Paskaita 5

K-artimiausių kaimynų metodas (K-nearest neighbors)

Šis metodas tinka tiek klasifikavimo, tiek regresijos skaičiavimo uždaviniams

Pirmas klausimas - koks modelis (parametrinis ar neparametrinis) yra naudojamas atlikti sprendimo priėmimą?

Atsakymas: labai naudingas, kai sekimasi sumažinti modelio apmokymo etapą (~~to~~ - apmokymo atlikti nereikia (tinginių algoritmas), o modelis yra visa ~~testavimui~~ apmokymui skirta duomenų aibė!

Lazda turi dvi galas - ar taip?

Taigi modelio sudarymo rezultatu ~~testavimo~~ apmokymui duomenų sutvarkymo, kad būtų galima efektyviai atlikti skaičiavimus (duomenų struktūrą) ir jų pastovų atnaujinimą, bei nereikalingų / netikslių duomenų pašalinimą (pvz. kartu dubliuojančių duomenų šalinimą)

K A K metodo algoritmas

- Tarkime nagrinėjamų duomenų X. Reikia rasti jiems atitinkančią kreipiamąją reikšmę (reprezentuojančią) arba nustatyti, kuriai klasei x priklauso (klasifikavimo uždavinys)

- Sekančiame K atitinkamų testavimo duomenų $(X_1, X_2, \dots, X_k)$.

Pastaba. X - gali būti apibrėžtas p komponentėmis $X = (x_1, x_2, \dots, x_p)$. 2 dimensija p. ?

Tai sudėtingiausia algoritmo dalis (kaimynų paieška) – prisimenkite skaldyk ir valdyk algoritmą greitimų taškų paieškai

- Tada regresijos uždaviniu skaičiuojame K ~~kaimyninių~~ artimiausių taškų reikšmių vidurkį (arba medianą), klasifikacijos uždaviniu randame dažniausiai pasitaikančią klasę (vykdome balsavimą).

Atstumo skaičiavimas

Žinome daug metrikų, kurias galime naudoti skaičiuojant atstumus. Patartina parenkant metriką atlikti analizę („apmokymą"), kai pasirenkame testuojamų taškų aibę, o likusius taškus laikome „apmokymo" taškais. Renkame tą metriką kuri geriausiai prognozuoja testuojamų taškų rezultatus

Prisiminkime populiariausias metrikas:

1. Vektorių Euklidinis atstumas:

$$X_1 = (x_1^1, x_2^1, \dots, x_p^1)$$

$$X_2 = (x_1^2, x_2^2, \dots, x_p^2).$$

$$ED(X_1, X_2) = \|X_1 - X_2\| = \left(\sum_{j=1}^{p} (x_j^1 - x_j^2)^2\right)^{1/2}$$

2. Miesto blokinis atstumas (rekomenduojamas kai tokios skirtingos koordinatės apibūdina skirtingos prigimties parametrus - amžių, svorį, ūgį ir t.t.

$$MBA(X_1, X_2) = \sum_{j=1}^{p} |x_j^1 - x_j^2|.$$

Parametro K (kaimynų skaičiaus) parinkimas

Šį parametrą irgi rekomenduojama parinkti „apmokymo" būdu - perrenkame K reikšmes iš tam tikros aibės, pvz (1, 2, ..., 6).

Pastaba. Jeigu K yra pakankamai didelės, o duomenų aibė (apmokymo imties aibė) irgi yra didelė, tai artimiausio kaimyno radimo užduotis tampa pakankamai sudėtingu (reikalauja didelės apimties skaičiavimų)

- Stochastinė paieška (pasirenkame apmokymo imties aibės poaibį)
- Lygiagretieji algoritmai