Paskaita 2

Įvadas apie tiesinės regresijos metodą

Šioje paskaitoje nagrinėsime gerai žinomą iš statistikos kurso tiesinės regresijos metodą, bet pagrindinį dėmesį skirsime tiems naujiems klausimams, kuriuos nagrinėjame MM metode. Mus domins tokie ~~klausimai~~ punktai:

a) duomenų skaidymo strategija
b) apmokymo tikslumo įvertinimas
c) statistikos metodai ir algoritmai, kurie esminiai naudojami MM technologijoje.

1. Nagrinėkime klasikinę (standartinę) situaciją, kai efektyviai naudojame MM.

Turime m eksperimentinių duomenų

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

kurie gauti „matuojant" funkcijos $f(x)$ reikšmes, $x \in X$ – gra nepriklausomo kintamojo reikšmių aibė (jos gra deterministinės),

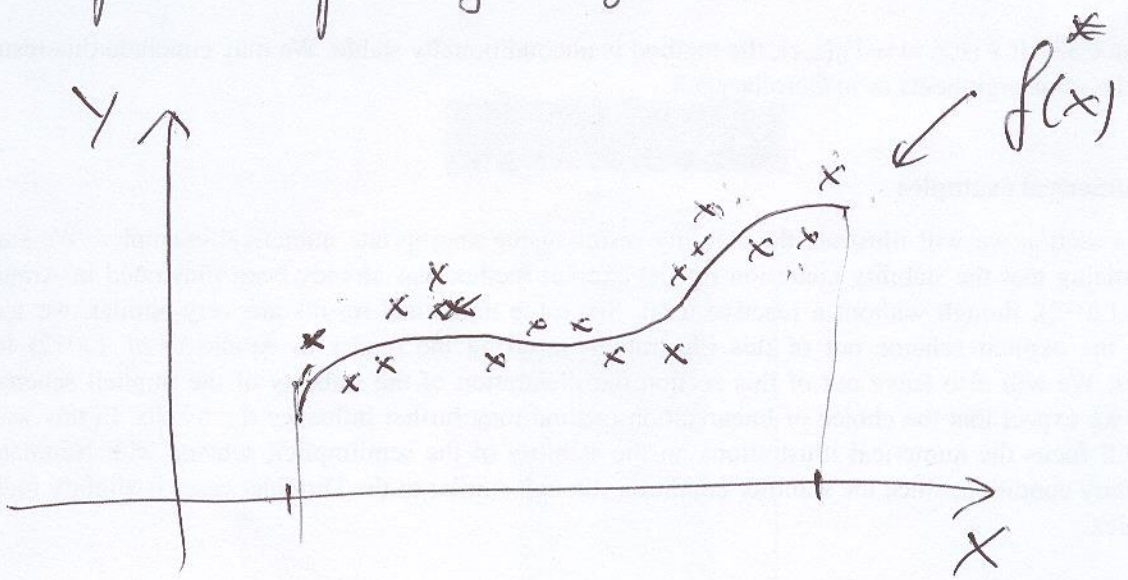
$y \in Y$ gra priklausomo kintamojo reikšmės (jos gra atsitiktinės, paveiktos „triukšmo").

Tikslas gra rasti ~~R~~ funkciją $f(x)$, kuri geriausiai nusakytų ryšį tarp ~~duomenų~~ eksperimentinių duomenų

$$Y = f(X) + \varepsilon$$

ε gra matavimų paklaidų, apie jos tikimybinį pasiskirstymo tankį tam tikrą informaciją (matavimo prietaisų charakteristikos).

Tipiški pavyzdys:



Priklausos apie paklaidas

1. ε_i yra normalusis atsitiktinis dydis
2. Paklaidų vidurkis lygus nuliui $E\varepsilon_i = 0$
3. ~~Dispersija, yra~~ Paklaidų dispersija yra lygi $D\varepsilon_i = \sigma^2$.
4. Visos atsitiktinės paklaidos yra nepriklausomos

Tolesnėje šios paskaitos medžiagoje prisiminsime paprastosios tiesinės regresijos analizę.

Darome prielaidą (hipotezę), kad eksperimentiniai duomenys gali būti aprašomi tokiu modeliu:

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i.$$

Čia α_0 ir α_1 yra parametrai kurių reikšmės yra nežinomos (jas siekiame įvertinti pasirinktu būdu).

