Probabilistic Classification

Probabilistic Classification – Idea

Imagine that

- I look out of a window and see a bird,
- it is black, approx. 25 cm long, and has a rather yellow beak.

Probabilistic Classification – Idea

Imagine that

- I look out of a window and see a bird,
- it is black, approx. 25 cm long, and has a rather yellow beak.

My daughter asks: What kind of bird is this?

Probabilistic Classification – Idea

Imagine that

- I look out of a window and see a bird,
- it is black, approx. 25 cm long, and has a rather yellow beak.

My daughter asks: What kind of bird is this?

My usual answer: This is *probably* a kind of blackbird (kos černý in Czech).

Probabilistic Classification – Idea

Imagine that

- I look out of a window and see a bird,
- it is black, approx. 25 cm long, and has a rather yellow beak.

My daughter asks: What kind of bird is this?

My usual answer: This is *probably* a kind of blackbird (kos černý in Czech).

Here *probably* means that out of my extensive catalogue of four kinds of birds that I am able to recognize, "blackbird" gets the highest degree of belief based on *features* of this particular bird.

Frequentists might say that the largest proportion of birds with similar features I have ever seen were blackbirds.

Probabilistic Classification – Idea

Imagine that

- I look out of a window and see a bird,
- it is black, approx. 25 cm long, and has a rather yellow beak.

My daughter asks: What kind of bird is this?

My usual answer: This is *probably* a kind of blackbird (kos černý in Czech).

Here *probably* means that out of my extensive catalogue of four kinds of birds that I am able to recognize, "blackbird" gets the highest degree of belief based on *features* of this particular bird.

Frequentists might say that the largest proportion of birds with similar features I have ever seen were blackbirds.

The degree of belief (Bayesians), or the relative frequency (frequentists) is the *probability*.

Basic Discrete Probability Theory

- A finite or countably infinite set Ω of *possible outcomes*, Ω is called *sample space*.

  Experiment: Roll one dice once. Sample space: $\Omega = \{1, \ldots, 6\}$

Basic Discrete Probability Theory

- A finite or countably infinite set Ω of *possible outcomes*, Ω is called *sample space*.

  Experiment: Roll one dice once. Sample space: $\Omega = \{1, \ldots, 6\}$
- Each element ω of Ω is assigned a "probability" value $f(\omega)$, here f must satisfy
  - $f(\omega) \in [0, 1]$ for all $\omega \in \Omega$,
  - $\sum_{\omega \in \Omega} f(\omega) = 1$.

  If the dice is fair, then $f(\omega) = \frac{1}{6}$ for all $\omega \in \{1, \ldots, 6\}$.

Basic Discrete Probability Theory

- A finite or countably infinite set Ω of *possible outcomes*, Ω is called *sample space*.

  Experiment: Roll one dice once. Sample space: $\Omega = \{1, \ldots, 6\}$
- Each element ω of Ω is assigned a "probability" value $f(\omega)$, here f must satisfy
  - $f(\omega) \in [0, 1]$ for all $\omega \in \Omega$,
  - $\sum_{\omega \in \Omega} f(\omega) = 1$.

  If the dice is fair, then $f(\omega) = \frac{1}{6}$ for all $\omega \in \{1, \ldots, 6\}$.
- An *event* is any subset E of Ω.
- The *probability* of a given event $E \subseteq \Omega$ is defined as

$$P(E) = \sum_{\omega \in E} f(\omega)$$

  Let E be the event that an odd number is rolled, i.e., $E = \{1, 3, 5\}$. Then $P(E) = \frac{1}{2}$.

Basic Discrete Probability Theory

- A finite or countably infinite set Ω of *possible outcomes*, Ω is called *sample space*.

  Experiment: Roll one dice once. Sample space: $\Omega = \{1, \dots, 6\}$
- Each element ω of Ω is assigned a "probability" value $f(\omega)$, here f must satisfy
  - $f(\omega) \in [0, 1]$ for all $\omega \in \Omega$,
  - $\sum_{\omega \in \Omega} f(\omega) = 1$.

  If the dice is fair, then $f(\omega) = \frac{1}{6}$ for all $\omega \in \{1, \dots, 6\}$.
- An *event* is any subset E of Ω.
- The *probability* of a given event $E \subseteq \Omega$ is defined as

$$P(E) = \sum_{\omega \in E} f(\omega)$$

  Let E be the event that an odd number is rolled, i.e., $E = \{1, 3, 5\}$. Then $P(E) = \frac{1}{2}$.
- Basic laws: $P(\Omega) = 1$, $P(\emptyset) = 0$, given disjoint sets A, B we have $P(A \cup B) = P(A) + P(B)$, $P(\Omega \setminus A) = 1 - P(A)$.

Conditional Probability and Independence

- $P(A \mid B)$ is the probability of A given B (assume $P(B) > 0$) defined by

  $$P(A \mid B) = P(A \cap B)/P(B)$$

  (We assume that B is all and only information known.)

  A fair dice: what is the probability that 3 is rolled assuming that an odd number is rolled? ... and assuming that an even number is rolled?

Conditional Probability and Independence

- $P(A \mid B)$ is the probability of A given B (assume $P(B) > 0$) defined by

  $$P(A \mid B) = P(A \cap B)/P(B)$$

  (We assume that B is all and only information known.)

  A fair dice: what is the probability that 3 is rolled assuming that an odd number is rolled? ... and assuming that an even number is rolled?

- A and B are independent if $P(A \cap B) = P(A) \cdot P(B)$.

