec{x}_k, c_k)$.

How good are the predictions $h_1, \ldots, h_p$ w.r.t. $c_1, \ldots, c_p$?

There are many possible metrics ...

Classifier

Assume binary classification into two classes $\{0, 1\}$.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in \{0, 1\}$ for all k.

Consider a sequence of predictions generated by a classifier:

$$h_1, \ldots, h_p \in \{0, 1\}$$

Here each h_k has been predicted for the k-the example $(\vec{x}_k, c_k)$.

How good are the predictions $h_1, \ldots, h_p$ w.r.t. $c_1, \ldots, c_p$?

There are many possible metrics ...

I will call the class 1 *positive* and the class 0 *negative*.

Note that the class 0 is not negative in the numerical sense but in the absence of something (e.g., predicted illness).

Confusion Matrix for Binary Classifier

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

Confusion Matrix for Binary Classifier

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

- TP = number of correctly classified examples with actual class 1

$$\text{TP} = |\{k \mid h_k = 1 \land c_k = 1\}|$$

Confusion Matrix for Binary Classifier

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

- TP = number of correctly classified examples with actual class 1

$$\mathrm{TP} = |\{k \mid h_k = 1 \wedge c_k = 1\}|$$

- TN = number of correctly classified examples with actual class 0

$$\mathrm{TN} = |\{k \mid h_k = 0 \wedge c_k = 0\}|$$

Confusion Matrix for Binary Classifier

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

- TP = number of correctly classified examples with actual class 1

$$\mathrm{TP} = |\{k \mid h_k = 1 \wedge c_k = 1\}|$$

- TN = number of correctly classified examples with actual class 0

$$\mathrm{TN} = |\{k \mid h_k = 0 \wedge c_k = 0\}|$$

- FP = number of incorrectly classified examples with actual class 0

$$\mathrm{FP} = |\{k \mid h_k = 1 \wedge c_k = 0\}|$$

Confusion Matrix for Binary Classifier

		Predicted	
		1	0
Actual	1	TP	FN
	0	FP	TN

- TP = number of correctly classified examples with actual class 1

  $$\text{TP} = |\{k \mid h_k = 1 \wedge c_k = 1\}|$$

- TN = number of correctly classified examples with actual class 0

  $$\text{TN} = |\{k \mid h_k = 0 \wedge c_k = 0\}|$$

- FP = number of incorrectly classified examples with actual class 0

  $$\text{FP} = |\{k \mid h_k = 1 \wedge c_k = 0\}|$$

- FN = number of correctly classified examples with actual class 1

  $$\text{FN} = |\{k \mid h_k = 0 \wedge c_k = 1\}|$$

Example

Given a sample of 12 individuals, eight have been diagnosed with cancer, and four are cancer-free.

Example

Given a sample of 12 individuals, eight have been diagnosed with cancer, and four are cancer-free.

Assume that we have trained a classifier with the following results:

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	1	1	1	0	0	0	0
Predicted	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Example

Given a sample of 12 individuals, eight have been diagnosed with cancer, and four are cancer-free.

Assume that we have trained a classifier with the following results:

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	1	1	1	0	0	0	0
Predicted	0	0	1	1	1	1	1	1	1	0	0	0
Result	FN	FN	TP	TP	TP	TP	TP	TP	FP	TN	TN	TN

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP $=6$	FN $=2$
Non-cancer	FP $=1$	TN $=3$
Total	$8+4=12$	

Terminology

- TP aka hit
- TN aka correct rejection
- FP aka type I error, false alarm, overestimation
- FN aka type II error, miss, underestimation

Usually, TP, TN, FP, and FN are used to denote the individual examples of a particular kind and the number of these examples.

Terminology

- TP aka hit
- TN aka correct rejection
- FP aka type I error, false alarm, overestimation
- FN aka type II error, miss, underestimation

Usually, TP, TN, FP, and FN are used to denote the individual examples of a particular kind and the number of these examples.

In what follows, we also use

- $\mathrm{P} = \mathrm{TP} + \mathrm{FN}$ of all cases with the *actual* class 1
- $\mathrm{N} = \mathrm{TN} + \mathrm{FP}$ of all cases with the *actual* class 0
- $\mathrm{PP} = \mathrm{TP} + \mathrm{FP}$ of all cases with the *predicted* class 1
- $\mathrm{PN} = \mathrm{TN} + \mathrm{FN}$ of all cases with the *predicted* class 0

Note that $\mathrm{P} + \mathrm{N}$ is the number of all cases.

Terminology

- TP aka hit
- TN aka correct rejection
- FP aka type I error, false alarm, overestimation
- FN aka type II error, miss, underestimation

Usually, TP, TN, FP, and FN are used to denote the individual examples of a particular kind and the number of these examples.

In what follows, we also use

- $\mathrm{P} = \mathrm{TP} + \mathrm{FN}$ of all cases with the *actual* class 1
- $\mathrm{N} = \mathrm{TN} + \mathrm{FP}$ of all cases with the *actual* class 0
- $\mathrm{PP} = \mathrm{TP} + \mathrm{FP}$ of all cases with the *predicted* class 1
- $\mathrm{PN} = \mathrm{TN} + \mathrm{FN}$ of all cases with the *predicted* class 0

Note that $\mathrm{P} + \mathrm{N}$ is the number of all cases.

There is a large number of derived metrics. We consider some of the most used in practice.

Accuracy

$$\text{Accuracy} = \frac{\mathsf{TP} + \mathsf{TN}}{\mathsf{P} + \mathsf{N}}$$

Intuitively, Accuracy is the proportion of correctly classified cases w.r.t. all cases.

Accuracy

$$\text{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{P} + \mathrm{N}}$$

Intuitively, Accuracy is the proportion of correctly classified cases w.r.t. all cases.

Example: Consider our cancer predictor with the confusion matrix

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	$\mathrm{TP} = 6$	$\mathrm{FN} = 2$
Non-cancer	$\mathrm{FP} = 1$	$\mathrm{TN} = 3$
Total	$8 + 4 = 12$	

Accuracy

$$\text{Accuracy} = \frac{\mathsf{TP} + \mathsf{TN}}{\mathsf{P} + \mathsf{N}}$$

Intuitively, Accuracy is the proportion of correctly classified cases w.r.t. all cases.

Example: Consider our cancer predictor with the confusion matrix

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP $= 6$	FN $= 2$
Non-cancer	FP $= 1$	TN $= 3$
Total	$8 + 4 = 12$	

The Accuracy is

$$\text{ACC} = \frac{\mathsf{TP} + \mathsf{TN}}{\mathsf{P} + \mathsf{N}} = \frac{6+3}{12} = \frac{3}{4}$$

Accuracy - Imbalanced Classes

Accuracy can be misleading when the classes are imbalanced:

- Consider 100 cases, 90 in the class 0 and 10 in the class 1,
- consider a classifier that returns 1 for a single sample of class 1 and 0 for all other samples.

Accuracy - Imbalanced Classes

Accuracy can be misleading when the classes are imbalanced:

- Consider 100 cases, 90 in the class 0 and 10 in the class 1,
- consider a classifier that returns 1 for a single sample of class 1 and 0 for all other samples.

Actual	**Predicted**	
	Pos	**Neg**
Pos	1	9
Neg	0	90
Total	$90 + 10 = 100$	

Accuracy - Imbalanced Classes

Accuracy can be misleading when the classes are imbalanced:

- Consider 100 cases, 90 in the class 0 and 10 in the class 1,
- consider a classifier that returns 1 for a single sample of class 1 and 0 for all other samples.

Actual	**Predicted**	
	Pos	**Neg**
Pos	1	9
Neg	0	90
Total	$90 + 10 = 100$	

The Accuracy is $91/100 > 0.9$. Pretty good, right?

