Strojno učenje

7. Regresija

prof. dr. sc. Bojana Dalbelo Bašić
doc. dr. sc. Jan Šnajder

Sveučilište u Zagrebu
Fakultet elektrotehnike i računarstva

Ak. god. 2012/13.

 v1.0

Danas...

Danas. . .

Regresija

$\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}, \quad \mathbf{x} \in \mathbb{R}^n, \quad y \in \mathbb{R}$

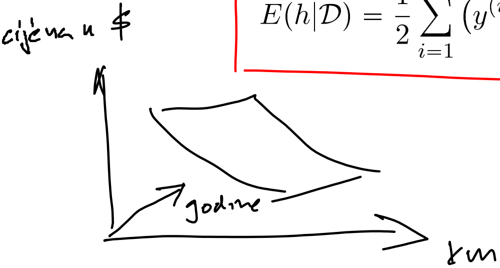
Hipoteza h aproksimira nepoznatu funkciju $f : \mathbb{R}^n \rightarrow \mathbb{R}$

$\mathbf{x}$ – ulazna varijabla (nezavisna, prediktorska)
y – izlazna varijabla (zavisna, kriterijska)

Funkcija pogreške:

$$\boxed{E(h|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - h(\mathbf{x}^{(i)})\right)^2}$$

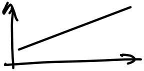
empirijska pogreška



Regresija

Linearan model regresije:

h je linearna funkcija parametara $\mathbf{w} = (w_0, \ldots, w_n)$



- Linearna regresija:
  $h(\mathbf{x}|\mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$

- Polinomijalna regresija:
  $h(\mathbf{x}|\mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d \quad (n = 1)$
  $h(\mathbf{x}|\mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2 \quad (n = 2, d = 2)$

- Općenite bazne funkcije:
  $h(\mathbf{x}|\mathbf{w}) = w_0 + w_1 \phi_1(\mathbf{x}) + \cdots + w_m \phi_m(\mathbf{x})$

međuovisnost
varijabli

Regresija

Broj ulaznih (nezavisnih) varijabli:

- Univarijatna (jednostavna, jednostruka) regresija: $n = 1$
- Multivarijatna (višestruka, multipla) regresija: $n > 1$

Broj izlaznih (zavisnih) varijabli:

- Jednoizlazna regresija: $f(\mathbf{x}) = y$ ← ovo radimo
- Višeizlazna regresija: $f(\mathbf{x}) = \mathbf{y}$

Danas. . .

Probabilistički model regresije

Ograničimo se BSO na univarijatnu ($n = 1$) linearnu regresiju:

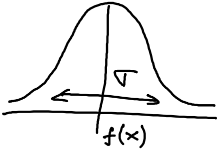
$$\boxed{h(x|\mathbf{w}) = w_0 + w_1 x} \qquad \boxed{\mathbf{w} = (w_0, w_1)}$$

Zbog šuma u $\mathcal{D}$:

$$y^{(i)} = f(x^{(i)}) + \varepsilon$$

Prepostavka:

$$\boxed{\varepsilon \sim \mathcal{N}(0, \sigma^2)}$$



Posjedično:

$$y|x \sim \mathcal{N}(h(x|\mathbf{w}), \sigma^2)$$

odnosno

$$p(y|x) = \mathcal{N}(h(x|\mathbf{w}), \sigma^2)$$

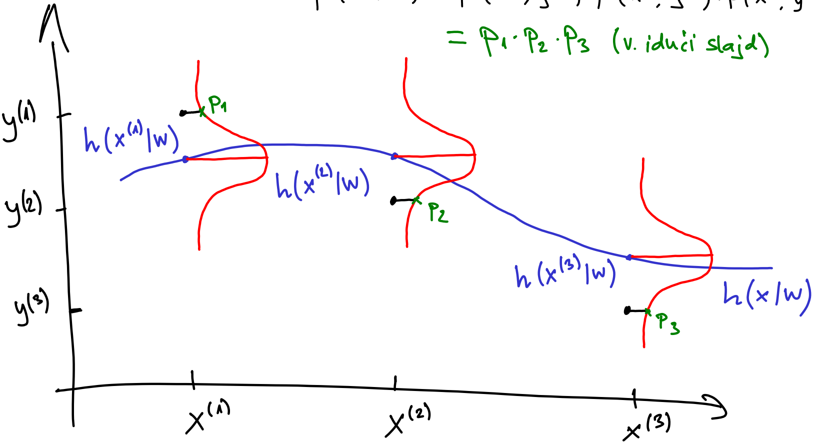
$$\mathbb{E}[y|x] = \mu = h(x|\mathbf{w})$$