  It is easy to show that if $P(B) > 0$, then

  $$A, B \text{ are independent iff } P(A \mid B) = P(A).$$

Random Variables and Random Vectors

- A *random variable* X is a function $X : \Omega \to \mathbb{R}$.
  A dice: $X : \{1, \dots, 6\} \to \{0, 1\}$ such that $X(n) = n \mod 2$.
- A *random vector* is a function $X : \Omega \to \mathbb{R}^d$.

Random Variables and Random Vectors

- A *random variable* X is a function $X : \Omega \to \mathbb{R}$.
  A dice: $X : \{1, \ldots, 6\} \to \{0, 1\}$ such that $X(n) = n \mod 2$.
- A *random vector* is a function $X : \Omega \to \mathbb{R}^d$.
  We use $X = (X_1, \ldots, X_d)$ where X_i is a random variable returning the i-th component of X.
- Consider random variables X_1, X_2 and Y. The variables X_1, X_2 are *conditionally independent given Y* if for all x_1, x_2 and y we have that

$$P(X_1 = x_1, X_2 = x_2 \mid Y = y) = P(X_1 = x_1 \mid Y = y) \cdot P(X_2 = x_2 \mid Y = y)$$

Random Vectors – Example

Let Ω be a space of colored geometric shapes that are divided into two categories ($\mathbf{1}$ and $\mathbf{0}$).

Assume a random vector $X = (X_{color}, X_{shape}, X_{cat})$ where

- $X_{color} : \Omega \rightarrow \{red, blue\}$,
- $X_{shape} : \Omega \rightarrow \{circle, square\}$,
- $X_{cat} : \Omega \rightarrow \{\mathbf{1}, \mathbf{0}\}$.

Probability distribution of values is given by the following tables:

category $\mathbf{1}$:

	circle	square
red	0.2	0.02
blue	0.02	0.01

category $\mathbf{0}$:

	circle	square
red	0.05	0.3
blue	0.2	0.2

Random Vectors – Example

Example:

$P(red, circle, \mathbf{1}) = P(X_{color} = red, X_{shape} = circle, X_{cat} = \mathbf{1}) = 0.2$

Random Vectors – Example

Example:

$P(red, circle, \mathbf{1}) = P(X_{color} = red, X_{shape} = circle, X_{cat} = \mathbf{1}) = 0.2$

"Summing over" all possible values of some variable(s) gives the distribution of the rest:

$$\begin{aligned} P(red, circle) &= P(X_{color} = red, X_{shape} = circle) \\ &= P(red, circle, \mathbf{1}) + P(red, circle, \mathbf{0}) \\ &= 0.2 + 0.05 = 0.25 \end{aligned}$$

Random Vectors – Example

Example:

$P(red, circle, \mathbf{1}) = P(X_{color} = red, X_{shape} = circle, X_{cat} = \mathbf{1}) = 0.2$

"Summing over" all possible values of some variable(s) gives the distribution of the rest:

$$\begin{aligned} P(red, circle) &= P(X_{color} = red, X_{shape} = circle) \\ &= P(red, circle, \mathbf{1}) + P(red, circle, \mathbf{0}) \\ &= 0.2 + 0.05 = 0.25 \end{aligned}$$

$$P(red) = 0.2 + 0.02 + 0.05 + 0.3 = 0.57$$

Random Vectors – Example

Example:
$P(red, circle, \mathbf{1}) = P(X_{color} = red, X_{shape} = circle, X_{cat} = \mathbf{1}) = 0.2$

"Summing over" all possible values of some variable(s) gives the distribution of the rest:

$$\begin{aligned} P(red, circle) &= P(X_{color} = red, X_{shape} = circle) \\ &= P(red, circle, \mathbf{1}) + P(red, circle, \mathbf{0}) \\ &= 0.2 + 0.05 = 0.25 \end{aligned}$$

$$P(red) = 0.2 + 0.02 + 0.05 + 0.3 = 0.57$$

Thus also all conditional probabilities can be computed:

$$P(\mathbf{1} \mid red, circle) = \frac{P(red, circle, \mathbf{1})}{P(red, circle)} = \frac{0.2}{0.25} = 0.8$$

Bayesian Classification

Let Ω be a sample space (a universum) of all objects that can be classified. We assume a probability P on Ω.

Bayesian Classification

Let Ω be a sample space (a universum) of all objects that can be classified. We assume a probability P on Ω.

We consider the problem of **binary classification**:

- Let Y be the random variable for the category which takes values in $\{\mathbf{0}, \mathbf{1}\}$.

Bayesian Classification

Let Ω be a sample space (a universum) of all objects that can be classified. We assume a probability P on Ω.

We consider the problem of **binary classification**:

- Let Y be the random variable for the category which takes values in $\{\mathbf{0}, \mathbf{1}\}$.
- Let X be the random vector describing n features of a given instance, i.e., $X = (X_1, \dots, X_n)$
  - Denote by $\vec{x} \in \mathbb{R}^n$ values of X,
  - and by $x_i \in \mathbb{R}$ values of X_i.

Bayesian Classification

Let Ω be a sample space (a universum) of all objects that can be classified. We assume a probability P on Ω.

We consider the problem of **binary classification**:

- Let Y be the random variable for the category which takes values in $\{\mathbf{0}, \mathbf{1}\}$.
- Let X be the random vector describing n features of a given instance, i.e., $X = (X_1, \ldots, X_n)$
  - Denote by $\vec{x} \in \mathbb{R}^n$ values of X,
  - and by $x_i \in \mathbb{R}$ values of X_i.

