

IV122 Zadání: Analýza dat

Jako vstupy pro obě úlohy použijte poskytnutá data i vlastní simulovaná data, tj. dílčí podúkol je napsat generátor simulovaných dat.

A) Lineární regrese

Vstup: seznam bodů x_i, y_i .

Výstup: přímka $ax + b$ minimalizující sumu čtverců chyb; znázorněno obrázkem (body i přímka).

Algoritmy: výpočet vzorcem, aproximativní řešení (grid search nebo gradient descent).

U simulovaných dat prozkoumejte vztah mezi pravými hodnotami a, b (těmi, které byly použity pro generování) a vypočtenými hodnotami. Jak tento vztah závisí na množství dat a na velikosti šumu? Jak robustní je výpočet při přítomnosti odlehlých bodů (body s výrazně vyšší odchylkou než ostatní body)?

B) Detekce shluků

Vstup: seznam bodů x_i, y_i ; žádaný počet shluků k .

Výstup: rozdělení na k shluků; znázorněno obrázkem (obarvení bodů podle shluků).

Algoritmus: k-means.

Prozkoumejte chování algoritmu (stabilita, závislost na iniciálních podmínkách, počet iterací potřebných pro konvergenci) pro různé typy vstupních dat, konkrétně například pro různé počty shluků a míra „překrývání“ shluků.