Accuracy - Imbalanced Classes

Accuracy can be misleading when the classes are imbalanced:

- Consider 100 cases, 90 in the class 0 and 10 in the class 1,
- consider a classifier that returns 1 for a single sample of class 1 and 0 for all other samples.

Actual	**Predicted**	
	Pos	**Neg**
Pos	1	9
Neg	0	90
Total	$90 + 10 = 100$	

The Accuracy is $91/100 > 0.9$. Pretty good, right?

However, the classifier is pretty bad in the positive cases.

In the case of cancer prediction, such a classifier would be a disaster.

Precision & Recall

To mitigate the defect of the Accuracy, we may compute the following metrics:

$$\text{Precision} = \frac{\text{TP}}{\text{PP}} \quad (= \text{how often is predicted positive actually positive})$$

Precision is also known as positive predictive value (PPV)

Precision & Recall

To mitigate the defect of the Accuracy, we may compute the following metrics:

$$\text{Precision} = \frac{\mathsf{TP}}{\mathsf{PP}} \quad (= \text{how often is predicted positive actually positive})$$

Precision is also known as positive predictive value (PPV)

$$\text{Recall} = \frac{\mathsf{TP}}{\mathsf{P}} \quad (= \text{how often is actually positive predicted positive})$$

Recall is also known as true positive rate, sensitivity, hit rate, and power.

Precision & Recall - Example

Example: In our cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP $= 6$	FN $= 2$
Non-cancer	FP $= 1$	TN $= 3$
Total	$8 + 4 = 12$	

Precision & Recall - Example

Example: In our cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP $= 6$	FN $= 2$
Non-cancer	FP $= 1$	TN $= 3$
Total	$8 + 4 = 12$	

- Precision measures how often is the patient predicted to be ill truly ill (in our case, $6/7$)

Precision & Recall - Example

Example: In our cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP = 6	FN = 2
Non-cancer	FP = 1	TN = 3
Total	$8 + 4 = 12$	

- Precision measures how often is the patient predicted to be ill truly ill (in our case, $6/7$)
- Recall measures how often is an ill patient found to be ill (in our case, $6/8$)

Precision & Recall - Imbalanced Classes

- Consider 100 cases, 90 in the class 0 and 10 in the class 1,

Precision & Recall - Imbalanced Classes

- Consider 100 cases, 90 in the class 0 and 10 in the class 1,
- consider a classifier that returns 1 for a single sample of class 1 and 0 for all other samples.

Actual	**Predicted**	
	Pos	**Neg**
Pos	1	9
Neg	0	90
Total	$90 + 10 = 100$	

Precision & Recall - Imbalanced Classes

- Consider 100 cases, 90 in the class 0 and 10 in the class 1,
- consider a classifier that returns 1 for a single sample of class 1 and 0 for all other samples.

Actual	**Predicted**	
	Pos	**Neg**
Pos	1	9
Neg	0	90
Total	$90 + 10 = 100$	

$$\text{Precision} = 1$$

$$\text{Recall} = \frac{1}{10}$$

You can see that the predictor is very precise (on the class 1) but useless due to the weak Recall.

Precision & Recall - Relative Importance

Let us get back to our cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	$\text{TP} = 6$	$\text{FN} = 2$
Non-cancer	$\text{FP} = 1$	$\text{TN} = 3$
Total	$8 + 4 = 12$	

Consider *Precision* and *Recall*.

By now, you should remember what they measure.

Precision & Recall - Relative Importance

Let us get back to our cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP $= 6$	FN $= 2$
Non-cancer	FP $= 1$	TN $= 3$
Total	$8 + 4 = 12$	

Consider *Precision* and *Recall*.

By now, you should remember what they measure.

Which of the two is more important in medicine?

Precision & Recall - Relative Importance

Let us get back to our cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	$\text{TP} = 6$	$\text{FN} = 2$
Non-cancer	$\text{FP} = 1$	$\text{TN} = 3$
Total	$8 + 4 = 12$	

Consider *Precision* and *Recall*.

By now, you should remember what they measure.

Which of the two is more important in medicine?

Which of the two is more important for plagiarism detectors?

Precision & Recall - Relative Importance

Let us get back to our cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	$\text{TP} = 6$	$\text{FN} = 2$
Non-cancer	$\text{FP} = 1$	$\text{TN} = 3$
Total	$8 + 4 = 12$	

Consider *Precision* and *Recall*.

By now, you should remember what they measure.

Which of the two is more important in medicine?

Which of the two is more important for plagiarism detectors?

Can we get a single number summarizing both Precision and Recall?

For example, to compare two classifiers.

F_1 Score

F_1 score is the harmonic mean of Recall and Precision:

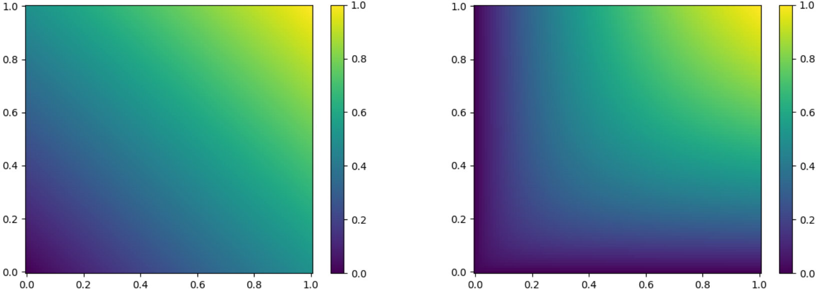
$$\mathsf{F}_1 = \frac{2}{\mathsf{Recall}^{-1} + \mathsf{Precision}^{-1}} = \frac{2\mathsf{TP}}{2\mathsf{TP} + \mathsf{FP} + \mathsf{FN}}$$

F_1 Score

F_1 score is the harmonic mean of Recall and Precision:

$$\mathrm{F}_1 = \frac{2}{\mathrm{Recall}^{-1} + \mathrm{Precision}^{-1}} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}$$

Compare the arithmetic (left) and harmonic (right) mean:



The harmonic mean prefers the two values closer to each other.

For example, the harmonic mean of $2/3$ and $1/3$ is (approx) 0.44444.

F_1 Score - Examples

Consider the cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP = 6	FN = 2
Non-cancer	FP = 1	TN = 3
Total	8 + 4 = 12	

Here $F_1 = \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} = (2 \cdot 6)/((2 \cdot 6) + 1 + 2) = 0.8$.

F_1 Score - Examples

Consider the cancer example:

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP = 6	FN = 2
Non-cancer	FP = 1	TN = 3
Total	8 + 4 = 12	

Here $F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} = (2 \cdot 6)/((2 \cdot 6) + 1 + 2) = 0.8.$

Our imbalanced example:

Actual	**Predicted**	
	Pos	**Neg**
Pos	1	9
Neg	0	90
Total	90 + 10 = 100	

Here $F_1 = \frac{2\mathrm{TP}}{2\mathrm{TP}+\mathrm{FP}+\mathrm{FN}} = (2 \cdot 1)/((2 \cdot 1) + 0 + 9) = 0.18.$

Note that the average of Precision and Recall is 0.55, which would give us a much less severe warning that the classifier is bad.

Imbalanced Classes Once More

Note that the standard definitions of Precision and Recall for binary classifiers reveal only part of the truth.

Imbalanced Classes Once More

Note that the standard definitions of Precision and Recall for binary classifiers reveal only part of the truth.

In particular, *false negatives are not used* in the definition of F_1.

Imbalanced Classes Once More

Note that the standard definitions of Precision and Recall for binary classifiers reveal only part of the truth.

In particular, *false negatives are not used* in the definition of F_1.

Consider

Actual	**Predicted**	
	Pos	**Neg**
Pos	90	0
Neg	9	1
Total	$90 + 10 = 100$	

Imbalanced Classes Once More

Note that the standard definitions of Precision and Recall for binary classifiers reveal only part of the truth.

