Pratybos (Lab.) 1

Tiesinės regresijos metodas

Duomenų aibė (x_i, y_i), $i = 1, \dots, m$.

Ieškome nežinomo atvaizdavimo

$$y = f(x)$$

Pasirenkame parametrinį metodą,

ir $f(x)$ ieškome tarp tiesių

$$f^*(x, \beta) = \beta_0 + \beta_1 x.$$

Parametrus β_0 ir β_1 randame sprendami variacinį (minimizav.) uždav.

$$\arg \min_{(\beta_0, \beta_1)} \sum_{i=1}^{m} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

Formulės, koeficientų skaičiav.

$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i, \quad \bar{y} = \frac{1}{m}\sum_{i=1}^{m} y_i$$

$$\boxed{\hat{\beta}_1 = \frac{\sum_{i=1}^{m}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{m}(x_i-\bar{x})^2} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}}$$

Duomenys

$(1,1), (2,3), (4,3), (3,2), (5,5)$

$m = 5.$

Ats.: $\boxed{\beta_0 = 0.4, \beta_1 = 0.8}$, $\boxed{RMSE = 0.69282}$

1) Sudarykite tiesinės regresijos modelį
2) Atvaizduokite informaciją grafike
3) Apskaičiuokite prognozę, kai

   $x = 1, 1.5, 2, 3, 4, 4.5, 5.$

4) Įvertinkite prognozės palyginę su apmokymo duomenimis. RMSE (root mean square error)

Gradientinis nusileidimo metodas

Turime dviejų kintamųjų f-ją

$F(\beta_0, \beta_1)$

Norime rasti minimumo reikšmę (lokalaus minimumo)

Surandame f-jos gradientą (tai vektorius, kurio kryptimi funkcija didėja greičiausiai)

$$\operatorname{grad} F = \left(\frac{\partial F}{\partial \beta_0}, \frac{\partial F}{\partial \beta_1}\right)$$

taško $\bar{\beta} = (\bar{\beta}_0, \bar{\beta}_1)$ aplinkoje.

Tada eidami antigradiento kryptimi mažiname f-jos reikšmę

$$\boxed{\beta^{n+1} = \beta^{n} - \gamma^{n} \operatorname{grad} F(\beta^{n})}$$

γ^n - pakankamai mažas parametras

Pritaikykime šį metodą tiesinės regresijos metodikai (tiesinės regresijos aproksimacijos skaičiavimui)

$$F(\beta_0, \beta_1) = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2$$

$$\frac{\partial F}{\partial \beta_0} = -2 \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i)\right]$$

$$\frac{\partial F}{\partial \beta_1} = -2 \sum_{i=1}^{n} \left[y_i - (\beta_0 + \beta_1 x_i)\right] x_i$$

Apskaičiuojame $F(\beta^{n+1})$:

- Jeigu $F(\beta^{n+1}) < F(\beta^n)$ tęsiame iteracinį procesą
- Jeigu $F(\beta^{n+1}) \geq F(\beta^n)$, tai sumažiname $\gamma := \frac{1}{2}\gamma$ ir kartojame β^{n+1} skaičiavimus.

Uždavinys

Išspręskite ankstesnį pavyzdį (suraskite tiesinės regresijos lygtį.) Pradinį artinį imkite

$$\beta_0^0 = 0, \quad \beta_1^0 = 0.$$

Iteracijas tęskite, kol bus išpildyta būtinoji minimumo sąlyga norimu tikslumu.

Pastaba: Jeigu n (duomenų skaičius) yra didelis (o mus domina BIG DATA uždav.), ~~tai~~ regresijos tiesės radimo algoritme kiekvienoje iteracijoje reikia perskaičiuoti gradiento vektorių ir apskaičiuoti $F(\beta^{n+1})$ reikšmes, kol surasime tikslumą jo reikšmę. Tai gali reikalauti pakankamai didelio apimties skaičiavimų

Stochastinis gradientinio nusileidimo metodas (stochastic gradient descent)

Koeficientus β_0, β_1 perskaičiuojame imdami po vieną ~~element~~ duomenų porą (x_i, y_i). Tokiu būdu smarkiai sumažiname vienos iteracijos realizavimo kaštus.

$$\beta^{n+1,i} = \beta^{n+1,i-1} - \gamma^i \operatorname{grad} F^i(\beta^{n+1,i-1})$$

$$\beta^{n+1,0} = \beta^n, \quad \beta^{n+1} = \beta^{n+1,m}$$

$$\frac{\partial F^i}{\partial \beta_0}(\beta) = -2\left(y_i - (\beta_0 + \beta_1 x)\right) \cdot 1.$$

$$\frac{\partial F^i}{\partial \beta_1}(\beta) = -2\left(y_i - (\beta_0 + \beta_1 x)\right) x.$$

Terminai: kai užbaigiame atnaujinti β^i reikšmes visiems $i=1,\dots,m$, sakome, kad įvykdėme vienos epochos skaičiavimus

Užduotis: Imkite ankstesnio pavyzdžio duomenis ir skaičiuokite tiesinės regresijos $f=j\beta$ (tiesę) stochastiniu gradientiniu nusileidimo metodu.

$\beta_0^0 = 0, \; \beta_1^0 = 0.$

Iteracinio parametro γ reikšmę imkite $\gamma = 0.01$

Apskaičiuokite 4 epochas (20 iteracijų)