Dėl to, kad ε_i yra atsitiktinis dydis, tai net ir tuomet kai pačių x_i reikšmių keli kartus: $x_i = x_j = \tilde{x}$, y_i ir y_j reikšmės gali būti skirtingos dėl matavimo paklaidos, t.y. $\varepsilon_i \neq \varepsilon_j$. Parametrų α_0, α_1 radimui formuluojame optimizavimo uždavinį

$$\arg\min_{\alpha_0, \alpha_1} \underbrace{\sum_{i=1}^{m} \left(y_i - (\alpha_0 + \alpha_1 x_i) \right)^2}_{J(\alpha_0, \alpha_1)} \quad (1)$$

$\overline{y_i}$

(1) uždavinio sprendinį (parametrus $\hat{\alpha}_0, \hat{\alpha}_1$) randame išreikštiniu būdu (2×2 - tiesinės lygčių sistemos sprendinį)

$$\frac{\partial J}{\partial \alpha_0} = -2\sum_{i=1}^{m}(y_i - \alpha_0 - \alpha_1 x_i) = 0$$

$$\frac{\partial J}{\partial \alpha_1} = -2\sum_{i=1}^{m}(y_i - \alpha_0 - \alpha_1 x_i)x_i = 0$$

$$\begin{cases} m\hat{\alpha}_0 + \left(\sum_{i=1}^{m} x_i\right)\hat{\alpha}_1 = \sum_{i=1}^{m} y_i \\ \left(\sum_{i=1}^{m} x_i\right)\hat{\alpha}_0 + \left(\sum_{i=1}^{m} x_i^2\right)\hat{\alpha}_1 = \sum_{i=1}^{m} x_i y_i \end{cases}$$

$$\hat{\alpha}_0 = \tilde{y} - \tilde{x}\,\hat{\alpha}_1,$$

$$\hat{\alpha}_1 = \frac{\sum_{i=1}^{m}(x_i - \tilde{x})(y_i - \tilde{y})}{\sum_{i=1}^{m}(x_i - \tilde{x})^2}$$

Mūsų tikslas yra nurodyti sudarytą funkciją

$$f(x) = \alpha_0 + \alpha_1 x,$$

kuri prognozuotume Y reikšmes naujuose taškuose. Tada svarbiausias teorinis klausimas – kokia bus tokios prognozės paklaida

$$error(x) = |f^*(x) - f(x)|.$$

Kadangi $f^*(x)$ nežinome, tai turime sudaryti apostteriorinius įverčius, kuriuose ~~progno~~ glūdi informacija apie apmokymo metu gautos funkcijos $f(x)$ tikslumą.

Tam naudojame – Duomenų skaidymą.

Duomenų skaidymo strategija 1

Visus turimus eksperimentinius duomenis dalijame į dvi ~~dvi~~ dalis.

1. Apmokymui skirti duomenys

(x_i, y_i), $i = 1, \dots, k$

2. Testavimui skirti duomenys

(x_i, y_i), $i = k+1, \dots, n$.

Tikslinga šias aibes parinkti, remiantis proporcija 70% : 30% arba 60% : 40%

Darbo eiga.

1. Apmokymo procese sudaromė f-jų $f(x)$ (pvz naudojame tiesinę regresiją) $f = f(x, \alpha)$.

2. Įvertinama gautosios f-jos $f(x)$ tikslumo $f = f(x, \alpha)$.

$$\hat{e} = \max_{k+1 \le i \le m} |y_i - f(x_i)|.$$

Jeigu šios tikslumas yra nepakankamas, tai remiamės kitų funkcijos f klasę – pvz. sudarome parabolę $f(x, \alpha) = \alpha_0 + \alpha_1 x + \alpha_2 x^2$ arba splainą

$$S(x, \alpha).$$

Vėl naudojame aproksimuoti skirtus duomenis pirmame etape ir testavimui skirtus duomenis, kai įvertiname parabolę.

Labai svarbu nesupainioti su netriviali (ir šiek tiek netikėta) išvada.

- Pasirinkdami įvairias f(x) aproksimavimo funkcijas galime gauti efektyviai realizuojamus algoritmus (tiek apriorinio stadijoje, tiek ir modeliavimo/paieškos metu.)

Jeigu f(x) yra paprasta (pvz. tiesinė f-ja), tai jos realizacijos kaštai yra maži. Tačiau tokios f-jos tikslumas gali būti nepakankamas, neefektyviai panaudojame turimus eksperimentinius duomenis. Tokią situaciją vadiname - nepakankamu apmokymu.

Apie tai gali signalizuoti ir parinkimo metu

$$\tilde{E} = \max_{1 \le i \le k} |y_i - f(x_i)|.$$

Veikia, tai patvirtina ir tikrinimo tikrinimas naudojant testavimui skirtus duomenis

$$\tilde{\tilde{e}} = \max_{k+1 \le i \le m} |y_i - f(x_i)|$$

(Dažnai $\tilde{\tilde{e}} \approx \tilde{e}$ arba net ir $\tilde{\tilde{e}} > \tilde{e}$)