Probabilistički model regresije

$$D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), (x^{(3)}, y^{(3)})\}$$

$$p(D|w) = p(x^{(1)}, y^{(1)}) \cdot p(x^{(2)}, y^{(2)}) \cdot p(x^{(3)}, y^{(2)})$$

$$= P_1 \cdot P_2 \cdot P_3 \quad \text{(v. idući slajd)}$$



P_i -> vjerojatnost da je y^i zbog šuma toliko udaljena od $h(x^i|w)$

Log-izglednost

$$\ln \mathcal{L}(\mathbf{w}|\mathcal{D}) = \ln p(\mathcal{D}|\mathbf{w}) = \ln \prod_{i=1}^{N} p(x^{(i)}, y^{(i)}) = \ln \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}) p(x^{(i)})$$

$$= \ln \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}) + \cancel{\ln \prod_{i=1}^{N} p(x^{(i)})}$$

ne ovisi o w

$$\Rightarrow \ln \prod_{i=1}^{N} p(y^{(i)}|x^{(i)}) = \ln \prod_{i=1}^{N} \mathcal{N}\big(\underbrace{h(x^{(i)}|\mathbf{w})}_{=\mu}, \sigma^2\big)$$

$$= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\big(y^{(i)} - h(x^{(i)}|\mathbf{w})\big)^2}{2\sigma^2}\right\}$$

$$= \underbrace{\cancel{-N \ln(\sqrt{2\pi}\sigma)}}_{\text{konst.}} - \frac{1}{2\cancel{\sigma^2}} \sum_{i=1}^{N} \big(y^{(i)} - h(x^{(i)}|\mathbf{w})\big)^2$$

$$\Rightarrow \boxed{\frac{1}{2} \sum_{i=1}^{N} \big(y^{(i)} - h(x^{(i)}|\mathbf{w})\big)^2}$$

zbroj kvadratnih odstupanja

MLE i postupak najmanjih kvadrata

Uz pretpostavku Gaussovog šuma, maksimizacija izglednosti odgovara minimizaciji funkcije pogreške definirane kao zbroj kvadratnih odstupanja:

$$\underset{\mathbf{w}}{\operatorname{argmax}} \ln \mathcal{L}(\mathbf{w}|\mathcal{D}) = \underset{\mathbf{w}}{\operatorname{argmin}}\, E(\mathbf{w}|\mathcal{D})$$

$$E(h|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(y^{(i)} - h(x^{(i)}|\mathbf{w})\right)^2$$

$$L\big(y, h(x|\mathbf{w})\big) \propto \big(y - h(x|\mathbf{w})\big)^2$$

$\Rightarrow$ Probabilističko opravdanje za kvadratnu funkciju gubitka

Rješenje MLE jednako je rješenju koje daje postupak najmanjih kvadrata!

$$\boxed{h(x|w_0, w_1) = w_0 + w_1 x}$$ model

$$\nabla_{w_0,w_1} E(h|\mathcal{D}) = 0$$

$$\vdots$$

$$\sum_{i=1}^{N} y^{(i)} = N w_0 + w_1 \sum_{i=1}^{N} x^{(i)}$$ (1.3) u skripti

$$\sum_{i=1}^{N} y^{(i)} x^{(i)} = w_0 \sum_{i=1}^{N} x^{(i)} + w_1 \sum_{i=1}^{N} (x^{(i)})^2$$ (1.5) u skripti

Matrični oblik: $\boxed{\mathbf{A}\mathbf{w} = \mathbf{z}}$

$$[\ A\ ][w] = [z]$$

$$\mathbf{A} = \begin{pmatrix} N & \sum_i x^{(i)} \\ \sum_i x^{(i)} & \sum_i (x^{(i)})^2 \end{pmatrix} \qquad \mathbf{w} = \begin{pmatrix} w_0 \\ w_1 \end{pmatrix} \qquad \mathbf{z} = \begin{pmatrix} \sum_i y^{(i)} \\ \sum_i y^{(i)} x^{(i)} \end{pmatrix}$$

Rješenje: $\boxed{\mathbf{w} = \mathbf{A}^{-1}\mathbf{z}}$

$$A^{-1} A w = A^{-1} z$$

Danas. . .

