4. Regresija II

Strojno učenje 1, UNIZG FER, ak. god. 2021./2022.

Jan Šnajder, natuknice s predavanja, v1.5

1 Nelinearna regresija

- Veza između nezavisnih varijabli i zavisne varijable često je **nelinearna**
- Neki nelinearni regresijski modeli:
  - Linearna višestruka regresija:

$$h(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$$

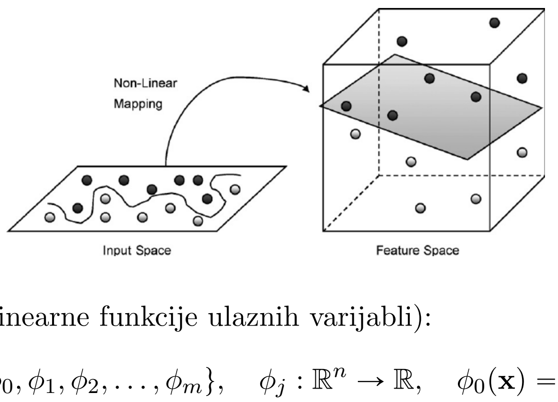
  - Jednostruka polinomijalna regresija ($n = 1$):

$$h(x; \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \cdots + w_d x^d$$

  - Višestruka polinomijalna regresija ($n = 2, d = 2$):

$$h(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + w_3 x_1 x_2 + w_4 x_1^2 + w_5 x_2^2$$

    gdje je $x_1 x_2$ **interakcijska značajka** (*cross-term*)
- Umjesto da mijenjamo model, mijenjamo podatke $\Rightarrow$ **preslikavanje u prostor značajki**



- **Bazne funkcije** (nelinearne funkcije ulaznih varijabli):

$$\{\phi_0, \phi_1, \phi_2, \ldots, \phi_m\}, \quad \phi_j : \mathbb{R}^n \to \mathbb{R}, \quad \phi_0(\mathbf{x}) = 1$$

- **Funkcija preslikavanja** u prostor značajki:

$$\boldsymbol{\phi} : \mathbb{R}^n \to \mathbb{R}^{m+1} :$$
$$\boldsymbol{\phi}(\mathbf{x}) = \big(\phi_0(\mathbf{x}), \ldots, \phi_m(\mathbf{x})\big)$$

- Model s ugrađenom funkcijom preslikavanja:

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

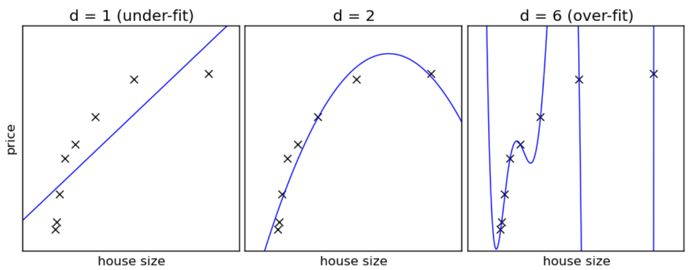
- Ova je **linearan model regresije** (linearan u parametrima) $\neq$ linearna regresija
- Uobičajene funkcije preslikavanja:
  - Linearna višestruka regresija: $\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, \dots, x_n)$
  - Jednostruka polinomijalna regresija: $\boldsymbol{\phi}(x) = (1, x, x^2, \dots, x^m)$
  - Višestruka polinomijalna regresija drugog stupnja: $\boldsymbol{\phi}(\mathbf{x}) = (1, x_1, x_2, x_1 x_2, x_1^2, x_2^2)$
- Matrica dizajna s preslikavanjem:

$$\boldsymbol{\Phi} = \begin{pmatrix} 1 & \phi_1(\mathbf{x}^{(1)}) & \dots & \phi_m(\mathbf{x}^{(1)}) \\ 1 & \phi_1(\mathbf{x}^{(2)}) & \dots & \phi_m(\mathbf{x}^{(2)}) \\ \vdots & & & \\ 1 & \phi_1(\mathbf{x}^{(N)}) & \dots & \phi_m(\mathbf{x}^{(N)}) \end{pmatrix}_{N\times(m+1)} = \begin{pmatrix} \phi(\mathbf{x}^{(1)})^{\mathrm{T}} \\ \phi(\mathbf{x}^{(2)})^{\mathrm{T}} \\ \vdots \\ \phi(\mathbf{x}^{(N)})^{\mathrm{T}} \end{pmatrix}_{N\times(m+1)}$$