In particular, *false negatives are not used* in the definition of F_1.

Consider

Actual	**Predicted**	
	Pos	**Neg**
Pos	90	0
Neg	9	1
Total	90 + 10 = 100	

$$\text{Precision} = 90/99 \qquad \text{Recall} = 90/90$$

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = (2 \cdot 90)/(2 \cdot 90 + 9 + 0) = 0.95$$

Imbalanced Classes Once More

Note that the standard definitions of Precision and Recall for binary classifiers reveal only part of the truth.

In particular, *false negatives are not used* in the definition of F_1.

Consider

Actual	**Predicted**	
	Pos	**Neg**
Pos	90	0
Neg	9	1
Total	$90 + 10 = 100$	

$$\text{Precision} = 90/99 \qquad \text{Recall} = 90/90$$

$$F_1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} = (2 \cdot 90)/(2 \cdot 90 + 9 + 0) = 0.95$$

All great, except that the classifier sucks on the negative cases.

If you are concerned with the negative cases, swap the classes and compute another set of metrics.

F_1 Score

- F_1 is often used as a summary score for binary classifiers instead of Accuracy.

  Works better with imbalanced classes.

F_1 Score

- F_1 is often used as a summary score for binary classifiers instead of Accuracy.
  Works better with imbalanced classes.
- Criticised for giving Precision and Recall the same importance.
- Is not symmetric, ignores true negatives, i.e., is misleading for some cases of imbalanced classes.

F_1 Score

- F_1 is often used as a summary score for binary classifiers instead of Accuracy.

  Works better with imbalanced classes.

- Criticised for giving Precision and Recall the same importance.
- Is not symmetric, ignores true negatives, i.e., is misleading for some cases of imbalanced classes.

- *Fowlkes-Mallows index* is a geometric mean of Precision and Recall (used in clustering).

  The geometric mean is between the arithmetic and harmonic mean. For example, the geometric mean of 2/3 and 1/3 is (approx) 0.4714.

More Derived Metrics

Positive predictive value (PPV), precision $= \frac{\mathrm{TP}}{\mathrm{PP}} = 1 - \mathrm{FDR}$	False omission rate (FOR) $= \frac{\mathrm{FN}}{\mathrm{PN}} = 1 - \mathrm{NPV}$
False discovery rate (FDR) $= \frac{\mathrm{FP}}{\mathrm{PP}} = 1 - \mathrm{PPV}$	Negative predictive value (NPV) $= \frac{\mathrm{TN}}{\mathrm{PN}} = 1 - \mathrm{FOR}$

You can see that the negative predictive value becomes the Precision when we swap the classes (and vice versa).

More Derived Metrics

True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{\mathrm{TP}}{\mathrm{P}} = 1 - \mathrm{FNR}$	False negative rate (FNR), miss rate $= \frac{\mathrm{FN}}{\mathrm{P}} = 1 - \mathrm{TPR}$
False positive rate (FPR), probability of false alarm, fall-out $= \frac{\mathrm{FP}}{\mathrm{N}} = 1 - \mathrm{TNR}$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{\mathrm{TN}}{\mathrm{N}} = 1 - \mathrm{FPR}$

Note that *specificity* becomes Recall when we swap the classes (and vice versa).

For example, medical doctors communicate in terms of *sensitivity* and *specificity*.

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP = 6	FN = 2
Non-cancer	FP = 1	TN = 3
Total	8 + 4 = 12	

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	$\text{TP} = 6$	$\text{FN} = 2$
Non-cancer	$\text{FP} = 1$	$\text{TN} = 3$
Total	$8 + 4 = 12$	

$$\text{TPR} = \text{Sensitivity} = \text{Recall} = \text{TP}/\text{P} = 6/8$$

How often is positive predicted positive?

Actual condition	Predicted condition	
	Cancer	**Non-cancer**
Cancer	$\mathrm{TP} = 6$	$\mathrm{FN} = 2$
Non-cancer	$\mathrm{FP} = 1$	$\mathrm{TN} = 3$
Total	$8 + 4 = 12$	

$$\mathrm{TPR} = \mathrm{Sensitivity} = \mathrm{Recall} = \mathrm{TP}/\mathrm{P} = 6/8$$

How often is positive predicted positive?

$$\mathrm{TNR} = \mathrm{Specificity} = \mathrm{TN}/\mathrm{N} = 3/4$$

How often is negative predicted negative?

Actual condition	**Predicted condition**	
	Cancer	**Non-cancer**
Cancer	TP = 6	FN = 2
Non-cancer	FP = 1	TN = 3
Total	8 + 4 = 12	

$$\text{TPR} = \text{Sensitivity} = \text{Recall} = \text{TP}/\text{P} = 6/8$$

How often is positive predicted positive?

$$\text{TNR} = \text{Specificity} = \text{TN}/\text{N} = 3/4$$

How often is negative predicted negative?

$$\text{FPR} = \text{Prob. of false alarm} = \text{FP}/\text{N} = 1/4$$

How often is negative predicted positive?

Actual condition	Predicted condition	
	Cancer	Non-cancer
Cancer	TP $= 6$	FN $= 2$
Non-cancer	FP $= 1$	TN $= 3$
Total	$8 + 4 = 12$	

$$\text{TPR} = \text{Sensitivity} = \text{Recall} = \text{TP}/\text{P} = 6/8$$

How often is positive predicted positive?

$$\text{TNR} = \text{Specificity} = \text{TN}/\text{N} = 3/4$$

How often is negative predicted negative?

$$\text{FPR} = \text{Prob. of false alarm} = \text{FP}/\text{N} = 1/4$$

How often is negative predicted positive?

$$\text{FNR} = \text{Miss rate} = \text{FN}/\text{P} = 2/8$$

How often is positive predicted negative?

Evaluating Multi-class Classifiers

Classification Into Multiple Classes

Assume classification into classes from a finite set C.

Classification Into Multiple Classes

Assume classification into classes from a finite set C.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in C$ for all k.

Classification Into Multiple Classes

Assume classification into classes from a finite set C.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in C$ for all k.

Consider a sequence of predictions generated by a classifier:

$$h_1, \ldots, h_p \in C$$

Here each h_k has been predicted for the k-the example $(\vec{x}_k, c_k)$.

Classification Into Multiple Classes

Assume classification into classes from a finite set C.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in C$ for all k.

Consider a sequence of predictions generated by a classifier:

$$h_1, \ldots, h_p \in C$$

Here each h_k has been predicted for the k-the example $(\vec{x}_k, c_k)$.

How good are the predictions $h_1, \ldots, h_p$ w.r.t. $c_1, \ldots, c_p$?

There are many possible metrics ...

Consider an arbitrary (finite) number of classes in C.

Confusion Matrix

Assume that $C = \{1, \ldots, m\}$.

Confusion Matrix

Assume that $C = \{1, \ldots, m\}$.

Now, given two classes $i, j \in C$ we denote by M_{ij} the number of samples of class i classified into the class j.

Confusion Matrix

Assume that $C = \{1, \ldots, m\}$.

Now, given two classes $i, j \in C$ we denote by M_{ij} the number of samples of class i classified into the class j.

Formally,

$$M_{ij} = |\{k \mid c_k = i \wedge h_k = j\}|$$

Actual	**Predicted**				
	1	$\cdots$	j	$\cdots$	m
1	M_{11}	$\cdots$	M_{1j}	$\cdots$	M_{1m}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
i	M_{i1}	$\cdots$	M_{ij}	$\cdots$	M_{im}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
m	M_{m1}	$\cdots$	M_{mj}	$\cdots$	M_{mm}

Example

Actual	Predicted
big	big
big	big
small	big
medium	medium
big	small
big	big
small	small
small	small
medium	medium
medium	small
small	small
big	big
medium	small
small	medium

Example

Actual	Predicted
big	big
big	big
small	big
medium	medium
big	small
big	big
small	small
small	small
medium	medium
medium	small
small	small
big	big
medium	small
small	medium

Actual	**Predicted**		
	big	**medium**	**small**
big	5	0	1
medium	0	2	2
small	1	1	3

Note that the diagonal counts the correctly classified samples.

