Paskaita 3

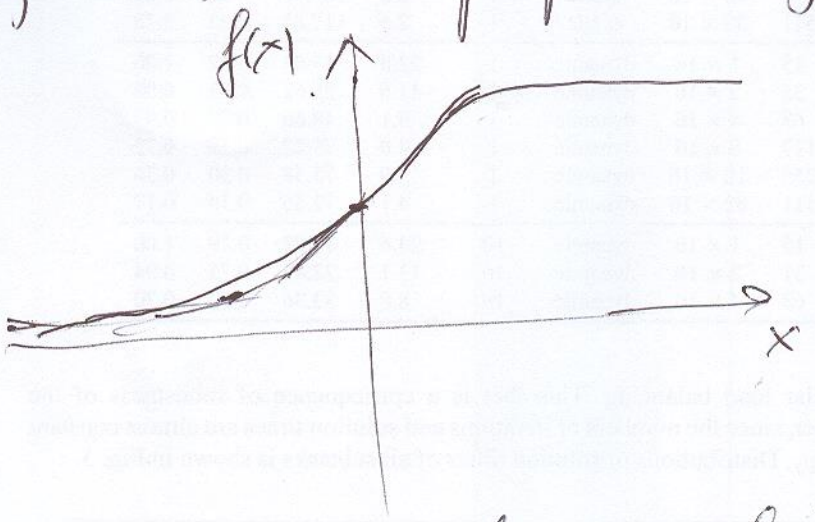
Klasifikavimo uždavinys

Nagrinėsime vieną iš pagrindinių MM uždavinių – duomenų klasifikavimą. Tai mokymosi su mokytoju algoritmų sritis ir naudojami parametriniai metodai tikslo funkcijos $f(x)$ aproksimavimui.

Turime duomenų aibę ir reikia juose gauti atsakymą į klausimą „taip/ne", „raudonas/juodas", (binarieji uždaviniai) ar atpažinti „šuo/katė/vištа/voverė" – m-galimybių klasifikavimo uždavinys.

Priminsime, kad anksčiau nagrinėjome prognozavimo uždavinius, kai $f(x)$ yra tolydi f-ja, įvertinanti pvz. svorį, kainą, temperatūrą ir pan. Susipažinome su tiesinės regresijos metodu

Vietoj tiesės naudojame S-formos kreivę, kuri dažnai natūraliai modeliuojant įvairias populiacijų dinamikas



Statistikoje aproksimuojančiu dažniausiai naudojamas modelis - sigmoidinė (sigmoid) funkcija

$$S(v) = \frac{1}{1 + e^{-v}},$$

v - neapribotas kintamasis reikšmė.

Tada klasifikatoriaus modeliuojame įvykio tikimybę

$$p(x) = \frac{1}{1 + e^{-x}}.$$

Jeigu $p(x) \geq 0.5$ - tai klasifikatorius patvirtina požymį

Logistinės regresijos aproksimacija

$$\hat{p}(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

Parametrai β_0, β_1 apibrėžia tiesinę priklausomybę, bet ji transformuojama panaudojant netiesinę funkciją

Taigi prognozės rezultatas įvairiai priklausomai nuo x.

Logistinės regresijos skaičiavimai

$$p(x) = \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}}$$

$$1 - p(x) = 1 - \frac{e^{\beta_0 + \beta_1 x}}{1 + e^{\beta_0 + \beta_1 x}} = \frac{1}{1 + e^{\beta_0 + \beta_1 x}}$$

$$\frac{p(x)}{1 - p(x)} = e^{\beta_0 + \beta_1 x}$$

$$\ln \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x$$

$\frac{p(x)}{1-p(x)}$ → šansas / odds

↑ atitinka tiesinės regresijos formulę.

$p = 0.8$

$$\text{odds} = \frac{0.8}{1 - 0.8} = 4.$$

Koef. β_0, β_1 surandame panaudojami išreikštines formules (iš eksp. duomenų)

Norėdami iš eksperimentinių duomenų surasti logistinę regresijos funkciją naudosime stochastinių gradientinių nuokrypių metodą.

Kaip pavyzdį turime tokius eksperimentinius duomenis

$$(x_{1i}, x_{2i}, y_i) \quad i = 1, \dots, m$$

$$\text{čia } y_i = \begin{cases} 0, \\ 1, \end{cases} \text{ binarine klasifikacija}$$

Aproksimacija:

$$y = \frac{1}{1 + \exp(-(\beta_0 + \beta_1 x_1 + \beta_2 x_2))}$$

Jeigu $y(x_1, x_2) < 0.5$, tai taškas priklauso 0

Jeigu $y(x_1, x_2) > 0.5$, tai taškas priklauso 1.