Bazne funkcije

Zanima nas poopćenje na $n > 1$ koje obuhvaća sve multivarijatne linearne modele regresije (univarijatna regresija, linearna regresija, polinomijalna regresija, ...).

Uvodimo fiksan skup baznih funkcija (nelinearne funkcije ulaznih varijabli):

$$\{\phi_j(\mathbf{x})\}_{j=1}^{m} \qquad \phi_j : \mathbb{R}^n \to \mathbb{R}$$

(Dogovorno: $\phi_0(\mathbf{x}) = 1$)

ulazni prostor

Funkcija preslikavanja (vektor baznih funkcija):

$$\boldsymbol{\phi}(\mathbf{x}) = \big(\phi_0(\mathbf{x}), \ldots, \phi_m(\mathbf{x})\big) \qquad \boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^m$$

prostor značajki (feature space)

Poopćen linearan model:

$$\boxed{h(\mathbf{x}|\mathbf{w}) = \sum_{j=0}^{m} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})}$$

Poopćeni linearan model

$$\boxed{h(\mathbf{x}|\mathbf{w}) = \sum_{j=0}^{m} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})}$$

- Linearna regresija:

$$\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, \ldots, x_n)$$ $n+1$

- Univarijatna polinomijalna regresija:

$$\boldsymbol{\phi}(x) = (1, x, x^2, \ldots, x^m)$$

- Polinomijalna regresija drugog stupnja:

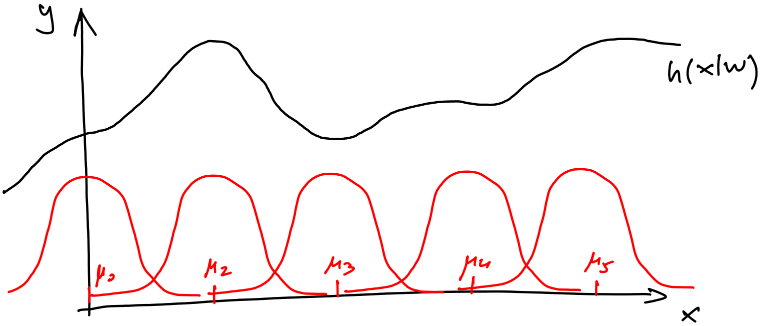
$$\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$$ $n=2$ $m=6$

- Gaussove bazne funkcije:

$$\phi_j(x) = \exp\left\{ -\frac{(x - \mu_j)^2}{2\sigma^2} \right\}$$

Gaussove bazne funkcije

(Općenito: radijalne bazne funkcije, RBF)



$$h(x|w) = w_1 \cdot \phi_1(x) + w_2 \cdot \phi_2(x) + w_3 \cdot \phi_3(x) + w_4 \cdot \phi_4(x) + w_5 \cdot \phi_5 = \underline{\Phi}(x)^T w$$

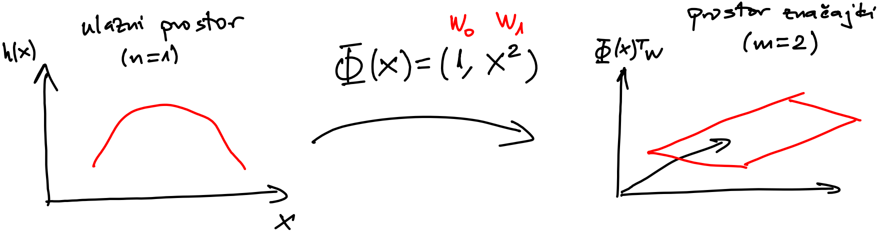
→ nelinearna fja u ulaznom prostoru

ali $\underline{\Phi}(x)^T w$ je linearna u prostoru značajki!