The off-diagonal elements correspond to misclassified samples.

Metrics

We can easily generalize Accuracy, Precision, Recall, and F_1-score from the binary classification to multiple classes.

Metrics

We can easily generalize Accuracy, Precision, Recall, and F_1-score from the binary classification to multiple classes.

Notation

- $M_{i\bullet} = \sum_{j=1}^{m} M_{ij}$

Metrics

We can easily generalize Accuracy, Precision, Recall, and F_1-score from the binary classification to multiple classes.

Notation

- $M_{i\bullet} = \sum_{j=1}^{m} M_{ij}$
- $M_{\bullet j} = \sum_{i=1}^{m} M_{ij}$

Metrics

We can easily generalize Accuracy, Precision, Recall, and F_1-score from the binary classification to multiple classes.

Notation

- $M_{i\bullet} = \sum_{j=1}^{m} M_{ij}$
- $M_{\bullet j} = \sum_{i=1}^{m} M_{ij}$
- $M_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{m} M_{ij}$

Metrics

We can easily generalize Accuracy, Precision, Recall, and F_1-score from the binary classification to multiple classes.

Notation

- $M_{i\bullet} = \sum_{j=1}^{m} M_{ij}$
- $M_{\bullet j} = \sum_{i=1}^{m} M_{ij}$
- $M_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{m} M_{ij}$

Now, the metrics:

$$\text{Accuracy} = \frac{\sum_{k=1}^{m} M_{kk}}{M_{\bullet\bullet}}$$

Metrics

We can easily generalize Accuracy, Precision, Recall, and F_1-score from the binary classification to multiple classes.

Notation

- $M_{i\bullet} = \sum_{j=1}^{m} M_{ij}$
- $M_{\bullet j} = \sum_{i=1}^{m} M_{ij}$
- $M_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{m} M_{ij}$

Now, the metrics:

$$\text{Accuracy} = \frac{\sum_{k=1}^{m} M_{kk}}{M_{\bullet\bullet}}$$

For a given class $i \in C$:

$$\text{Precision}[i] = \frac{M_{ii}}{M_{\bullet i}} \qquad \text{Recall}[i] = \frac{M_{ii}}{M_{i\bullet}}$$

Metrics

We can easily generalize Accuracy, Precision, Recall, and F_1-score from the binary classification to multiple classes.

Notation

- $M_{i\bullet} = \sum_{j=1}^{m} M_{ij}$
- $M_{\bullet j} = \sum_{i=1}^{m} M_{ij}$
- $M_{\bullet\bullet} = \sum_{i=1}^{m} \sum_{j=1}^{m} M_{ij}$

Now, the metrics:

$$\text{Accuracy} = \frac{\sum_{k=1}^{m} M_{kk}}{M_{\bullet\bullet}}$$

For a given class $i \in C$:

$$\text{Precision}[i] = \frac{M_{ii}}{M_{\bullet i}} \qquad \text{Recall}[i] = \frac{M_{ii}}{M_{i\bullet}}$$

$$F_1[i] = \frac{2 * \text{Precision}[i] * \text{Recall}[i]}{\text{Precision}[i] + \text{Recall}[i]}$$

Note that Precision, Recall, and F_1 can be defined only for a given class!

Example

Actual	**Predicted**		
	big	**medium**	**small**
big	5	0	1
medium	0	2	2
small	1	1	3

Compute the metrics.

Example

$$\text{Accuracy} = (5+2+3)/15 = 0.66$$

$$\text{Precision[big]} = 5/6$$

$$\text{Precision[medium]} = 2/3$$

$$\text{Precision[small]} = 3/6$$

$$\text{Recall[big]} = 5/6$$

$$\text{Recall[medium]} = 2/4$$

$$\text{Recall[small]} = 3/5$$

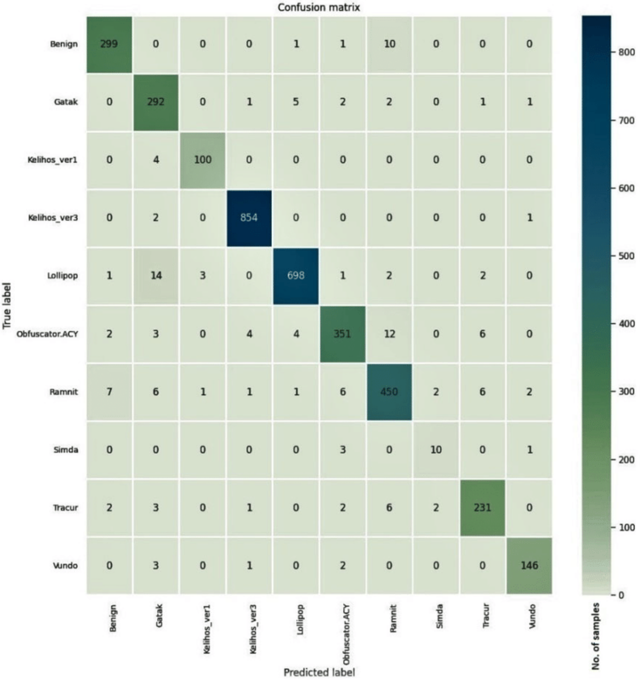
Actual	**Predicted**		
	big	**medium**	**small**
big	5	0	1
medium	0	2	2
small	1	1	3

$$F_1[\text{big}] = \frac{2 * (5/6) * (5/6)}{(5/6) + (5/6)} = 5/6 = 0.83$$

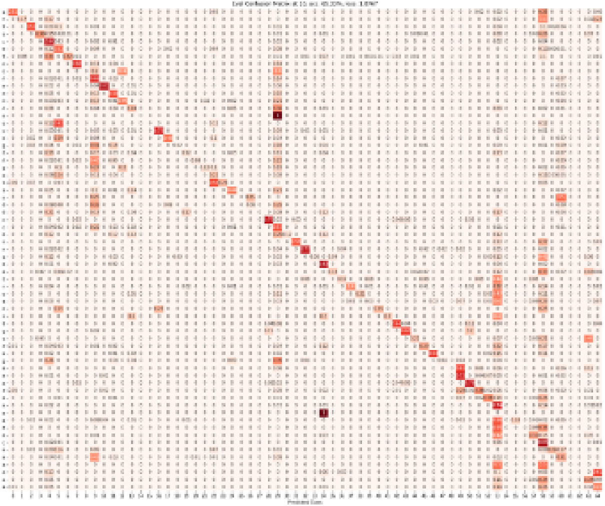
$$F_1[\text{medium}] = 0.57$$

$$F_1[\text{medium}] = 0.54$$

How do you get a single number out of these? Average Precision, Recall, and F_1 are usually computed, but one needs to be careful about the variance.



True label \ Predicted label	Benign	Gatak	Kelihos_ver1	Kelihos_ver3	Lollipop	Obfuscator.ACY	Ramnit	Simda	Tracur	Vundo
Benign	299	0	0	0	1	1	10	0	0	0
Gatak	0	292	0	1	5	2	2	0	1	1
Kelihos_ver1	0	4	100	0	0	0	0	0	0	0
Kelihos_ver3	0	2	0	854	0	0	0	0	0	1
Lollipop	1	14	3	0	698	1	2	0	2	0
Obfuscator.ACY	2	3	0	4	4	351	12	0	6	0
Ramnit	7	6	1	1	1	6	450	2	6	2
Simda	0	0	0	0	0	3	0	10	0	1
Tracur	2	3	0	1	0	2	6	2	231	0
Vundo	0	3	0	1	0	2	0	0	0	146



Machine learning/data mining is needed to understand the matrix.