Bayes classifier: Given a vector of feature values $\vec{x}$,

$$C^{Bayes}(\vec{x}) := \begin{cases} \mathbf{1} & \text{if } P(Y = \mathbf{1} \mid X = \vec{x}) \geq P(Y = \mathbf{0} \mid X = \vec{x}) \\ \mathbf{0} & \text{otherwise.} \end{cases}$$

Intuitively, C^{Bayes} assigns to $\vec{x}$ the most probable category it might be in.

Bayesian Classification – Example

Imagine a conveyor belt with apples and apricots.

A machine is supposed to correctly distinguish apples from apricots based on their weight and diameter.

Bayesian Classification – Example

Imagine a conveyor belt with apples and apricots.

A machine is supposed to correctly distinguish apples from apricots based on their weight and diameter.

That is,

- $Y \in \{\mathbf{1}, \mathbf{0}\}$
  (here our interpretation is $\mathbf{1}$ = apple, $\mathbf{0}$ = appricot)
- $X = (X_{weight}, X_{diam})$

Bayesian Classification – Example

Imagine a conveyor belt with apples and apricots.

A machine is supposed to correctly distinguish apples from apricots based on their weight and diameter.

That is,

- $Y \in \{\mathbf{1}, \mathbf{0}\}$
  (here our interpretation is $\mathbf{1}$ = apple, $\mathbf{0}$ = appricot)
- $X = (X_{weight}, X_{diam})$

We are given a fruit of a diameter $5cm$ that weighs $40g$.

Bayesian Classification – Example

Imagine a conveyor belt with apples and apricots.

A machine is supposed to correctly distinguish apples from apricots based on their weight and diameter.

That is,

- $Y \in \{\mathbf{1}, \mathbf{0}\}$
  (here our interpretation is $\mathbf{1}$ = apple, $\mathbf{0}$ = appricot)
- $X = (X_{weight}, X_{diam})$

We are given a fruit of a diameter $5cm$ that weighs $40g$.

The Bayes classifier compares $P(Y = \mathbf{1} \mid X = (40g, 5cm))$ with $P(Y = \mathbf{0} \mid X = (40g, 5cm))$ and selects the more probable category given the features.

Crucial question: Is such a classifier good?

There are other classifiers, e.g., one which compares the weight divided by 10 with the diameter and decides based on the answer, or maybe a classifier which sums the weight and the diameter and compares the result with a constant, etc.

Bayes Classifier

Let C be an arbitrary *classifier*, that is a function that to every feature vector $\vec{x} \in \mathbb{R}^n$ assigns a class from $\{\mathbf{0}, \mathbf{1}\}$.

Bayes Classifier

Let C be an arbitrary *classifier*, that is a function that to every feature vector $\vec{x} \in \mathbb{R}^n$ assigns a class from $\{\mathbf{0}, \mathbf{1}\}$.

Define the error of the classifier C by

$$E_C = P(Y \neq C)$$

(Here we slightly abuse notation and apply C to samples, technically we apply the composition $C \circ X$ of C and X which first determines the features using X and then classifies according to C).

Bayes Classifier

Let C be an arbitrary *classifier*, that is a function that to every feature vector $\vec{x} \in \mathbb{R}^n$ assigns a class from $\{\mathbf{0}, \mathbf{1}\}$.

Define the error of the classifier C by

$$E_C = P(Y \neq C)$$

(Here we slightly abuse notation and apply C to samples, technically we apply the composition $C \circ X$ of C and X which first determines the features using X and then classifies according to C).

Theorem

The Bayes classifier C^{Bayes} minimizes E_C, that is

$$E_{C^{Bayes}} = \min_{C \text{ is a classifier}} E_C$$

Practical Use of Bayes Classifier

The crucial problem: The probability P is not known!
In particular, where to get $P(Y = \mathbf{1} \mid X = \vec{x})$?
Note that $P(Y = \mathbf{0} \mid X = \vec{x}) = 1 - P(Y = \mathbf{1} \mid X = \vec{x})$

Practical Use of Bayes Classifier

The crucial problem: The probability P is not known!
In particular, where to get $P(Y = \mathbf{1} \mid X = \vec{x})$?
Note that $P(Y = \mathbf{0} \mid X = \vec{x}) = 1 - P(Y = \mathbf{1} \mid X = \vec{x})$

Given no other assumptions, this requires a table giving
the probability of the category $\mathbf{1}$ for each possible feature vector $\vec{x}$.

Where to get these probabilities?

Practical Use of Bayes Classifier

The crucial problem: The probability P is not known!
In particular, where to get $P(Y = \mathbf{1} \mid X = \vec{x})$?
Note that $P(Y = \mathbf{0} \mid X = \vec{x}) = 1 - P(Y = \mathbf{1} \mid X = \vec{x})$

Given no other assumptions, this requires a table giving the probability of the category $\mathbf{1}$ for each possible feature vector $\vec{x}$.

Where to get these probabilities?

In some cases the probabilities might come from the knowledge of the solved problem (e.g., applications in physics might be supported by a theory giving the probabilities).

In most cases, however, P is estimated from sampled data by

$$\bar{P}(Y = \mathbf{1} \mid X = \vec{x}) = \frac{\text{number of samples with } Y = \mathbf{1} \text{ and } X = \vec{x}}{\text{number of samples with } X = \vec{x}}$$

(We use $\bar{P}$ to denote an estimate of P from data.)