Preslikavanje značajki

Funkcija preslikavanja značajki $\phi : \mathbb{R}^n \to \mathbb{R}^m$ preslikava primjere iz n-dimenzijskog ulaznog prostora (engl. *input space*) u m-dimenzijski prostor značajki (engl. *feature space*).

Tipično je $m > n$. Funkcija koja je linearna u prostoru značajki je nelinearna u ulaznom prostoru.



- sada možemo koristiti linearan model za nelinearne probleme
- imamo jedinstven matematički tretman, neovisan o funkciji ϕ

Dizajn-matrica:

$$
\begin{array}{r l}
\text{1. primjer} \rightarrow & \\
 & \boldsymbol{\Phi} = \begin{pmatrix} 1 & \phi_1(\mathbf{x}^{(1)}) & \dots & \phi_m(\mathbf{x}^{(1)}) \\ 1 & \phi_1(\mathbf{x}^{(2)}) & \dots & \phi_m(\mathbf{x}^{(2)}) \\ \vdots & & & \\ 1 & \phi_1(\mathbf{x}^{(N)}) & \dots & \phi_m(\mathbf{x}^{(N)}) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\phi}(\mathbf{x}^{(1)})^{\mathrm{T}} \\ \boldsymbol{\phi}(\mathbf{x}^{(2)})^{\mathrm{T}} \\ \vdots \\ \boldsymbol{\phi}(\mathbf{x}^{(N)})^{\mathrm{T}} \end{pmatrix}_{N\times m} \\
\text{N-ti primjer} \rightarrow &
\end{array}
$$

Vektor izlaznih vrijednosti:

$$
\mathbf{y} = \begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix}
$$

$$
\underset{\Phi}{\overset{N\times m}{[\qquad]}}\,\underset{w}{\overset{m\times 1}{[\ \ ]}} = \underset{y}{\overset{N\times 1}{[\ \ ]}}
$$

Matrična jednadžba (N jednadžbi s m nepoznanica):

$$
\boxed{\boldsymbol{\Phi}\mathbf{w} = \mathbf{y}}
$$

Matrično rješenje najmanjih kvadrata

Rješenje je $\boxed{\mathbf{w} = \mathbf{\Phi}^{-1}\mathbf{y}}$

Međutim, rješenja nema ili ono nije jedinstveno ako:

(1) $\mathbf{\Phi}$ nije kvadratna, pa nema inverz. U pravilu:

- $N > m$
  $\Rightarrow$ sustav je preodređen (engl. *overdetermined*) i nema rješenja
- $N < m$
  $\Rightarrow$ sustav je pododređen (engl. *underdetermined*) i ima višestruka rješenja

(2) $\mathbf{\Phi}$ jest kvadratna ($N = m$), ali ipak nema inverz
$\Rightarrow$ sustav je nekonzistentan

Približno rješenje možemo naći postupkom najmanjih kvadrata
$\Rightarrow$ minimizacija kvadrata pogreške svake od jednadžbi

u smislu najmanjih kvadrata

Matrično rješenje najmanjih kvadrata

Funkcija pogreške:

$$\boxed{E(\mathbf{w}|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}^{(i)}) - y^{(i)}\right)^2}$$

Matrični oblik:

$$
\begin{aligned}
E(\mathbf{w}|\mathcal{D}) &= \boxed{\frac{1}{2}(\boldsymbol{\Phi}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\boldsymbol{\Phi}\mathbf{w} - \mathbf{y})}\\
&= \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} - \mathbf{y}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \mathbf{y}^{\mathrm{T}}\mathbf{y})\\
&= \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - 2\mathbf{y}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \mathbf{y}^{\mathrm{T}}\mathbf{y})
\end{aligned}
$$

$= (w^{T}\Phi^{T}y)^{T} = y^{T}\Phi w$

$(AB)^{T} = B^{T}A^{T}$

Matrično rješenje najmanjih kvadrata

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} - 2\mathbf{y}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} + \mathbf{y}^{\mathrm{T}}\mathbf{y})$$

$$\nabla_{\mathbf{w}}E = \frac{1}{2}\Big(\mathbf{w}^{\mathrm{T}}\big(\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} + (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{\mathrm{T}}\big) - 2\mathbf{y}^{\mathrm{T}}\mathbf{\Phi}\Big) = \boxed{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{y} = 0}$$

$$\frac{\mathrm{d}}{\mathrm{d}x}x^{\mathrm{T}}Ax = x^{\mathrm{T}}(A + A^{\mathrm{T}}) \qquad \frac{\mathrm{d}}{\mathrm{d}x}Ax = A$$