Probabilistic Classifier Evaluation

Binary Probabilistic Classifier

Assume binary classification into two classes $\{0, 1\}$.

Binary Probabilistic Classifier

Assume binary classification into two classes $\{0, 1\}$.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in C$ for all k.

Binary Probabilistic Classifier

Assume binary classification into two classes $\{0, 1\}$.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in C$ for all k.

Consider a sequence of predictions generated by a classifier.
Now the classifier returns *probability of class* 1 for a given input:

$$h_1, \ldots, h_p \in [0, 1]$$

Here each h_k has been predicted for the k-the example $(\vec{x}_k, c_k)$.

Binary Probabilistic Classifier

Assume binary classification into two classes $\{0, 1\}$.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in C$ for all k.

Consider a sequence of predictions generated by a classifier.
Now the classifier returns *probability of class* 1 for a given input:

$$h_1, \ldots, h_p \in [0, 1]$$

Here each h_k has been predicted for the k-the example $(\vec{x}_k, c_k)$.

How to interpret the predictions $h_1, \ldots, h_p$?

Binary Probabilistic Classifier

Assume binary classification into two classes $\{0, 1\}$.

Consider a classification dataset:

$$\{(\vec{x}_k, c_k) \mid k = 1, \ldots, p\}$$

Here $\vec{x}_k$ is a vector of attributes/features and $c_k \in C$ for all k.

Consider a sequence of predictions generated by a classifier.
Now the classifier returns *probability of class* 1 for a given input:

$$h_1, \ldots, h_p \in [0, 1]$$

Here each h_k has been predicted for the k-the example $(\vec{x}_k, c_k)$.

How to interpret the predictions $h_1, \ldots, h_p$?

How good are the predictions $h_1, \ldots, h_p$ w.r.t. $c_1, \ldots, c_p$?

Probabilistic Classifier

Let us fix predictions $h_1, \ldots, h_p$.

Probabilistic Classifier

Let us fix predictions $h_1, \ldots, h_p$.

Given a threshold $T \in [0, 1]$ we define

$$h_k^T = \begin{cases} 1 & \text{if } h_k \geq T \\ 0 & \text{if } h_k < T \end{cases}$$

For every T we can compute all the metrics (Precision, Recall, etc.)

Probabilistic Classifier

Let us fix predictions $h_1, \ldots, h_p$.

Given a threshold $T \in [0, 1]$ we define

$$h_k^T = \begin{cases} 1 & \text{if } h_k \geq T \\ 0 & \text{if } h_k < T \end{cases}$$

For every T we can compute all the metrics (Precision, Recall, etc.)

Given a metric MET and a threshold T, we denote by MET[T] the metric MET evaluated on $h_1^T, \ldots, h_p^T$.

Probabilistic Classifier

Let us fix predictions $h_1, \ldots, h_p$.

Given a threshold $T \in [0,1]$ we define

$$h_k^T = \begin{cases} 1 & \text{if } h_k \geq T \\ 0 & \text{if } h_k < T \end{cases}$$

For every T we can compute all the metrics (Precision, Recall, etc.)

Given a metric MET and a threshold T, we denote by $\mathrm{MET}[T]$ the metric MET evaluated on $h_1^T, \ldots, h_p^T$.

We obtain

$$\mathrm{TP}[T] = |\{k \mid h_k^T = 1 \wedge c_k = 1\}|$$

Probabilistic Classifier

Let us fix predictions $h_1, \ldots, h_p$.

Given a threshold $T \in [0,1]$ we define

$$h_k^T = \begin{cases} 1 & \text{if } h_k \geq T \\ 0 & \text{if } h_k < T \end{cases}$$

For every T we can compute all the metrics (Precision, Recall, etc.)

Given a metric MET and a threshold T, we denote by $\text{MET}[T]$ the metric MET evaluated on $h_1^T, \ldots, h_p^T$.

We obtain

$$\text{TP}[T] = |\{k \mid h_k^T = 1 \wedge c_k = 1\}|$$

and

$\text{TN}[T], \text{FP}[T], \text{FN}[T], \text{Accuracy}[T], \text{Precision}[T], \text{Recall}[T], F_1[T], \ldots$

Probabilistic Classifier

Let us fix predictions $h_1, \dots, h_p$.

Given a threshold $T \in [0, 1]$ we define

$$h_k^T = \begin{cases} 1 & \text{if } h_k \geq T \\ 0 & \text{if } h_k < T \end{cases}$$

For every T we can compute all the metrics (Precision, Recall, etc.)

Given a metric MET and a threshold T, we denote by $\text{MET}[T]$ the metric MET evaluated on $h_1^T, \dots, h_p^T$.

We obtain

$$\text{TP}[T] = |\{k \mid h_k^T = 1 \wedge c_k = 1\}|$$

and
$\text{TN}[T], \text{FP}[T], \text{FN}[T], \text{Accuracy}[T], \text{Precision}[T], \text{Recall}[T], F_1[T], \dots$

However, all metrics are now functions of the threshold T.

Thresholded Classifier Metrics

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05
T=0.5	TP	TP	TP	TP	TP	TN	TN	FN	FN	TN	TN	TN
T=0.42	TP	TP	TP	TP	TP	FP	FP	TP	FN	TN	TN	TN
T=0.1	TP	TP	TP	TP	TP	FP	FP	TP	TP	FP	FP	TN

Thresholded Classifier Metrics

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05
T=0.5	TP	TP	TP	TP	TP	TN	TN	FN	FN	TN	TN	TN
T=0.42	TP	TP	TP	TP	TP	FP	FP	TP	FN	TN	TN	TN
T=0.1	TP	TP	TP	TP	TP	FP	FP	TP	TP	FP	FP	TN

For example, consider $T = 0.42$, then

Thresholded Classifier Metrics

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05
T=0.5	TP	TP	TP	TP	TP	TN	TN	FN	FN	TN	TN	TN
T=0.42	TP	TP	TP	TP	TP	FP	FP	TP	FN	TN	TN	TN
T=0.1	TP	TP	TP	TP	TP	FP	FP	TP	TP	FP	FP	TN

For example, consider $T = 0.42$, then

$$\mathrm{TP}[T] = 6 \quad \mathrm{FP}[T] = 2 \quad \mathrm{FN}[T] = 1 \quad \mathrm{TN}[T] = 3$$

Thresholded Classifier Metrics

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05
T=0.5	TP	TP	TP	TP	TP	TN	TN	FN	FN	TN	TN	TN
T=0.42	TP	TP	TP	TP	TP	FP	FP	TP	FN	TN	TN	TN
T=0.1	TP	TP	TP	TP	TP	FP	FP	TP	TP	FP	FP	TN

For example, consider $T = 0.42$, then

$$\text{TP}[T] = 6 \quad \text{FP}[T] = 2 \quad \text{FN}[T] = 1 \quad \text{TN}[T] = 3$$

$$\text{Accuracy}[T] = \frac{3+6}{12} \quad \text{Precision}[T] = \frac{6}{6+2} \quad \text{Recall}[T] = \frac{5}{6+1}$$

Thresholded Classifier Metrics

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05
T=0.5	TP	TP	TP	TP	TP	TN	TN	FN	FN	TN	TN	TN
T=0.42	TP	TP	TP	TP	TP	FP	FP	TP	FN	TN	TN	TN
T=0.1	TP	TP	TP	TP	TP	FP	FP	TP	TP	FP	FP	TN

