3. Regresija

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

Jan Šnajder, natuknice s predavanja, v1.5

1 Jednostavna regresija

- Označen skup podataka: $\mathcal{D} = \{(\mathbf{x}^{(i)}, y^{(i)})\}, \mathbf{x} \in \mathbb{R}^n, y \in \mathbb{R}$
- Hipoteza $h : \mathbb{R}^n \to \mathbb{R}$
- $\mathbf{x}$ – ulazne/nezavisne/prediktorske varijable; y – izlazna/zavisna/kriterijska varijabla
- **Linearna regresija**:

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

- **Jednostavna regresija** ($n = 1$):

$$h(x; w_0, w_1) = w_0 + w_1 x$$

- Funkcija gubitka je **kvadratni gubitak**: $L(y, h(x)) = (y - h(x))^2$
- Funkcija pogreške je zbroj kvadratnih gubitaka (**reziduala**):

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

- Optimizacijski postupak:

$$h^* = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, E(h|\mathcal{D}) = \underset{h \in \mathcal{H}}{\operatorname{argmin}}\, \frac{1}{2} \sum_{i=1}^{N} \big(y^{(i)} - h(x^{(i)})\big)^2$$

- Za jednostavnu regresiju:

$$\begin{aligned}
&\nabla_{w_0,w_1} E(h|\mathcal{D}) = 0 \\
&\frac{\partial}{\partial w_0}\Big[\frac{1}{2} \sum_{i}^{N} \big(y^{(i)} - (w_1 x^{(i)} + w_0)\big)^2\Big] = 0 \\
&\frac{\partial}{\partial w_1}\Big[\frac{1}{2} \sum_{i}^{N} \big(y^{(i)} - (w_1 x^{(i)} + w_0)\big)^2\Big] = 0 \\
&\vdots \\
&w_0 = \bar{y} - w_1 \bar{x} \\
&w_1 = \frac{\sum_i^N x^{(i)} y^{(i)} - N\bar{x}\bar{y}}{\sum_i^N (x^{(i)})^2 - N\bar{x}^2}
\end{aligned}$$

2 Vrste regresije

- Ulazne varijable: **jednostavna** ($n = 1$) ili **višestruka** ($n > 1$)
- Izlazne varijable: **univarijatna** ($f(\mathbf{x}) = y$) ili **multivarijatna** ($f(\mathbf{x}) = \mathbf{y}$)

	Jedan izlaz	Više izlaza
Jedan ulaz	**(Univarijatna) jednostavna**	**Multivarijatna jednostavna**
Više ulaza	**(Univarijatna) višestruka**	**Multivarijatna višestruka**

- Mi radimo samo univarijatnu regresiju

3 Tri komponente linearne regresije

- (1) Model:

$$h(\mathbf{x}; \mathbf{w}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n = \sum_{i=1}^{n} w_i x_i + w_0 = h(\mathbf{x}; \mathbf{w}) = \mathbf{w}^{\mathrm{T}} \mathbf{x}$$

- (2) Funkcija gubitka i funkcija pogreške:

$$L\big(y^{(i)}, h(\mathbf{x}^{(i)})\big) = \big(y^{(i)} - h(\mathbf{x}^{(i)})\big)^2$$

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

- (3) Optimizacijski postupak:

$$\operatorname*{argmin}_{\mathbf{w}} E(\mathbf{w}|\mathcal{D})$$

  $\Rightarrow$ **metoda najmanjih kvadrata** (*ordinary least squares, OLS*)

- Postoji rješenje u **zatvorenoj formi**

4 Postupak najmanjih kvadrata

- Označeni primjeri daju N jednadžbi s $(n + 1)$ nepoznanica:

$$\forall (\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{D}.\ \mathbf{w}^{\mathrm{T}} \mathbf{x} = y^{(i)}$$

- Matrično:

$$\underbrace{\begin{pmatrix} 1 & x_1^{(1)} & x_2^{(1)} \dots & x_n^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} \dots & x_n^{(2)} \\ \vdots & & & \\ 1 & x_1^{(N)} & x_2^{(N)} \dots & x_n^{(N)} \end{pmatrix}}_{\mathbf{X}} \cdot \underbrace{\begin{pmatrix} w_0 \\ w_1 \\ \vdots \\ w_n \end{pmatrix}}_{\mathbf{w}} = \underbrace{\begin{pmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{pmatrix}}_{\mathbf{y}}$$

- Matrica $\mathbf{X}$ je **matrica dizajna**
- Egzaktno rješenje je $\mathbf{w} = \mathbf{X}^{-1}\mathbf{y}$, ali ono ne postoji ako:
  - $\mathbf{X}$ nije kvadratna $\Rightarrow$ pre/pododređenost sustava
  - $\mathbf{X}$ je kvadratna, ali je sustav nekonzistentan
- Umjesto egzaktnog, tražimo približno rješenje (najmanja kvadratna odstupanja)
- Funkcija pogreške u matričnom obliku:

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

  uz $(A^{\mathrm{T}})^{\mathrm{T}} = A$ i $(AB)^{\mathrm{T}} = B^{\mathrm{T}}A^{\mathrm{T}}$
- Minimizacija:

$$
\begin{aligned}
\nabla_{\mathbf{w}}E &= \frac{1}{2}\Big(\mathbf{w}^{\mathrm{T}}\big(\mathbf{X}^{\mathrm{T}}\mathbf{X}+(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{\mathrm{T}}\big)-2\mathbf{y}^{\mathrm{T}}\mathbf{X}\Big) = \mathbf{w}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}-\mathbf{y}^{\mathrm{T}}\mathbf{X} = \mathbf{0} \\
\mathbf{w}^{\mathrm{T}} &= \mathbf{y}^{\mathrm{T}}\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \\
\mathbf{w} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y} = \mathbf{X}^{+}\mathbf{y}
\end{aligned}
$$

  uz $\frac{\mathrm{d}}{\mathrm{d}x}Ax = A$ i $\frac{\mathrm{d}}{\mathrm{d}x}x^{\mathrm{T}}Ax = x^{\mathrm{T}}(A+A^{\mathrm{T}})$
- Matrica $\mathbf{X}^{+} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$ je Moore-Penroseov **pseudoinverz** matrice dizajna $\mathbf{X}$
- Pseudoinverz minimizira normu $\|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2$
- Ako je $\mathbf{X}$ kvadratna i punog ranga, onda $\mathbf{X}^{+} = \mathbf{X}^{-1}$
- $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ je **Gramova matrica**; $\mathrm{rang}(\mathbf{X}^{\mathrm{T}}\mathbf{X}) = \mathrm{rang}(\mathbf{X})$
- Ako je $\mathrm{rang}(\mathbf{X}) = n+1$, onda $\mathbf{X}^{+} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}$
- Dimenzija Gramove matrice je $(n+1)\times(n+1) \Rightarrow$ izračun inverza je moguće skup
- Ako $\mathrm{rang}(\mathbf{X}) < n+1$ (plitka matrica), onda $\mathbf{X}^{+}$ računamo pomoću SVD-a

5 Probabilistička interpretacija regresije

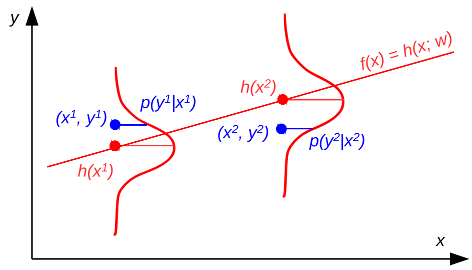
- Opažena oznaka je zbroj vrijednosti funkcije i šuma: $y^{(i)} = f(\mathbf{x}^{(i)}) + \varepsilon_i$
- Šum modeliramo kao normalno distribuiranu **slučajnu varijablu**: $\varepsilon \sim \mathcal{N}(0,\sigma^2)$
- **Normalna razdioba**:

$$
p(Y=y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\Big(-\frac{(y-\mu)^2}{2\sigma^2}\Big)
$$

- Vjerojatnost oznake za zadani primjer: $p(y|\mathbf{x}) = \mathcal{N}\big(f(\mathbf{x}), \sigma^2\big)$
- Vjerojatnost da je cijeli skup primjera $\mathbf{X}$ označen oznakama $\mathbf{y}$ (uz pretpostavku **iid**):

$$p(\mathbf{y}|\mathbf{X}) = \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)})$$

$\Rightarrow$ **izglednost** (*likelihood*) (vjerojatnost oznaka pod modelom)



- Radi matematičke jednostavnosti, radimo s logaritmom izglednosti $\Rightarrow$ **log-izglednost**
- Tražimo $\mathbf{w}$ koji oznake čini najvjerojatnijim $\Leftrightarrow$ maksimizacija log-izglednosti
- Vrijedi $h(\mathbf{x};\mathbf{w}) = f(\mathbf{x})$ (hipoteza treba aproksimirati funkciju $f(\mathbf{x})$)
- Log-izglednost težina $\mathbf{w}$:

$$\begin{aligned}
\ln p(\mathbf{y}|\mathbf{X},\mathbf{w}) &= \ln \prod_{i=1}^{N} p(y^{(i)}|\mathbf{x}^{(i)}) = \ln \prod_{i=1}^{N} \mathcal{N}\big(f(\mathbf{x}^{(i)}), \sigma^2\big) \\
&= \ln \prod_{i=1}^{N} \mathcal{N}\big(h(\mathbf{x}^{(i)};\mathbf{w}), \sigma^2\big) \\
&= \ln \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\Big(-\frac{\big(y^{(i)} - h(\mathbf{x}^{(i)};\mathbf{w})\big)^2}{2\sigma^2}\Big) \\
&= \underbrace{-N\ln(\sqrt{2\pi}\sigma)}_{=\text{konst.}} - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\big(y^{(i)} - h(\mathbf{x}^{(i)};\mathbf{w})\big)^2 \\
&\propto -\frac{1}{2}\sum_{i=1}^{N}\big(y^{(i)} - h(\mathbf{x}^{(i)};\mathbf{w})\big)^2
\end{aligned}$$

$\Rightarrow$ maksimizacija izglednosti $\Leftrightarrow$ minimizacija pogreške kvadratnog gubitka