Sustav normalnih jednadžbi:

$$\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} = \mathbf{\Phi}^{\mathrm{T}}\mathbf{y} \quad /\cdot(\Phi^{T}\Phi)^{-1}$$

Rješenje:

$$\boxed{\mathbf{w} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y} = \mathbf{\Phi}^{+}\mathbf{y}}$$

$\mathbf{\Phi}^{+} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}}$ je pseudoinverz (Moore-Penroseov inverz) matrice $\mathbf{\Phi}$

Danas. . .

Odabir modela

Linearan model regresije ima hiperparametre:

- izgled baznih funkcija ϕ_j
- broj baznih funkcija m (dimenzija prostora značajki)

Parametre treba namjestiti tako da odgovaraju podatcima, odnosno treba dobro odabrati model. U suprotnom model može biti podnaučen ili prenaučen. Ako model ima mnogo parametra, lako ga je prenaučiti.

Sprečavanje prenaučenosti:

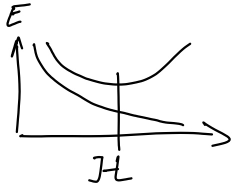
(1) koristiti više primjera za učenje

(2) odabrati model unakrsnom provjerom

(3) regularizacija ← danas

(4) bayesovska regresija

→ SU 2



Danas. . .

Regularizacija – motivacija

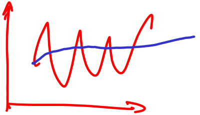
Opažanje: kod linearnih modela, što je model složeniji, to ima veće vrijednosti parametara $\mathbf{w}$.

$$h(x|w) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

Prenaučeni linearni modeli imaju:

(1) ukupno previše parametara (težina) i/ili

(2) prevelike vrijednosti pojedinačnih parametara.



Ideja: ograničiti rast vrijednosti parametara kažnjavanjem hipoteza s visokim vrijednostima parametara.

Time se ostvaruje kompromis između točnosti i jednostavnosti modela i to već pri samom učenju modela. Efektivno se ograničava složenost modela i sprečava se prenaučenost.

Cilj: što više parametara (težina) pritegnuti na nulu ⇒ rijetki modeli (engl. *sparse models*). interpretabilni + računalno jednostavniji

Složenost modela ugrađena je u funkciju pogreške:

$$E' = \text{empirijska pogreška} + \lambda \times \text{složenost modela}$$

regularizacijski faktor regularizacijski izraz

$$E'(\mathbf{w}|\mathcal{D}) = E(\mathbf{w}|\mathcal{D}) + \lambda E_w(\mathbf{w})$$

Veća vrijednost regularizacijskog faktora λ uzrokuje smanjuje efektivne složenost modela.

$\lambda = 0 \;\Rightarrow$ neregularizirana funkcija pogreške

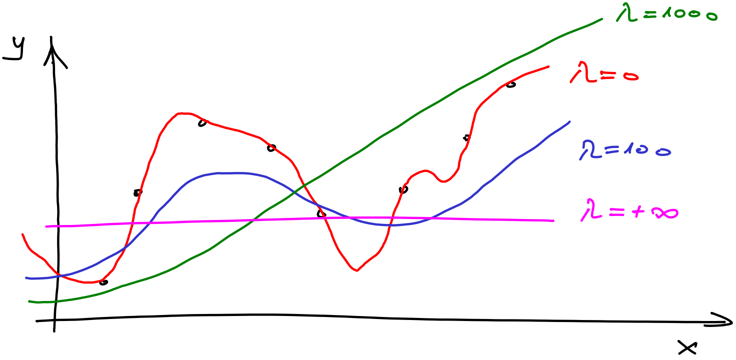
Općenit regularizacijski izraz:

$$\boxed{E_w(\mathbf{w}) = \frac{1}{2} \sum_{j=1}^{m} |w_j|^q} \longrightarrow q\text{-norma}$$

ne regularizirati w_0!