Estimating P

Consider a problem with $X = (X_1, X_2, X_3)$ where each X_i returns either 0 or 1. What the data might look like?

Estimating P

Consider a problem with $X = (X_1, X_2, X_3)$ where each X_i returns either 0 or 1. What the data might look like?

Part of the data table:

Y	X_1	X_2	X_3
1	1	0	1
1	0	1	1
0	1	0	1
0	0	0	1
1	0	0	0
0	1	1	1
...			

All data with $X_1 = 1$, $X_2 = 0$, $X_3 = 1$:

Y	X_1	X_2	X_3
1	1	0	1
1	1	0	1
0	1	0	1
0	1	0	1
1	1	0	1
1	1	0	1

Estimate: $\bar{P}(\mathbf{1} \mid 1, 0, 1) = 2/3$

Estimating P

Consider a problem with $X = (X_1, X_2, X_3)$ where each X_i returns either 0 or 1. What the data might look like?

Part of the data table:

Y	X_1	X_2	X_3
1	1	0	1
1	0	1	1
0	1	0	1
0	0	0	1
1	0	0	0
0	1	1	1
...			

All data with $X_1 = 1$, $X_2 = 0$, $X_3 = 1$:

Y	X_1	X_2	X_3
1	1	0	1
1	1	0	1
0	1	0	1
0	1	0	1
1	1	0	1
1	1	0	1

Estimate: $\bar{P}(\mathbf{1} \mid 1, 0, 1) = 2/3$

The probability table and hence also the necessary data are typically too large!

Concretely, if all $X_1, \ldots, X_n$ are binary, there are 2^n probabilities $P(Y = \mathbf{1} \mid X = \vec{x})$, one for each possible $\vec{x} \in \{0, 1\}^n$.

Let's Look at It the Other Way Round

Theorem (Bayes,1764)

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Let's Look at It the Other Way Round

Theorem (Bayes,1764)

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

Proof.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{P(A \cap B)}{P(A)} \cdot P(A)}{P(B)} = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

$\square$

Bayesian Classification

Determine the category for $\vec{x}$ by computing

$$P(Y = y \mid X = \vec{x}) = \frac{P(Y = y) \cdot P(X = \vec{x} \mid Y = y)}{P(X = \vec{x})}$$

for both $y \in \{\mathbf{0}, \mathbf{1}\}$ and deciding whether or not the following holds:

$$P(Y = \mathbf{1} \mid X = \vec{x}) \geq P(Y = \mathbf{0} \mid X = \vec{x})$$

Bayesian Classification

Determine the category for $\vec{x}$ by computing

$$P(Y = y \mid X = \vec{x}) = \frac{P(Y = y) \cdot P(X = \vec{x} \mid Y = y)}{P(X = \vec{x})}$$

for both $y \in \{\mathbf{0}, \mathbf{1}\}$ and deciding whether or not the following holds:

$$P(Y = \mathbf{1} \mid X = \vec{x}) \geq P(Y = \mathbf{0} \mid X = \vec{x})$$

So in order to make the classifier we need to compute:

- The prior $P(Y = \mathbf{1})$ (then $P(Y = \mathbf{0}) = 1 - P(Y = \mathbf{1})$)
- The conditionals $P(X = \vec{x} \mid Y = y)$ for $y \in \{\mathbf{0}, \mathbf{1}\}$ and for every $\vec{x}$

Estimating the Prior and Conditionals

- $P(Y = \mathbf{1})$ can be easily estimated from data by

$$\bar{P}(Y = \mathbf{1}) = \frac{\text{number of samples with } Y = \mathbf{1}}{\text{number of all samples}}$$

Estimating the Prior and Conditionals

- $P(Y = \mathbf{1})$ can be easily estimated from data by

$$\bar{P}(Y = \mathbf{1}) = \frac{\text{number of samples with } Y = \mathbf{1}}{\text{number of all samples}}$$

- If the dimension of features is small, $P(X = \vec{x} \mid Y = y)$ can be estimated from data similarly as $P(Y = \mathbf{1} \mid X = \vec{x})$ by

$$\bar{P}(X = \vec{x} \mid Y = y) = \frac{\text{number of samples with } Y = y \text{ and } X = \vec{x}}{\text{number of samples with } Y = y}$$

Unfortunately, for higher dimensional data too many samples are needed to estimate all $P(X = \vec{x} \mid Y = y)$ (there are too many $\vec{x}$'s).

So where is the advantage of using the Bayes thm.??

Estimating the Prior and Conditionals

- $P(Y = \mathbf{1})$ can be easily estimated from data by

$$\bar{P}(Y = \mathbf{1}) = \frac{\text{number of samples with } Y = \mathbf{1}}{\text{number of all samples}}$$

- If the dimension of features is small, $P(X = \vec{x} \mid Y = y)$ can be estimated from data similarly as $P(Y = \mathbf{1} \mid X = \vec{x})$ by

$$\bar{P}(X = \vec{x} \mid Y = y) = \frac{\text{number of samples with } Y = y \text{ and } X = \vec{x}}{\text{number of samples with } Y = y}$$

Unfortunately, for higher dimensional data too many samples are needed to estimate all $P(X = \vec{x} \mid Y = y)$ (there are too many $\vec{x}$'s).

So where is the advantage of using the Bayes thm.??

We introduce *independence assumptions* about the features!