Imkime paklaidos f-ją

$$F(\beta_0, \beta_1, \beta_2; x_1, x_2) = \left(y - \frac{1}{1+\exp(-\beta_0 + \beta_1 x_1 + \beta_2 x_2)} \right)^2$$

Radę skaičiuosim gradiento komponentes

$$\frac{\partial F}{\partial \beta_0} = 2\left(y - \frac{1}{1+\exp(\ )}\right)(-1) \cdot$$

$$\cdot \frac{\partial}{\partial \beta_0}\left[1 + \exp(-\beta_0 + \beta_1 x_1 + \beta_2 x_2)\right]^{-1} \cdot$$

$$= -2\left(y - \underbrace{\frac{1}{1+\exp(\ )}}_{p(\beta_0, \beta_1, \beta_2; x)}\right) \underbrace{\left(1+\exp(\ )\right)^{-2} \cdot \exp(\ )}_{p \cdot (1-p)}$$

$$\boxed{\frac{\partial F}{\partial \beta_0} = -2\left(y - p(\beta_0, \beta_1, \beta_2; x)\right) \cdot p(\ ) \cdot (1 - p(\ ))}$$

Analogiškai apskaičiuojame

$$\frac{\partial F}{\partial \beta_k} = -2\,(y - p(t))\cdot p(t)\cdot (1-p(t))\cdot x_k,$$

$$k = 1, 2.$$

Gauname iteracinį algoritmą

$$\beta_0^{n+1} = \beta_0^n - \gamma \cdot \frac{\partial F}{\partial \beta_0}\left(\beta_0^n, \beta_1^n, \beta_2^n; x_1^n, x_2^n\right)$$

$$\beta_k^{n+1} = \beta_k^n - \gamma \frac{\partial F}{\partial \beta_k}\left(\beta_0^n, \beta_1^n \beta_2^n; x_1^n, x_2^n\right)\cdot x_k^n,$$

$$k = 1, 2.$$

Tiesinė Diskriminantinė Analizė
(Linear Discriminant Analysis).

Anksčiau metodai leido klasifikuoti modelius, kai turėjome tik du požymius (klases).

LDA klasifikatorius tinka ir tada, kai turime daugiau nei 2 klases.

LDA modelių atvaizdavimas yra grafinės. Jis priklauso nuo turimų duomenų statistinių savybių.

Jeigu turime tik vieną įvesties parametrą (x), kuris ~~duomenis~~ reikia klasifikuoti, tai kiekvienam klasei apibrėžiame du parametrus (charakteristikas)

- vidurkį
- dispersiją.

Jeigu turime daugiau įvesties kintamųjų, tai skaičiuojame

- vidurkį
- kovariacijų matricą (tai yra daugiamatės dispersijos apibendrinimas).

Turėdami šią statistinę informaciją, ją įstatome į LDA lygtį ir prognozuojame, kuriai klasei priklauso (x) – atliekame klasifikavimą.

Sprendžiant klasifikavimo uždavinį LDA metodu daroma prielaida:

- visi kintamieji (x) yra pasiskirstę pagal normalųjį pasiskirstymo dėsnį
- visų kintamųjų (x) (jei jis yra keli) variacijos (dispersijos) yra vienodos.

Todėl patartina duomenis patikrinti ir suformuluoti dar iki LDA algoritmo panaudojimo.

Algoritmas:

1. ~~Skaičiuoja~~ Apmokymo duomenimis skaičiuojame koeficientus klasei k duomenų vektoriui:

$$a_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i, \quad x_i - \text{priklausos klasei } k.$$

2. Skaičiuojame visos imties dispersiją (remiamės prielaida, kad visuose klasėse duomenų pasiskirstymas yra vienodas)

$$\sigma^2 = \frac{1}{n-K} \sum_{i=1}^{n} (x_i - a_k)^2,$$

čia K yra skirtingų klasių skaičius, iš statistikos kurso žinome tokią nepaslinktą dispersijos vertinimo formulę

Dalej zwiększamy liczbę, kod nowej obserwacji (x) przypisano klasie k. Aproksymujemy liczbę + klasie, a pozostałe tą, której liczba obserwacji jest największa.

Naudosime Bayes teoremą:

$$P(Y=k \mid X=x) = \frac{P(k) \cdot P(x \mid k)}{\sum_{j=1}^{K} P(j)\, P(x \mid j)}$$

- $P(Y=k \mid X=x)$ yra tikimybė, kad kintamojo X reikšmė x atitinka klasę k.
- $P(k)$ atitinka tikimybę kad aproksimuojamo skaičiaus obserwacijos k-klasei tikimybė (dažnis) bus $P(k)$.

- $P(x|k)$ yra tikimybė, kad x priklauso klasei k. (eksperimentinė)

Jo aproksimuojame remdamiesi normaliojo dėsnio ir prielaida, kad (x) yra pasiskirstę pagal normalųjį dėsnį.

Gauname diskriminantinę funkciją

$$D_k(x) = x \frac{a_k}{\sigma^2} - \frac{a_k^2}{2\sigma^2} + \ln(P(k)).$$

Dabar jau galime atlikti klasifikavimo sprendimą (žr. Lab3.txt duomenų failą).