- Rješenje najmanjih kvadrata: $\mathbf{w} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} = \boldsymbol{\Phi}^{+}\mathbf{y}$

2 Prenaučenost

- Odabir preslikavanja $\boldsymbol{\phi}$ je **hiperparametar** modela
- Nelinearan model je složeniji od linearnog $\Rightarrow$ sklonost **prenaučenosti**



- Rješenje: učiti na više primjera, odabir modela, regularizacija, bayesovska regresija

3 Regularizacija

- Složeniji model $\Leftrightarrow$ veće magnitude parametara (težina) $\mathbf{w}$
- Ograničavanje rasta parametara pri učenju $\Rightarrow$ **regularizacija**
- **Rijetki modeli** (*sparse models*) – modeli s težinama pritegnutima na nulu

- **Regularizirana funkcija pogreške**:

$$E_R(\mathbf{w}|\mathcal{D}) = E(\mathbf{w}|\mathcal{D}) + \underbrace{\lambda\Omega(\mathbf{w})}_{\text{reg. izraz}}$$

  gdje je λ **regularizacijski faktor** $\Rightarrow$ kompromis između jednostavnosti i složenosti

- Regularizacijski izraz je p-norma vektora težina:

$$\Omega(\mathbf{w}) = \|\mathbf{w}\|_p = \Big(\sum_{j=1}^{m} |w_j|^p\Big)^{\frac{1}{p}}$$

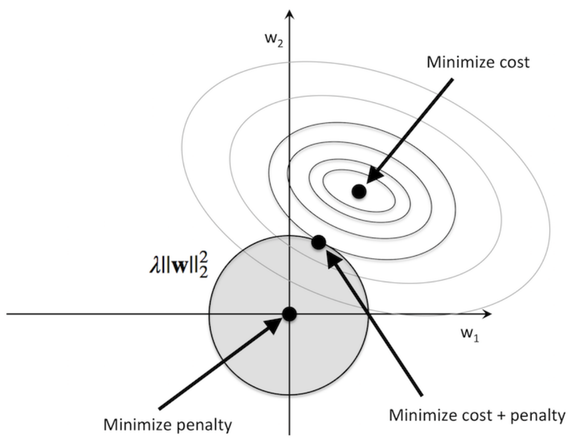
  - L2-norma ($p = 2$): $\|\mathbf{w}\|_2 = \sqrt{\sum_{j=1}^{m} w_j^2} = \sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$
  - L1-norma ($p = 1$): $\|\mathbf{w}\|_1 = \sum_{j=1}^{m} |w_j|$
  - L0-norma ($p = 0$): $\|\mathbf{w}\|_0 = \sum_{j=1}^{m} \mathbf{1}\{w_j \neq 0\}$

- **L2-regularizacija** (Tikhonovljeva regularizacija):

$$E_R(\mathbf{w}|\mathcal{D}) = \frac{1}{2}\sum_{i=1}^{N} \big(\mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x}^{(i)}) - y^{(i)}\big)^2 + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

  $\Rightarrow$ **hrbatna regresija** (*ridge regression*)

- Skica: izokonture L2-regularizirane funkcije pogreške

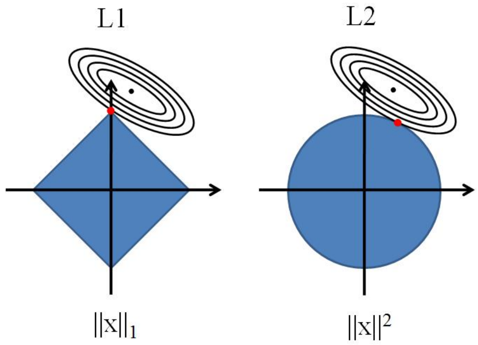


- **L1-regularizacija (LASSO)**:

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

  $\Rightarrow$ daje rijetke modele

- Skica: usporedba izokontura L1- i L2-regulariziranih funkcija pogreške



- **L0-regularizacija**:

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

$\Rightarrow$ efektivno provodi **odabir značajki**

- L0-regularizacija je NP-potpuna, L1-regularizacija nema rješenje u zatvorenoj formi
- Rješenje najmanjih kvadrata s L2-regularizacijom:

$$\begin{aligned}
E_R(\mathbf{w}|\mathcal{D}) &= \frac{1}{2}(\boldsymbol{\Phi}\mathbf{w} - \mathbf{y})^{\mathrm{T}}(\boldsymbol{\Phi}\mathbf{w} - \mathbf{y}) + \frac{\lambda}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} \\
&= \frac{1}{2}(\mathbf{w}^{\mathrm{T}}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - 2\mathbf{y}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} + \mathbf{y}^{\mathrm{T}}\mathbf{y} + \lambda\mathbf{w}^{\mathrm{T}}\mathbf{w}) \\
\nabla_{\mathbf{w}}E_R &= \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\mathbf{w} - \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} + \lambda\mathbf{w} \\
&= (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda\mathbf{I})\mathbf{w} - \boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y} = 0 \\
\mathbf{w} &= (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda\mathbf{I})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y}
\end{aligned}$$

gdje $\lambda\mathbf{I} = \mathrm{diag}(0, \lambda, \ldots, \lambda)$ (težinu w_0 ne regulariziramo)

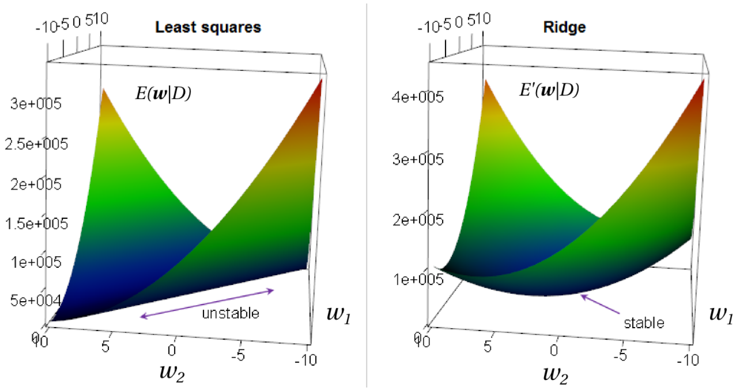
4 Regularizacija i kondicija matrice

- Rješenje najmanjih kvadrata: $\mathbf{w} = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\mathbf{y}$
- $(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi})^{-1}$ definiran $\Leftrightarrow$ $\mathrm{rang}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}) = \mathrm{rang}(\boldsymbol{\Phi}) = m + 1$ $\Leftrightarrow$ linearno nezavisni stupci
- Linearno zavisni stupci $\Leftrightarrow$ redundantne značajke $\Leftrightarrow$ **savršena multikolinearnost**
- **Multikolinearnost** – dvije varijable ili više njih su visoko korelirane
- Multikolinearnost daje **numerički nestabilno rješenje** $\Rightarrow$ **prenaučenost**
- Nestabilnost rješenja iskazuje se **kondicijskim brojem** matrice
- $m \gg N$ $\Leftrightarrow$ "široka i plitka" matrica dizajna $\Rightarrow$ $\mathrm{rang}(\boldsymbol{\Phi}) < m+1$ $\Rightarrow$ multikolinearnost
- Regularizacija smanjuje multikolinearnost:

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

$\Rightarrow$ dodavanje dijagonale smanjuje linearnu zavisnost $\Rightarrow$ **rekondicioniranje matrice**

- Regularizacijom funkcija pogreške postaje konveksnija (nestaje hrbat)



5 Napomene

- Magnituda parametra w_i odgovara **važnosti značajke**, osim ako je model prenaučen
- Regularizacija **sprječava prenaučenost** prigušujući vrijednosti značajki
- Ako je model nelinearan, regularizacijom smanjujemo nelinearnost
- Težinu w_0 treba izuzeti iz regularizacijskog izraza ili treba centrirati podatke
- Odabir hiperparametra λ najčešće se provodi **unakrsnom provjerom**