Naive Bayes

- We assume that features are (conditionally) independent *given the category*. That is for all $\vec{x} = (x_1, \ldots, x_n)$ and $y \in \{\mathbf{0}, \mathbf{1}\}$ we **assume**:

$$
\begin{aligned}
P(X = x \mid Y = y) &= P(X_1 = x_1, \cdots, X_n = x_n \mid Y) \\
&= \prod_{i=1}^{n} P(X_i = x_i \mid Y = y)
\end{aligned}
$$

Naive Bayes

- We assume that features are (conditionally) independent *given the category*. That is for all $\vec{x} = (x_1, \dots, x_n)$ and $y \in \{\mathbf{0}, \mathbf{1}\}$ we **assume**:

$$
\begin{aligned}
P(X = x \mid Y = y) &= P(X_1 = x_1, \cdots, X_n = x_n \mid Y) \\
&= \prod_{i=1}^{n} P(X_i = x_i \mid Y = y)
\end{aligned}
$$

- Therefore, we only need to specify $P(X_i = x_i \mid Y = y)$ for each possible pair of a feature-value x_i and $y \in \{\mathbf{0}, \mathbf{1}\}$.

Naive Bayes

- We assume that features are (conditionally) independent *given the category*. That is for all $\vec{x} = (x_1, \ldots, x_n)$ and $y \in \{\mathbf{0}, \mathbf{1}\}$ we **assume**:

$$P(X = x \mid Y = y) = P(X_1 = x_1, \cdots, X_n = x_n \mid Y)$$
$$= \prod_{i=1}^{n} P(X_i = x_i \mid Y = y)$$

- Therefore, we only need to specify $P(X_i = x_i \mid Y = y)$ for each possible pair of a feature-value x_i and $y \in \{\mathbf{0}, \mathbf{1}\}$.

  Note that if all X_i are binary (values in $\{0, 1\}$), this requires specifying only $2n$ parameters:

$$P(X_i = 1 \mid Y = \mathbf{1}) \text{ and } P(X_i = 1 \mid Y = \mathbf{0}) \text{ for each } X_i$$

  as $P(X_i = 0 \mid Y = y) = 1 - P(X_i = 1 \mid Y = y)$ for $y \in \{\mathbf{0}, \mathbf{1}\}$.

Compared to specifying 2^n parameters without any independence assumption.

Estimating the marginal probabilities

Estimate the probabilities $P(X_i = x_i \mid Y = y)$ by

$$\bar{P}(X_i = x_i \mid Y = y) = \frac{\text{number of samples with } X_i = x_i \text{ and } Y = y}{\text{number of samples with } Y = y}$$

Estimating the marginal probabilities

Estimate the probabilities $P(X_i = x_i \mid Y = y)$ by

$$\bar{P}(X_i = x_i \mid Y = y) = \frac{\text{number of samples with } X_i = x_i \text{ and } Y = y}{\text{number of samples with } Y = y}$$

Example: Consider a problem with $X = (X_1, X_2, X_3)$ where each X_i returns either 0 or 1. The data is

Y	X_1	X_2	X_3
1	1	0	1
1	0	1	1
0	1	0	1
0	0	0	1
1	0	0	0
0	1	1	1

$\bar{P}(X_1 = 1 \mid Y = \mathbf{1}) = 1/3$ $\quad$ $\bar{P}(X_1 = 1 \mid Y = \mathbf{0}) = 2/3$
$\bar{P}(X_2 = 1 \mid Y = \mathbf{1}) = 1/3$ $\quad$ $\bar{P}(X_2 = 1 \mid Y = \mathbf{0}) = 1/3$
$\bar{P}(X_3 = 1 \mid Y = \mathbf{1}) = 2/3$ $\quad$ $\bar{P}(X_3 = 1 \mid Y = \mathbf{0}) = 1$

Naive Bayes – Example

Consider classification of geometric shapes:

$X_1 \in \{small, medium, large\}$

$X_2 \in \{red, blue, green\}$

$X_3 \in \{square, triangle, circle\}$

Naive Bayes – Example

Consider classification of geometric shapes:

$X_1 \in \{small, medium, large\}$

$X_2 \in \{red, blue, green\}$

$X_3 \in \{square, triangle, circle\}$

We have already estimated the following probabilities:

	$Y = \mathbf{1}$	$Y = \mathbf{0}$
$\bar{P}(Y)$	0.5	0.5
$\bar{P}(small \mid Y)$	0.4	0.4
$\bar{P}(medium \mid Y)$	0.1	0.2
$\bar{P}(large \mid Y)$	0.5	0.4
$\bar{P}(red \mid Y)$	0.9	0.3
$\bar{P}(blue \mid Y)$	0.05	0.3
$\bar{P}(green \mid Y)$	0.05	0.4
$\bar{P}(square \mid Y)$	0.05	0.4
$\bar{P}(triangle \mid Y)$	0.05	0.3
$\bar{P}(circle \mid Y)$	0.9	0.3

Does $(medium, red, circle)$ belong to the category $\mathbf{1}$?

	$Y = \mathbf{1}$	$Y = \mathbf{0}$
$\bar{P}(Y)$	0.5	0.5
$\bar{P}(medium \mid Y)$	0.1	0.2
$\bar{P}(red \mid Y)$	0.9	0.3
$\bar{P}(circle \mid Y)$	0.9	0.3

Denote $\vec{x} = (medium, red, circle)$.