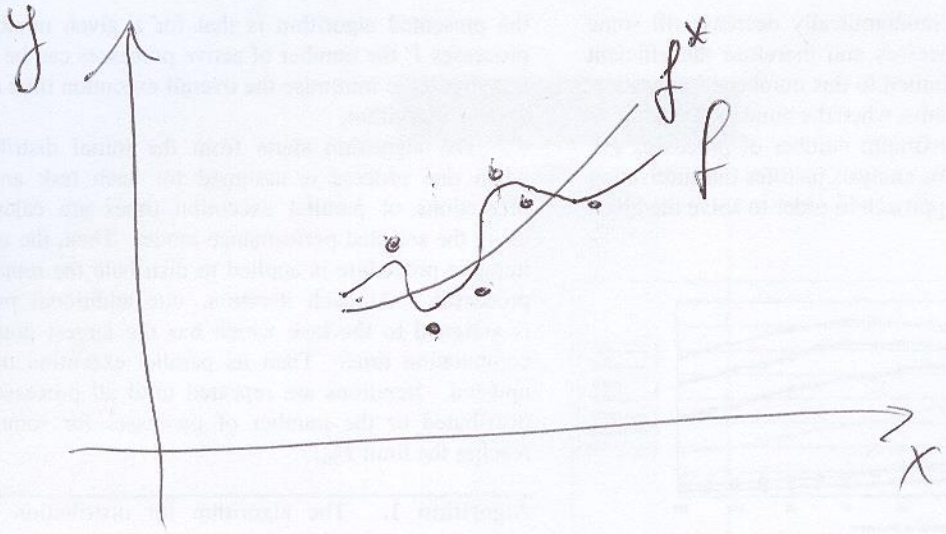
- Pasirinkdami specialios struktūros f-jas (splainai, racionalios trupmenos, Padė aproksimacijos) parinkimo metu suderiname $f(x)$ f-jos, kuri labai gerai aproksimuoja duomenis, ir y_i aproksimuoja skirtis $\tilde{e} << 1$ (t. maža pakanka)

Tačiau atlikę testavimą su antrąja duomenų aibę gaunama, kad paklaida $\tilde{\tilde{E}}$ yra didelė.

Tokią situaciją vadiname persimokymu apmokymo proceso metu.

Tai nesunkiai paaiškinti - apmokymo metu per daug tiksliai aproksimuojame eksperimentinius duomenis ir, kaip matyti, patys yra netikslūs (jie išmatuojami su paklaida ε).

Reguliarizavimo principas - apmokymo paklaida turi būti tos pačios eilės, kaip ir duomenų matavimo paklaida.



Modelio įvertinimas

Sukonstravus modelį (pvz. tiesinės regresijos modelį), jo tinkamumą prieš praktinį panaudojimą galime įvertinti naudodami įvairius statistinius metodus (Fišerio ir Stjudento kriterijus).

Apibrėžiame tokius dydžius: (statistinės imties dispersijos) (n - laisvės laipsnių skaičius).

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$ (nenormuota dispersija)

Normuoti d. $\frac{1}{n}$

arba patikl. $\frac{1}{n-1}$

Regresijos $f(x)$ ~~kai~~ skirtumų nuo vidutinės kvadratų reikšmė

$$SSR = \sum_{i=1}^{n} \left(f_{\alpha}(\alpha, x_i) - \bar{y} \right)^2$$

$\hat{y}_i$

Regresijos paklaidų nuokrypių kvadratų

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} SST - SSR$$

Determinacijos koeficientas

$$r^2 = \frac{SSR}{SST}$$

Jeigu $r^2 < 0.25$, ~~tai~~ priklauso, kad duomenys gali būti aproksimuoti tiesine regresijos kreive (tiese) nėra patikimas (nėra tinkamas prognozei)

Gautojo modelio savybių aptarimas (santrauka)

1. Modelio sudėtingumui didėjant apmokymo paklaida mažėja

$$MSE = \sum_{i=1}^{n}\left(\hat{y}_i - y_i\right)^2$$

(modelio reikšmės vis artimesnės eksperimentiniams duomenims, bet nereikšminga, kad y_i yra matuojami su matavimo paklaida!)

Svarbi charakteristika – modelio variacija. Ji nusako, kaip pasikeičia modelio prognozė, pakeitus apmokymo duomenų imtį. Labai tiksliems (lankstiems) modeliams variacija yra didelė (The Variance).

(The Bias)

Antroji charakteristika - modelio poslinkis. Jis parodo, kiek geri duotasis modelis gali aproksimuoti tikrąją f-ją $f^*(x)$. ~~Regos~~ Dažniausiai tikrieji, nuolatiniai modeliai (laukiamieji modeliai) teoriškai gali tikrąjį aproksimuoti ir ~~tesiškai~~ apmokymo duomenis ir tikrąją f-ją $f^*(x)$. Problema - turimi tik su pakankamai matematiniu duomenų modeliu per tikrąjį aproksimuoti apmokymo duomenys gali blogai aproksimuoti f-ją $f^*(x)$, o tai lems, kad netikslia prognozuosime y reikšmes nežinomiems x.