Npr. $n=1$, $m=6$:

$$h(x|w) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 + \ldots + w_6 x^6$$



Regularizirani linearni model regresije

L2-regularizacija (engl. *ridge regularization*) $(q = 2)$

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

(ima rješenje u zatvorenoj formi)

L1-regularizacija (engl. *LASSO regularization*) $(q = 1)$

(nema rješenja u zatvorenoj formi)

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_1$$

L0-regularizacija $(q = 0)$

(NP-complete)

$$E(\mathbf{w}|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)}) - y^{(i)}\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{m}\mathbf{1}\{w_j \neq 0\}$$

L2-regularizacija

Linearna regresija s L2-regularizacijom ima rješenje u zatvorenoj formi:

$$
\begin{aligned}
E'(\mathbf{w}|\mathcal{D}) &= \frac{1}{2}(\mathbf{\Phi}\mathbf{w}-\mathbf{y})^{\mathrm{T}}(\mathbf{\Phi}\mathbf{w}-\mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\
&= \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} - 2\mathbf{y}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} + \mathbf{y}^{\mathrm{T}}\mathbf{y} + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w})
\end{aligned}
$$

$\rightarrow w^T w = \|w\|^2$

$$
\begin{aligned}
\nabla_{\mathbf{w}}E' &= \mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{w} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{y} + \lambda\mathbf{w} \\
&= (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} + \lambda\mathbf{I})\mathbf{w} - \mathbf{\Phi}^{\mathrm{T}}\mathbf{y} = 0
\end{aligned}
$$

$$
\boxed{\mathbf{w} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi} + \lambda\mathbf{I})^{-1}\mathbf{\Phi}^{\mathrm{T}}\mathbf{y}}
$$

$\rightarrow I = \begin{bmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{bmatrix}$

L2-regularizacija

Ima li L2-regularizirana pogreška probabilističku interpretaciju?

Umjesto MLE, izračunajmo MAP:

apriorna razdioba težina

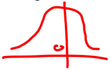
$$p(\mathbf{w}|\mathcal{D}) \propto p(\mathcal{D}|\mathbf{w})p(\mathbf{w})$$

Izglednost je normalna sa srednjom vrijednošću $h(\mathbf{x}|\mathbf{w})$:

$$p(\mathcal{D}|\mathbf{w}) = \prod_{i=1}^{N} \mathcal{N}\big(\underbrace{h(\mathbf{x}^{(i)}|\mathbf{w})}_{\mu}, \underbrace{\beta^{-1}}_{\sigma^2}\big) \qquad (\beta = 1/\sigma^2)$$

↳ "preciznost"

Pretpostavimo da apriorna PDF parametara normalna sa srednjom vrijednošću nula i izotropnom kov. matricom:

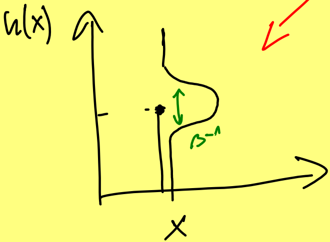


$$p(\mathbf{w}) = \mathcal{N}(0, \alpha^{-1}\mathbf{I}) \overset{\text{DZ}}{=} \left(\frac{\alpha}{2\pi}\right)^{m/2} \exp\left\{-\frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}\right\} \qquad (\alpha = 1/\sigma^2)$$

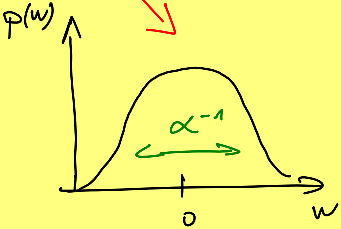
(α i β su hiperparametri modela)

MAP procjena

$$\boxed{p(w|D) \propto p(D|w) \cdot p(w)} \quad \text{MAP}$$



×



konjugatni par

Gauss ⟷ Gauss

$\mu = h(x|w)$ | $\mu = 0$

$\sigma^2 = \beta^{-1}$ | $\sigma^2 = \alpha^{-1}$

L2-regularizacija

Maksimizacija aposteriorne vjerojatnosti istovjetna je minimizaciji logaritma aposteriorne vjerojatnosti:

$$\underset{\mathbf{w}}{\operatorname{argmax}}\, p(\mathcal{D}|\mathbf{w})p(\mathbf{w}) = \underset{\mathbf{w}}{\operatorname{argmin}} \big(\ln p(\mathcal{D}|\mathbf{w}) + \ln p(\mathbf{w}) \big)$$

$$\ln p(\mathcal{D}|\mathbf{w}) = \frac{\beta}{2} \sum_{i=1}^{N} \big(y^{(i)} - h(\mathbf{x}|\mathbf{w})\big)^2 + \cancel{\text{konst.}} \qquad \ln p(\mathbf{w}) = \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$$