	$Y=\mathbf{1}$	$Y=\mathbf{0}$
$\bar{P}(Y)$	0.5	0.5
$\bar{P}(medium \mid Y)$	0.1	0.2
$\bar{P}(red \mid Y)$	0.9	0.3
$\bar{P}(circle \mid Y)$	0.9	0.3

Denote $\vec{x} = (medium, red, circle)$.

$$
\begin{aligned}
&P(Y=\mathbf{1} \mid X=\vec{x}) = \\
&\qquad = P(\mathbf{1}) \cdot P(medium \mid \mathbf{1}) \cdot P(red \mid \mathbf{1}) \cdot P(circle \mid \mathbf{1}) \,/\, P(X=\vec{x}) \\
&\qquad \doteq 0.5 \cdot 0.1 \cdot 0.9 \cdot 0.9 \,/\, P(X=\vec{x}) = 0.0405/P(X=\vec{x})
\end{aligned}
$$

	$Y = \mathbf{1}$	$Y = \mathbf{0}$
$\bar{P}(Y)$	0.5	0.5
$\bar{P}(medium \mid Y)$	0.1	0.2
$\bar{P}(red \mid Y)$	0.9	0.3
$\bar{P}(circle \mid Y)$	0.9	0.3

Denote $\vec{x} = (medium, red, circle)$.

$$P(Y = \mathbf{1} \mid X = \vec{x}) =$$
$$= P(\mathbf{1}) \cdot P(medium \mid \mathbf{1}) \cdot P(red \mid \mathbf{1}) \cdot P(circle \mid \mathbf{1}) \,/\, P(X = \vec{x})$$
$$\doteq 0.5 \cdot 0.1 \cdot 0.9 \cdot 0.9 \,/\, P(X = \vec{x}) = 0.0405/P(X = \vec{x})$$

$$P(Y = \mathbf{0} \mid X = \vec{x}) =$$
$$= P(\mathbf{0}) \cdot P(medium \mid \mathbf{0}) \cdot P(red \mid \mathbf{0}) \cdot P(circle \mid \mathbf{0}) \,/\, P(X = \vec{x})$$
$$\doteq 0.5 \cdot 0.2 \cdot 0.3 \cdot 0.3 \,/\, P(X = \vec{x}) = 0.009/P(X = \vec{x})$$

(Note that we used the estimates $\bar{P}$ of P to finish the computation above.)

	$Y=\mathbf{1}$	$Y=\mathbf{0}$
$\bar{P}(Y)$	0.5	0.5
$\bar{P}(medium \mid Y)$	0.1	0.2
$\bar{P}(red \mid Y)$	0.9	0.3
$\bar{P}(circle \mid Y)$	0.9	0.3

Denote $\vec{x} = (medium, red, circle)$.

$$
\begin{aligned}
P(Y = \mathbf{1} \mid X = \vec{x}) &= \\
&= P(\mathbf{1}) \cdot P(medium \mid \mathbf{1}) \cdot P(red \mid \mathbf{1}) \cdot P(circle \mid \mathbf{1}) \,/\, P(X = \vec{x}) \\
&\doteq 0.5 \cdot 0.1 \cdot 0.9 \cdot 0.9 \,/\, P(X = \vec{x}) = 0.0405/P(X = \vec{x})
\end{aligned}
$$

$$
\begin{aligned}
P(Y = \mathbf{0} \mid X = \vec{x}) &= \\
&= P(\mathbf{0}) \cdot P(medium \mid \mathbf{0}) \cdot P(red \mid \mathbf{0}) \cdot P(circle \mid \mathbf{0}) \,/\, P(X = \vec{x}) \\
&\doteq 0.5 \cdot 0.2 \cdot 0.3 \cdot 0.3 \,/\, P(X = \vec{x}) = 0.009/P(X = \vec{x})
\end{aligned}
$$

(Note that we used the estimates $\bar{P}$ of P to finish the computation above.)
Apparently,

$$P(Y = \mathbf{1} \mid X = \vec{x}) \doteq 0.0405/P(X = \vec{x}) > 0.009/P(X = \vec{x}) \doteq P(\mathbf{0} \mid X = \vec{x})$$

So we classify $\vec{x}$ to the category $\mathbf{1}$.

Estimating Probabilities in Practice

We already know that $P(X_i = x_i \mid Y = y)$ can be estimated by

$$\bar{P}(X_i = x_i \mid Y = y) = \ell_{y,x_i} \,/\, \ell_y$$

where

- ℓ_{y,x_i} = number of samples with $Y = y$ and $X_i = x_i$
- ℓ_y = number of samples with $Y = y$

Estimating Probabilities in Practice

We already know that $P(X_i = x_i \mid Y = y)$ can be estimated by

$$\bar{P}(X_i = x_i \mid Y = y) = \ell_{y,x_i} \,/\, \ell_y$$

where

- ℓ_{y,x_i} = number of samples with $Y = y$ and $X_i = x_i$
- ℓ_y = number of samples with $Y = y$

Problem: If, by chance, a rare value x_i of a feature X_i never occurs in the training data, we get

$$\bar{P}(X_i = x_i \mid Y = y) = 0 \quad \text{for both } y \in \{\mathbf{0}, \mathbf{1}\}$$

But then $\bar{P}(X = x) = 0$ for x containing the value x_i for X_i, and thus $\bar{P}(Y = y \mid X = x)$ is not well defined.
Moreover, $\bar{P}(Y = y) \cdot \bar{P}(X = x \mid Y = y) = 0$ (for $y \in \{\mathbf{0}, \mathbf{1}\}$) so even this cannot be used for classification.