For example, consider $T = 0.42$, then

$$\mathrm{TP}[T] = 6 \quad \mathrm{FP}[T] = 2 \quad \mathrm{FN}[T] = 1 \quad \mathrm{TN}[T] = 3$$

$$\mathrm{Accuracy}[T] = \frac{3+6}{12} \quad \mathrm{Precision}[T] = \frac{6}{6+2} \quad \mathrm{Recall}[T] = \frac{5}{6+1}$$

$$F_1[T] = \frac{2 \cdot 6/8 \cdot 5/7}{6/8 + 5/7} = 0.73$$

Receiver Operating Characteristic (ROC)

Consider two metrics for a given T:

$$\mathrm{TPR}[T] = \frac{\mathrm{TP}[\mathrm{T}]}{\mathrm{P}[T]} \qquad \text{(True Positive Rate)}$$

Receiver Operating Characteristic (ROC)

Consider two metrics for a given T:

$$\mathrm{TPR}[T] = \frac{\mathrm{TP[T]}}{\mathrm{P}[T]} \qquad \text{(True Positive Rate)}$$

$$\mathrm{FPR}[T] = \frac{\mathrm{FP}[T]}{\mathrm{N}[T]} \qquad \text{(False Positive Rate)}$$

Receiver Operating Characteristic (ROC)

Consider two metrics for a given T:

$$\mathrm{TPR}[T] = \frac{\mathrm{TP[T]}}{\mathrm{P}[T]} \qquad \text{(True Positive Rate)}$$

$$\mathrm{FPR}[T] = \frac{\mathrm{FP}[T]}{\mathrm{N}[T]} \qquad \text{(False Positive Rate)}$$

ROC curve is then a function $\mathrm{ROC} : [0,1] \to [0,1]^2$ defined by

$$\mathrm{ROC}(T) = (\mathrm{TPR}[T], \mathrm{FPR}[T])$$

Receiver Operating Characteristic (ROC)

Consider two metrics for a given T:

$$\mathrm{TPR}[T] = \frac{\mathrm{TP[T]}}{\mathrm{P}[T]} \qquad \text{(True Positive Rate)}$$

$$\mathrm{FPR}[T] = \frac{\mathrm{FP}[T]}{\mathrm{N}[T]} \qquad \text{(False Positive Rate)}$$

ROC curve is then a function $\mathrm{ROC} : [0,1] \to [0,1]^2$ defined by

$$\mathrm{ROC}(T) = (\mathrm{TPR}[T], \mathrm{FPR}[T])$$

Observe that

$$\mathrm{ROC}(0) = (1,1)$$

Because the classifier with $T = 0$ simply classifies everything as positive, i.e., into the class 1.

Both $\mathrm{TPR}[T]$ and $\mathrm{FPR}[T]$ are non-increasing in T.

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \leq 0.36$: TPR $= 1$ and FPR $= 2/5$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \leq 0.36$: TPR $= 1$ and FPR $= 2/5$
- $0.36 < T \leq 0.40$: TPR $= 6/7$ and FPR $= 2/5$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \leq 0.36$: TPR $= 1$ and FPR $= 2/5$
- $0.36 < T \leq 0.40$: TPR $= 6/7$ and FPR $= 2/5$
- $0.40 < T \leq 0.42$: TPR $= 5/7$ and FPR $= 2/5$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \leq 0.36$: TPR $= 1$ and FPR $= 2/5$
- $0.36 < T \leq 0.40$: TPR $= 6/7$ and FPR $= 2/5$
- $0.40 < T \leq 0.42$: TPR $= 5/7$ and FPR $= 2/5$
- $0.42 < T \leq 0.48$: TPR $= 5/7$ and FPR $= 1/5$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \leq 0.36$: TPR $= 1$ and FPR $= 2/5$
- $0.36 < T \leq 0.40$: TPR $= 6/7$ and FPR $= 2/5$
- $0.40 < T \leq 0.42$: TPR $= 5/7$ and FPR $= 2/5$
- $0.42 < T \leq 0.48$: TPR $= 5/7$ and FPR $= 1/5$
- $0.48 < T \leq 0.66$: TPR $= 5/7$ and FPR $= 0$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \le T \le 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \le 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \le 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \le 0.36$: TPR $= 1$ and FPR $= 2/5$
- $0.36 < T \le 0.40$: TPR $= 6/7$ and FPR $= 2/5$
- $0.40 < T \le 0.42$: TPR $= 5/7$ and FPR $= 2/5$
- $0.42 < T \le 0.48$: TPR $= 5/7$ and FPR $= 1/5$
- $0.48 < T \le 0.66$: TPR $= 5/7$ and FPR $= 0$
- $0.66 < T \le 0.86$: TPR $= 4/7$ and FPR $= 0$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \leq 0.36$: TPR $= 1$ and FPR $= 2/5$
- $0.36 < T \leq 0.40$: TPR $= 6/7$ and FPR $= 2/5$
- $0.40 < T \leq 0.42$: TPR $= 5/7$ and FPR $= 2/5$
- $0.42 < T \leq 0.48$: TPR $= 5/7$ and FPR $= 1/5$
- $0.48 < T \leq 0.66$: TPR $= 5/7$ and FPR $= 0$
- $0.66 < T \leq 0.86$: TPR $= 4/7$ and FPR $= 0$
- $0.86 < T \leq 0.90$: TPR $= 3/7$ and FPR $= 0$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

- $0.00 \leq T \leq 0.05$: $\text{TPR} = 1$ and $\text{FPR} = 1$
- $0.05 < T \leq 0.10$: $\text{TPR} = 1$ and $\text{FPR} = 4/5$
- $0.10 < T \leq 0.15$: $\text{TPR} = 1$ and $\text{FPR} = 3/5$
- $0.15 < T \leq 0.36$: $\text{TPR} = 1$ and $\text{FPR} = 2/5$
- $0.36 < T \leq 0.40$: $\text{TPR} = 6/7$ and $\text{FPR} = 2/5$
- $0.40 < T \leq 0.42$: $\text{TPR} = 5/7$ and $\text{FPR} = 2/5$
- $0.42 < T \leq 0.48$: $\text{TPR} = 5/7$ and $\text{FPR} = 1/5$
- $0.48 < T \leq 0.66$: $\text{TPR} = 5/7$ and $\text{FPR} = 0$
- $0.66 < T \leq 0.86$: $\text{TPR} = 4/7$ and $\text{FPR} = 0$
- $0.86 < T \leq 0.90$: $\text{TPR} = 3/7$ and $\text{FPR} = 0$
- $0.90 < T \leq 0.95$: $\text{TPR} = 2/7$ and $\text{FPR} = 0$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

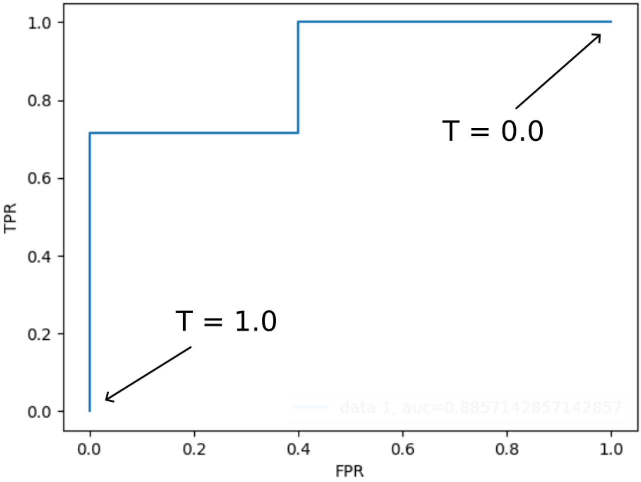
- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \leq 0.36$: TPR $= 1$ and FPR $= 2/5$
- $0.36 < T \leq 0.40$: TPR $= 6/7$ and FPR $= 2/5$
- $0.40 < T \leq 0.42$: TPR $= 5/7$ and FPR $= 2/5$
- $0.42 < T \leq 0.48$: TPR $= 5/7$ and FPR $= 1/5$
- $0.48 < T \leq 0.66$: TPR $= 5/7$ and FPR $= 0$
- $0.66 < T \leq 0.86$: TPR $= 4/7$ and FPR $= 0$
- $0.86 < T \leq 0.90$: TPR $= 3/7$ and FPR $= 0$
- $0.90 < T \leq 0.95$: TPR $= 2/7$ and FPR $= 0$
- $0.95 < T \leq 0.98$: TPR $= 1/7$ and FPR $= 0$