Dakle, pogreška koju minimiziramo je:

$$\boxed{\frac{\beta}{2} \sum_{i=1}^{N} \big(y^{(i)} - h(\mathbf{x}|\mathbf{w})\big)^2 + \frac{\alpha}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}} \qquad / : \beta$$

a to je upravo L2-regularizirana pogreška! (uz $\lambda = \alpha/\beta$)

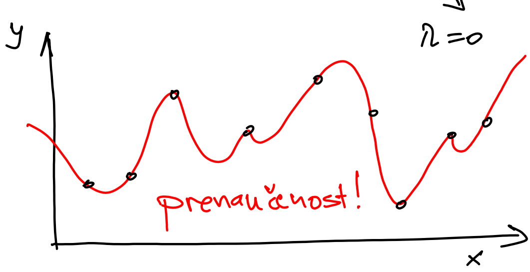
Regularizacija – napomene

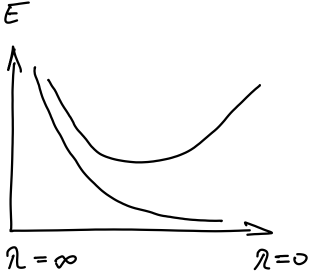
- Iznos parametra w_j odgovara važnosti značajke, a predznak upućuje na njezin utjecaj (pozitivan ili negativan) na izlaznu vrijednost
- Regularizacija smanjuje složenost modela na način da prigušuje vrijednosti pojedinih značajki, odnosno efektivno ih izbacuje (kada $w_j \to 0$). Ako je model nelinearan, to znači smanjivanje nelinearnosti
- Težinu w_0 treba izuzeti iz regularizacijskog izraza (jer ona definira pomak) ili treba centrirati podatke tako da $\overline{y} = 0$, jer onda $w_0 \to 0$
- L2-regularizacija kažnjava težine proporcionalno njihovom iznosu (velike težine više, a manje težine manje). Teško će parametri biti pritegnuti baš na nulu. Zato L2-regularizacija ne rezultira rijetkim modelima
- L1-regularizacija rezultira rijetkim modelima, ali nema rješenja u zatvorenoj formi (međutim mogu se koristiti iterativni optimizacijski postupci)

Regularizacija – napomene

- Regularizacija je korisna kod modela s puno parametara, jer je takve modele lako prenaučiti
- Regularizacija smanjuje mogućnost prenaučenosti, ali ostaje problem odabira hiperparametra λ te drugih hiperparametara (broj i oblik baznih funkcija). Taj se odabir najčešće radi unakrsnom provjerom

Q: Koju optimalnu vrijednost za λ bismo dobili kada bismo optimizaciju radili na skupu za učenje?





Sažetak

- **Linearan model regresije** linearan je u parametrima
- Nelinearnost regresijske funkcije ostvaruje se uporabom nelinearnih **baznih funkcija** (preslikavanjem ulaznog prostora u prostor značajki)
- Uz pretpostavku normalno distribuiranog šuma, **MLE je istovjetan postupku najmanjih kvadrata**, što daje probabilističko opravdanje za uporabu kvadratne funkcije gubitka
- Parametri linearnog modela uz kvadratnu funkciju gubitka imaju rješenje u zatvorenoj formi u obliku **pseudoinverza dizajn-matrice**
- **Regularizacija smanjuje prenaučenost** ugradnjom dodatnog izraza u funkciju pogreške kojim se kažnjava složenost modela
- **L2-regularizirana regresija istovjetna je MAP-procjeni** parametara (uz pretpostavku Gaussovog šuma) te ima rješenje u zatvorenoj formi

Sljedeća tema: Linearni diskriminativni modeli