22. Vrednovanje modela II

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

Jan Šnajder, natuknice s predavanja, v1.3

1 Statističko zaključivanje – ideja

- Točnost modela mjerimo na slučajnome uzorku $\Rightarrow$ točnost je **slučajna varijabla** (s.v.)
- **Statističko zaključivanje** omogućava zaključivanje na temelju slučajnog uzorka
- Osnovni pristupi: (1) **interval pouzdanosti** i (2) **statističko testiranje hipoteze**
- Osnovni pojmovi:
  - **populacija** – konačan ili beskonačan skup svih objekata od interesa
  - **uzorak** – podskup populacije veličine N dobiven (slučajnim) uzorkovanjem
  - **statistika** – procjenitelj (funkcija uzorka) koji odgovara parametru populacije
- Parametarsko statističko zaključivanje $\Rightarrow$ temeljeno na distr. uzorkovanja statistike
- **Distribucija uzorkovanja** (*sampling distribution*) – distr. statistike na temelju sl. uzorka
- **Standardna pogreška (SE)** (*standard error*) – stand. devijacija distr. uzorkovanja
- Za neke je statistike distr. uzorkovanja u zatvorenoj formi, npr., srednja vrijednost
- **Distribucija uzorkovanja srednje vrijednosti**:

$$\text{populacija: } x \sim \mathcal{N}(\mu, \sigma^2) \quad \Rightarrow \quad \text{statistika: } \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$$

  $\Rightarrow \text{SE} = \sqrt{\sigma^2/N} = \sigma/\sqrt{N}$

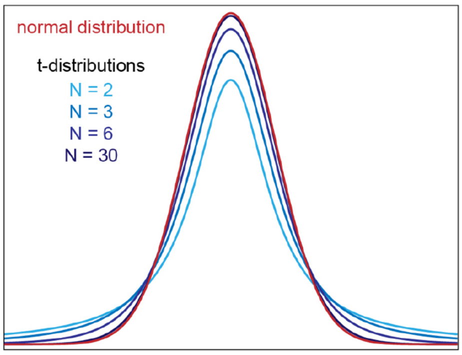
- **Središnji granični teorem**: za $N \to \infty$ vrijedi $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$, neovisno o distr. od x
- U praksi, već za $N \geqslant 30$ možemo pretpostaviti $\bar{x} \sim \mathcal{N}(\mu, \frac{\sigma^2}{N})$
- **Standardizacijom** od $\bar{x}$ dobivamo **z-vrijednost** s distribucijom uzorkovanja:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1)$$

- Ako je σ^2 populacije **nepoznata**, procjenjujemo $\hat{\sigma}^2$ iz uzorka i koristimo **t-statistiku**:

$$t = \frac{\bar{x} - \mu}{\hat{\sigma}/\sqrt{N}} \sim t(N-1)$$

gdje je $t(N-1)$ **Studentova t-distribucija** sa $N-1$ stupnjeva slobode



- Za $N \geqslant 30$, t-distribucija praktički je identična normalnoj
- Sažetak pravila (za statistiku $\bar{x}$ i varijancu $\hat{\sigma}^2$ procijenjenu iz uzorka):
  - $N \geqslant 30$ ("velik uzorak") $\Rightarrow$ koristimo z-statistiku ili t-statistiku (svejedno)
  - $N < 30$ i populacija je normalna $\Rightarrow$ koristimo t-statistiku
  - $N < 30$ i populacija nije normalna $\Rightarrow$ ne radimo param. stat. zaključivanje!
- Provjera normalnosti: Shapiro-Wilkov test ili **Q-Q plot** za normalnu distribuciju

2 Statističko zaključivanje za vrednovanje modela

- Distribucija uzorkovanja nije poznata za sve mjere vrednovanja (npr., za F1-mjeru)
- Ideja: koristiti srednju vrijednost odabrane mjere izračunatu na K preklopa:
  - **populacija** – svi mogući primjeri (moguće beskonačno)
  - **uzorak** – vrijednosti mjere na K preklopa višestruke unakrsne provjere
  - **statistika** – srednja vrijednost mjere kroz K preklopa
- NB: Veličina statističkog uzorka je K (broj preklopa), a ne N (broj primjera)!
- Tipično $K < 30$, pa treba provjeriti normalnost

3 Interval pouzdanosti

- **Interval pouzdanosti** srednje vrijednosti populacije μ na temelju sredine uzorka $\bar{x}$:

$$\mu = \bar{x} \pm \text{SE} \quad \text{(sa X\% pouzdanosti)}$$

- Budući da

$$z = \frac{\bar{x} - \mu}{\mathrm{SE}} \sim \mathcal{N}(0,1)$$

  to (za $X = 95\%$) vrijedi

$$P(-1.96 \leqslant (\bar{x} - \mu)/\mathrm{SE} \leqslant 1.96) = 0.95$$

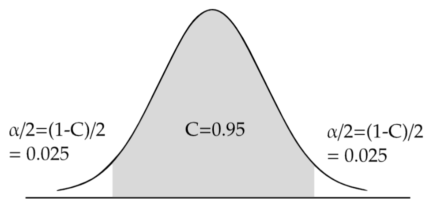
  što (uz $\mathrm{SE} = \sigma/\sqrt{N}$) daje

$$P(\bar{x} - 1.96\sigma/\sqrt{N} \leqslant \mu \leqslant \bar{x} + 1.96\sigma/\sqrt{N}) = 0.95$$

  odnosno

$$\mu = \bar{x} \pm 1.96\sigma/\sqrt{N} \quad \text{(sa 95\% pouzdanosti)}$$

- Veza između **razine pouzdanosti** $C \in [0,1]$ i **razine značajnosti** $\alpha \in [0,1]$: $C = 1 - \alpha$