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

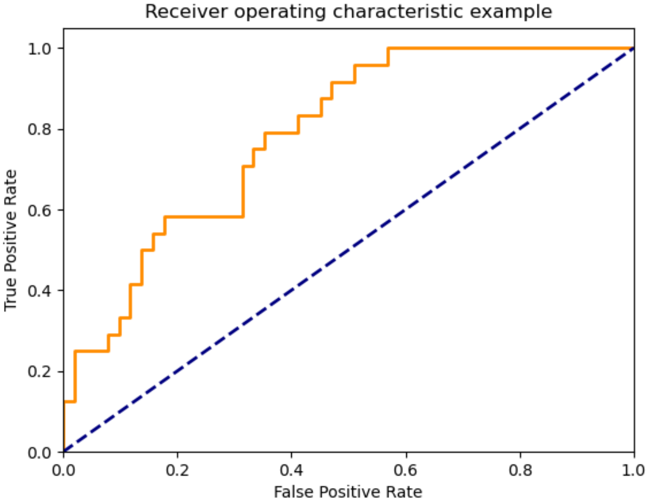
- $0.00 \leq T \leq 0.05$: TPR $= 1$ and FPR $= 1$
- $0.05 < T \leq 0.10$: TPR $= 1$ and FPR $= 4/5$
- $0.10 < T \leq 0.15$: TPR $= 1$ and FPR $= 3/5$
- $0.15 < T \leq 0.36$: TPR $= 1$ and FPR $= 2/5$
- $0.36 < T \leq 0.40$: TPR $= 6/7$ and FPR $= 2/5$
- $0.40 < T \leq 0.42$: TPR $= 5/7$ and FPR $= 2/5$
- $0.42 < T \leq 0.48$: TPR $= 5/7$ and FPR $= 1/5$
- $0.48 < T \leq 0.66$: TPR $= 5/7$ and FPR $= 0$
- $0.66 < T \leq 0.86$: TPR $= 4/7$ and FPR $= 0$
- $0.86 < T \leq 0.90$: TPR $= 3/7$ and FPR $= 0$
- $0.90 < T \leq 0.95$: TPR $= 2/7$ and FPR $= 0$
- $0.95 < T \leq 0.98$: TPR $= 1/7$ and FPR $= 0$
- $0.98 < T \leq 1.00$: TPR $= 0$ and FPR $= 0$

ROC

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

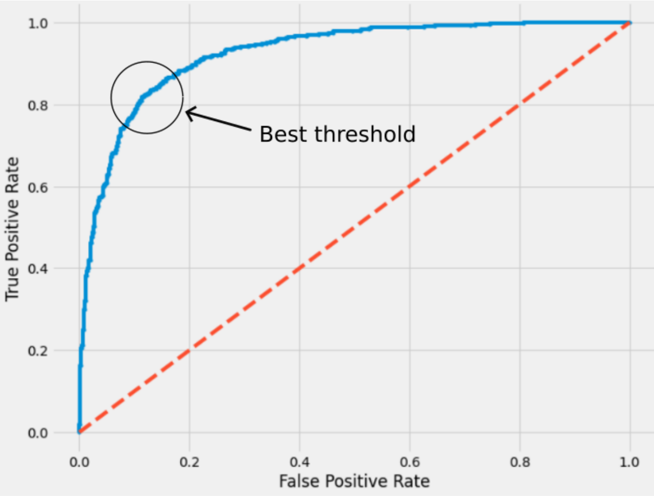


Iris Dataset - A Classifier



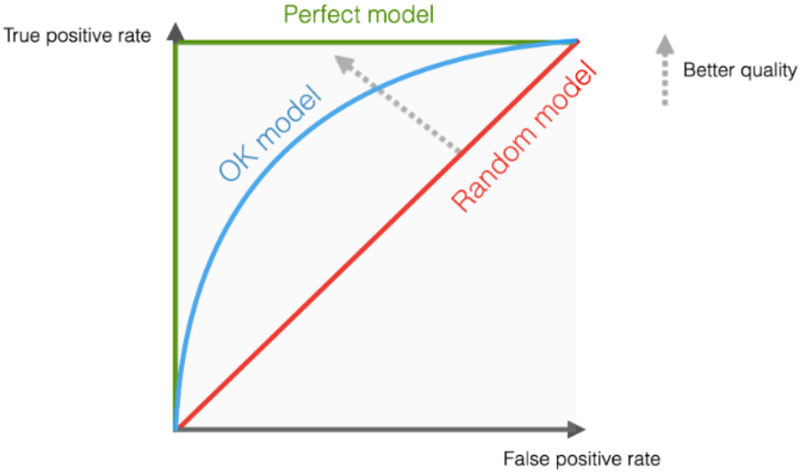
Example from the scikit-learn manual - SVM classifier trained in Iris

Using ROC and Threshold



Search for the best threshold at the elbow of the ROC curve.

ROC - Explanation

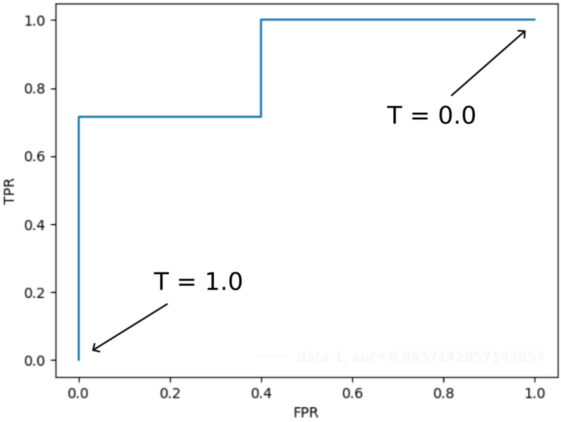


The larger the *area under the ROC curve (ROC-AUC)*, the better.

ROC-AUC ranges from 0 to 1. ROC-AUC ≈ 0.5 indicates random guessing.

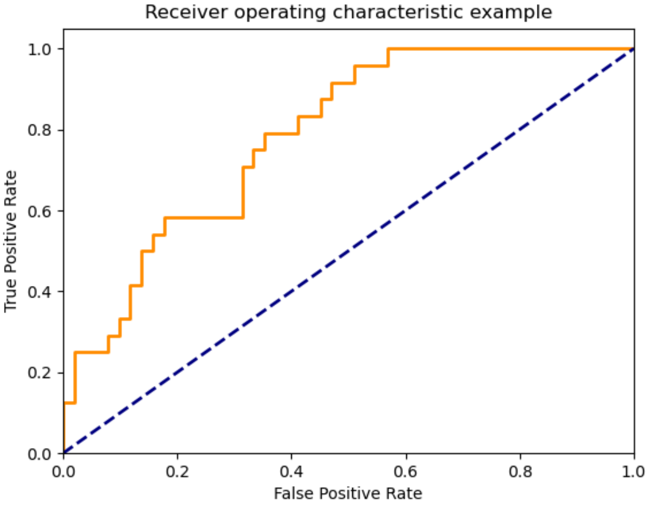
ROC-AUC

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05



ROC-AUC $= 0.8857$

Iris - ROC-AUC



ROC-AUC $= 0.79$

ROC-AUC - Probabilistic Interpretation

How is the ROC-AUC connected with the samples?