Probability Estimation Example

Training data:

Size	Color	Shape	Class
small	red	circle	**1**
large	red	circle	**1**
small	red	triangle	**0**
large	blue	circle	**0**

Estimated probabilities:

	$Y = \mathbf{1}$	$Y = \mathbf{0}$
$\bar{P}(Y)$	0.5	0.5
$\bar{P}(small \mid Y)$	0.5	0.5
$\bar{P}(medium \mid Y)$	0	0
$\bar{P}(large \mid Y)$	0.5	0.5
$\bar{P}(red \mid Y)$	1	0.5
$\bar{P}(blue \mid Y)$	0	0.5
$\bar{P}(green \mid Y)$	0	0
$\bar{P}(square \mid Y)$	0	0
$\bar{P}(triangle \mid Y)$	0	0.5
$\bar{P}(circle \mid Y)$	1	0.5

Note that $\bar{P}(medium \mid \mathbf{1}) = P(medium \mid \mathbf{0}) = 0$ and thus also $\bar{P}(medium, red, circle) = 0$.

So what is $\bar{P}(\mathbf{1} \mid medium, red, circle)$?

Smoothing

- To account for estimation from small samples, probability estimates are adjusted or *smoothed*.

Smoothing

- To account for estimation from small samples, probability estimates are adjusted or *smoothed*.
- *Laplace smoothing* adds one to every count of feature values

$$\tilde{P}(X_i = x_i \mid Y = y) = \frac{\ell_{y,x_i} + 1}{\ell_y + v_i}$$

where

  - ℓ_y = number of training samples with $Y = y$,
  - ℓ_{y,x_i} = number of training samples with $Y = y$ and $X_i = x_i$,
  - v_i is the number of all distinct values of the variable X_i.

To understand note that

$$\ell_y = \sum_{x_i \text{ is a value of } X_i} \ell_{y,x_i}$$

and thus

$$\bar{P}(X_i = x_i \mid Y = y) = \ell_{y,x_i} \,/ \sum_{x_i \text{ is a value of } X_i} \ell_{y,x_i}$$

$$\tilde{P}(X_i = x_i \mid Y = y) = (\ell_{y,x_i} + 1) \,/ \sum_{x_i \text{ is a value of } X_i} (\ell_{y,x_i} + 1)$$

Laplace Smoothing Example

- Assume training set contains 10 samples of category **1**:
  - 4 small
  - 0 medium
  - 6 large

Laplace Smoothing Example

- Assume training set contains 10 samples of category **1**:
  - 4 small
  - 0 medium
  - 6 large
- Estimate parameters as follows
  - $\tilde{P}(small \mid \mathbf{1}) = (4+1)/(10+3) = 0.384$
  - $\tilde{P}(medium \mid \mathbf{1}) = (0+1)/(10+3) = 0.0769$
  - $\tilde{P}(large \mid \mathbf{1}) = (6+1)/(10+3) = 0.538$

Continuous Features

Ω may be (potentially) continuous, X_i may assign a continuum of values in $\mathbb{R}$.

Continuous Features

Ω may be (potentially) continuous, X_i may assign a continuum of values in $\mathbb{R}$.

- The probabilities are computed using *probability density* $p : \mathbb{R} \to \mathbb{R}^+$.
  A random variable $X : \Omega \to \mathbb{R}^+$ has a density $p : \mathbb{R} \to \mathbb{R}^+$ if for every interval $[a, b]$ we have

  $$P(a \leq X \leq b) = \int_a^b p(x)dx$$

  Usually, $P(X_i \mid Y = y)$ is used to denote the *density* of X_i conditioned on $Y = y$.

Continuous Features

Ω may be (potentially) continuous, X_i may assign a continuum of values in $\mathbb{R}$.

- The probabilities are computed using *probability density* $p : \mathbb{R} \to \mathbb{R}^+$.
  A random variable $X : \Omega \to \mathbb{R}^+$ has a density $p : \mathbb{R} \to \mathbb{R}^+$ if for every interval $[a, b]$ we have
  $$P(a \leq X \leq b) = \int_a^b p(x)dx$$
  Usually, $P(X_i \mid Y = y)$ is used to denote the *density* of X_i conditioned on $Y = y$.
- The densities $P(X_i \mid Y = y)$ are usually estimated using Gaussian densities as follows:
  - Estimate the mean μ_{iy} and the standard deviation σ_{iy} based on training data.
  - Then put
  $$\bar{P}(X_i \mid Y = y) = \frac{1}{\sigma_{iy}\sqrt{2\pi}} \exp\left(\frac{-(X_i - \mu_{iy})^2}{2\sigma_{iy}^2}\right)$$

Comments on Naive Bayes

- Tends to work well despite rather strong assumption of conditional independence of features.

Comments on Naive Bayes

- Tends to work well despite rather strong assumption of conditional independence of features.
- Experiments show that it is quite competitive with other classification methods.

  Even if the probabilities are not accurately estimated, it often picks the correct maximum probability category.

Comments on Naive Bayes

- Tends to work well despite rather strong assumption of conditional independence of features.
- Experiments show that it is quite competitive with other classification methods.

  Even if the probabilities are not accurately estimated, it often picks the correct maximum probability category.
- Directly constructs a hypothesis from parameter estimates that are calculated from the training data.

Comments on Naive Bayes

- Tends to work well despite rather strong assumption of conditional independence of features.
- Experiments show that it is quite competitive with other classification methods.

  Even if the probabilities are not accurately estimated, it often picks the correct maximum probability category.
- Directly constructs a hypothesis from parameter estimates that are calculated from the training data.
- Typically handles outliers and noise well.

Comments on Naive Bayes

- Tends to work well despite rather strong assumption of conditional independence of features.
- Experiments show that it is quite competitive with other classification methods.