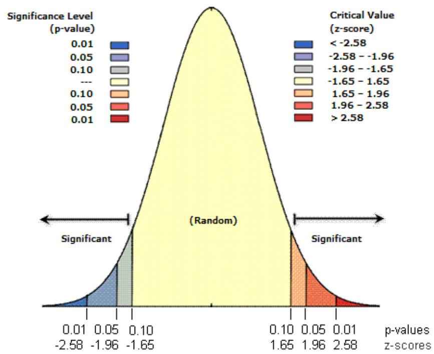


- Općenito, normalan interval pouzdanosti razine $C = 1 - \alpha$:

$$P(\bar{x} - z_{\alpha/2}\sigma/\sqrt{N} \leqslant \mu \leqslant \bar{x} + z_{\alpha/2}\sigma/\sqrt{N}) = 1 - \alpha$$

  gdje je $z_{\alpha/2}$ **kritična vrijednost** distribucije $\mathcal{N}(0,1)$ za razinu značajnosti α, tj.:

$$P(|z| \geqslant z_{\alpha/2}) = \alpha$$



- Ako je $\hat{\sigma}^2$ procijenjen iz uzorka, umjesto z-statistike treba upotrijebiti t-statistiku:

$$P(\bar{x} - t_{\alpha/2}\hat{\sigma}/\sqrt{N} \leqslant \mu \leqslant \bar{x} + t_{\alpha/2}\hat{\sigma}/\sqrt{N}) = 1 - \alpha$$

  gdje je $t_{\alpha/2}$ **kritična vrijednost** distribucije $t(N-1)$ za razinu značajnosti α

- Kritične vrijednosti z-distribucije i t-distribucije očitavaju se iz tablica

4 Statističko testiranje hipoteze

- Pretpostavljamo da parametar populacije μ ima neku vrijednost (**hipoteza**)
- Možemo li **odbaciti** tu hipotezu na temelju opažanja $\bar{x}$, koje donekle odstupa od μ?
- **p-vrijednost**: vjerojatnost da smo opazili $\bar{x}$ ili ekstremnije, ako je hipoteza istinita
- Hipotezu odbacujemo ako je p-vrijednost manja od odabrane **razine značajnosti** α
- **t-test** za srednju vrijednost: koristi t-statistiku kao testnu statistiku

t-test za srednju vrijednost

- Korak 1: Iskazati hipotezu H_0 i njoj alternativnu hipotezu H_1:

$$\begin{aligned} H_0 :& \mu = ... \\ H_1 :& \mu \neq ... \end{aligned}$$

- Korak 2: Odabrati nivo značajnosti α, npr. $\alpha = 1\%$ (ili 5%)
- Korak 3: Izračunati t-statistiku pod hipotezom H_0: $t = \frac{\bar{x}-\mu}{\hat{\sigma}/\sqrt{N}}$
- Korak 4: Na distribuciji $t(N-1)$ provjeriti:
  - Varijanta A (provjera kritičnog područja): je li $|t| \geqslant t_{\alpha/2}$?
  - Varijanta B (provjera p-vrijednosti): je li $P(|X| > t) \leqslant \alpha$?
- Korak 5:
  - Ako da: odbaciti hipotezu H_0 i prihvatiti hipotezu H_1
  - Ako ne: ne možemo odbaciti (ali niti prihvatiti) hipotezu H_0
- Korak 6: Formulirati zaključak

- **Jednostrani test** (*one-tailed test*):
  - Testiramo je li $\bar{x}$ veće/manje od μ
  - Hipoteza H_0 je ista, alternativna hipoteza je $H_1 : \mu > \ldots$ ili $H_1 : \mu < \ldots$
  - p-vrijednost je polovica p-vrijednosti za dvostrani test $\Rightarrow$ lakše je odbaciti H_0
  - Kod vrednovanja modela u načelu treba izbjegavati jednostrani test

5 Usporedba modela

- Je li točnost modela A **statistički značajno** različita/bolja od točnosti modela B?

- **Upareni t-test** (*matched-pair t-test*): testiranje razlika u točnosti kroz K preklopa
- Uzorak je $\{d_k\}_{i=1}^K$, gdje je $d_i = m_i^A - m_i^B$ razlika u mjeri m na preklopu i
- Izračunavamo srednju vrijednost razlika, $\bar{d} = \overline{m}^A - \overline{m}^B = \frac{1}{K}\sum_{i=1}^K d_i$
- Hipoteze:

$$
\begin{aligned}
H_0 &: \overline{m}^A - \overline{m}^B = \bar{d} = 0 && \text{točnosti su iste} \\
H_1 &: \overline{m}^A - \overline{m}^B \neq 0 && \text{točnosti su različite (dvostrani test)} \\
\text{ili } H_1 &: \overline{m}^A - \overline{m}^B \lesseqgtr 0 && \text{točnost od A je manja/veća od B (jednostrani test)}
\end{aligned}
$$

- t-statistika:

$$t = \frac{\bar{d} - 0}{\hat{\sigma}/\sqrt{K}}, \quad \text{gdje} \quad \hat{\sigma} = \sqrt{\frac{\sum_{i=1}^K (d_i - \bar{d})^2}{K-1}}$$

- Ako je $K < 30$, treba provjeriti vrijedi li normalnost razlika d_i