ROC-AUC - Probabilistic Interpretation

How is the ROC-AUC connected with the samples?

Consider our cancer detection example:

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

ROC-AUC - Probabilistic Interpretation

How is the ROC-AUC connected with the samples?

Consider our cancer detection example:

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

AUC has a probabilistic explanation:

Consider the following experiment:

ROC-AUC - Probabilistic Interpretation

How is the ROC-AUC connected with the samples?

Consider our cancer detection example:

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

AUC has a probabilistic explanation:

Consider the following experiment:

- Choose randomly a patient i from positive patients
  Each positive patient has the same probability of being chosen.

ROC-AUC - Probabilistic Interpretation

How is the ROC-AUC connected with the samples?

Consider our cancer detection example:

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

AUC has a probabilistic explanation:

Consider the following experiment:

- Choose randomly a patient i from positive patients
  Each positive patient has the same probability of being chosen.
- Choose randomly a patient j from negative patients
  Each negative patient has the same probability of being chosen.

ROC-AUC - Probabilistic Interpretation

How is the ROC-AUC connected with the samples?

Consider our cancer detection example:

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

AUC has a probabilistic explanation:

Consider the following experiment:

- Choose randomly a patient i from positive patients
  Each positive patient has the same probability of being chosen.
- Choose randomly a patient j from negative patients
  Each negative patient has the same probability of being chosen.
- Check if $h_i \geq h_j$.

ROC-AUC - Probabilistic Interpretation

How is the ROC-AUC connected with the samples?

Consider our cancer detection example:

Index	1	2	3	4	5	6	7	8	9	10	11	12
Actual	1	1	1	1	1	0	0	1	1	0	0	0
Predicted	.98	.95	.9	.86	.66	.48	.42	.4	.36	.15	.1	.05

AUC has a probabilistic explanation:

Consider the following experiment:

- Choose randomly a patient i from positive patients
  Each positive patient has the same probability of being chosen.
- Choose randomly a patient j from negative patients
  Each negative patient has the same probability of being chosen.
- Check if $h_i \geq h_j$.

The ROC-AUC is the probability of succeeding in the $h_i \geq h_j$ test.

Summary

We have discussed various metrics that can be used to evaluate the quality of a classifier.

The metrics summarize the results of evaluation on a given dataset.

Summary

We have discussed various metrics that can be used to evaluate the quality of a classifier.

The metrics summarize the results of evaluation on a given dataset.

We have discussed metrics for evaluating

- binary classifiers,
  Accuracy, Precision, Recall, F_1, and few more
- multi-class classifiers,
  Accuracy, Precision, Recall, F_1
- probabilistic classifiers,
  parametrized metrics, ROC-AUC

Summary

We have discussed various metrics that can be used to evaluate the quality of a classifier.

The metrics summarize the results of evaluation on a given dataset.

We have discussed metrics for evaluating

- binary classifiers,
  Accuracy, Precision, Recall, F_1, and few more
- multi-class classifiers,
  Accuracy, Precision, Recall, F_1
- probabilistic classifiers,
  parametrized metrics, ROC-AUC

There are still several questions unanswered:

- When to use the metrics.
- How to estimate the influence of sampling the dataset.

Use of Evaluation Metrics

In our case, the following scenarios are typical:

- **Final test**: Evaluate the model on the test set (separated at the beginning of training) and then compute the metrics. May inform the user about the quality of the model.

Use of Evaluation Metrics

In our case, the following scenarios are typical:

- **Final test**: Evaluate the model on the test set (separated at the beginning of training) and then compute the metrics. May inform the user about the quality of the model.
- **Validation**: Evaluate models on a separate validation set and use the metrics to compare models.
  There are (at least) two scenarios in which this happens:
  - Hyperparameter fine-tuning.
  - Comparison of different models (e.g., KNN and decision trees).

Use of Evaluation Metrics

In our case, the following scenarios are typical:

- **Final test**: Evaluate the model on the test set (separated at the beginning of training) and then compute the metrics. May inform the user about the quality of the model.
- **Validation**: Evaluate models on a separate validation set and use the metrics to compare models.
  There are (at least) two scenarios in which this happens:
  - Hyperparameter fine-tuning.
  - Comparison of different models (e.g., KNN and decision trees).

Keep in mind that the metrics are artificial, and the results of the model are roughly summarized.

It would be best if you always strived to test the proper functionality of your model in as natural conditions as possible.

For example, a model for medical diagnosis should be evaluated by medical doctors who may observe many features of its behavior that are difficult to express quantitatively.

How to Estimate Significance

Machine learning models are typically trained on (pseudo) random samples of data objects.

For example, a set of patients treated by the concrete hospital.

How to Estimate Significance

Machine learning models are typically trained on (pseudo) random samples of data objects.

For example, a set of patients treated by the concrete hospital.

However, the purpose of testing/evaluation is to get information about the whole population (i.e., all possible patients).

How to Estimate Significance

Machine learning models are typically trained on (pseudo) random samples of data objects.

For example, a set of patients treated by the concrete hospital.

However, the purpose of testing/evaluation is to get information about the whole population (i.e., all possible patients).

How do we estimate how much specific properties of the given sample influence our model?

This is a challenging question; methods of inferential statistics are needed to get the answer.

How to Estimate Significance

Machine learning models are typically trained on (pseudo) random samples of data objects.

For example, a set of patients treated by the concrete hospital.

However, the purpose of testing/evaluation is to get information about the whole population (i.e., all possible patients).

How do we estimate how much specific properties of the given sample influence our model?

This is a challenging question; methods of inferential statistics are needed to get the answer.

We will consider these issues in some later lecture. Concretely,

- *Bias-variance* tradeoff
- *Statistical tests* for testing
  - significance of the metrics values,
  - paired t-tests for comparing models.

How to Compare Classifiers

Let us consider two classifiers. How do you compare them?

How to Compare Classifiers

Let us consider two classifiers. How do you compare them?

Accuracies and F_1 scores can be compared easily (they are just numbers).

How to Compare Classifiers

Let us consider two classifiers. How do you compare them?

Accuracies and F_1 scores can be compared easily (they are just numbers).

How to compare $(\mathrm{Precision}_1, \mathrm{Recall}_1)$ of the fist classifier with $(\mathrm{Precision}_2, \mathrm{Recall}_2)$ of the second classifier?

How to Compare Classifiers

Let us consider two classifiers. How do you compare them?

Accuracies and F_1 scores can be compared easily (they are just numbers).

How to compare $(\text{Precision}_1, \text{Recall}_1)$ of the fist classifier with $(\text{Precision}_2, \text{Recall}_2)$ of the second classifier?

Thresholding

- Introduce a threshold $0 \leq t \leq 1$
- Demand, one of the two metrics (typically the Recall), to be at least t. That is

$$\text{Recall}_1 \geq t \qquad \text{Recall}_2 \geq t$$

- Compare the values of the other metric numerically. In our case, decide whether

$$\text{Precision}_1 \geq \text{Precision}_2$$

(Still need to be concerned about the statistical significance.)

Example

Actual condition	**Predicted condition**	
	Canc.	**Non-canc.**
Cancer	6	2
Non-canc.	1	3
Total	$8 + 4 = 12$	

Actual condition	**Predicted condition**	
	Canc.	**Non-canc.**
Cancer	5	3
Non-canc.	0	4
Total	$8 + 4 = 12$	

$$\text{Precision}_1 = \frac{6}{7} \qquad \text{Recall}_1 = \frac{6}{8}$$

$$\text{Precision}_2 = \frac{5}{5} = 1 \qquad \text{Recall}_2 = \frac{5}{8}$$

Consider a threshold t on the Recall.

The second classifier is better if the threshold t is $5/8$, then the second classifier is better.

If the threshold t is $6/8$, then the second classifier is unacceptable.