  Even if the probabilities are not accurately estimated, it often picks the correct maximum probability category.
- Directly constructs a hypothesis from parameter estimates that are calculated from the training data.
- Typically handles outliers and noise well.
- Missing values are easy to deal with (simply average over all missing values in feature vectors).

Bayesian Networks (Basic Information)

In the Naive Bayes we have assumed that *all* features $X_1, \ldots, X_n$ are independent.

Bayesian Networks (Basic Information)

In the Naive Bayes we have assumed that *all* features $X_1, \ldots, X_n$ are independent.

This is usually not realistic.

E.g. Variables "rain" and "grass wet" are (usually) strongly dependent.

Bayesian Networks (Basic Information)

In the Naive Bayes we have assumed that *all* features $X_1, \ldots, X_n$ are independent.

This is usually not realistic.

E.g. Variables "rain" and "grass wet" are (usually) strongly dependent.

What if we return some dependencies back?

(But now in a well-defined sense.)

Bayesian Networks (Basic Information)

In the Naive Bayes we have assumed that *all* features $X_1, \ldots, X_n$ are independent.

This is usually not realistic.

E.g. Variables "rain" and "grass wet" are (usually) strongly dependent.

What if we return some dependencies back?

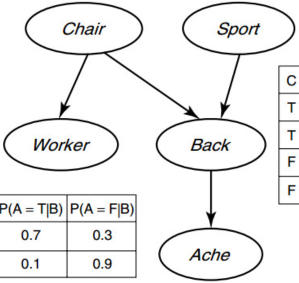
(But now in a well-defined sense.)

Bayesian networks are a graphical model that uses a directed acyclic graph to specify dependencies among variables.

Bayesian Networks – Example

P(C = T)	P(C = F)
0.8	0.2

P(S = T)	P(S = F)
0.02	0.98



C	P(W = T|C)	P(W = F|C)
T	0.9	0.1
F	0.01	0.99

C	S	P(B = T|C,S)	P(B = F|C,S)
T	T	0.9	0.1
T	F	0.2	0.8
F	T	0.9	0.1
F	F	0.01	0.99

B	P(A = T|B)	P(A = F|B)
T	0.7	0.3
F	0.1	0.9

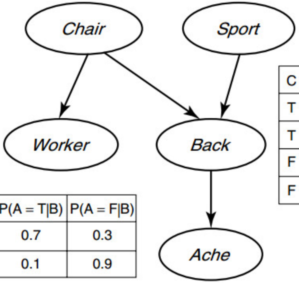
Now, e.g.,

$$P(C, S, W, B, A) = P(C) \cdot P(S) \cdot P(W \mid C) \cdot P(B \mid C, S) \cdot P(A \mid B)$$

Bayesian Networks – Example

P(C = T)	P(C = F)
0.8	0.2

P(S = T)	P(S = F)
0.02	0.98



C	P(W = T|C)	P(W = F|C)
T	0.9	0.1
F	0.01	0.99

C	S	P(B = T|C,S)	P(B = F|C,S)
T	T	0.9	0.1
T	F	0.2	0.8
F	T	0.9	0.1
F	F	0.01	0.99

B	P(A = T|B)	P(A = F|B)
T	0.7	0.3
F	0.1	0.9

Now, e.g.,

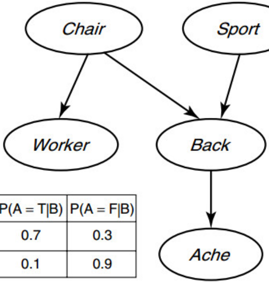
$$P(C, S, W, B, A) = P(C) \cdot P(S) \cdot P(W \mid C) \cdot P(B \mid C, S) \cdot P(A \mid B)$$

Now we may, e.g., infer what is the probability $P(C = T \mid A = T)$ that we sit in a bad chair assuming that our back aches.

Bayesian Networks – Example

P(C = T)	P(C = F)
0.8	0.2

P(S = T)	P(S = F)
0.02	0.98



C	P(W = T|C)	P(W = F|C)
T	0.9	0.1
F	0.01	0.99

C	S	P(B = T|C,S)	P(B = F|C,S)
T	T	0.9	0.1
T	F	0.2	0.8
F	T	0.9	0.1
F	F	0.01	0.99

B	P(A = T|B)	P(A = F|B)
T	0.7	0.3
F	0.1	0.9

Now, e.g.,

$$P(C, S, W, B, A) = P(C) \cdot P(S) \cdot P(W \mid C) \cdot P(B \mid C, S) \cdot P(A \mid B)$$

Now we may, e.g., infer what is the probability $P(C = T \mid A = T)$ that we sit in a bad chair assuming that our back aches.

We have to store only 10 numbers as opposed to $2^5 - 1$ possible probabilities for all vectors of values of C, S, W, B, A.

Bayesian Networks – Learning & Naive Bayes

Many algorithms have been developed for learning:

- the structure of the graph of the network,
- the *conditional probability tables*.

The methods are based on maximum-likelihood estimation, gradient descent, etc.

Automatic procedures are usually combined with expert knowledge.

Bayesian Networks – Learning & Naive Bayes

Many algorithms have been developed for learning:

- the structure of the graph of the network,
- the *conditional probability tables*.

The methods are based on maximum-likelihood estimation, gradient descent, etc.

Automatic procedures are usually combined with expert knowledge.

Can you express the naive Bayes for $Y, X_1, \ldots, X_n$ using a Bayesian network?