

Princípy dátovej vedy

Informácie o skúške

(snáď) finálna verzia, 10.1.2025

Termíny:

- sú uvedené v AISE
- prihlasovať/odhlasovať sa dá vždy do 9:00 predchádzajúceho dňa

Miestnosť:

- je uvedená v AISE

Harmonogram:

Na každom termíne môže byť maximálne 8 študentov, skúšať sa bude podľa harmonogramu nižšie. Môže sa stať, že počas dňa vzniknú časové sklzy.

č. študenta	príprava	odpoveď
1	8:50-9:30	9:30-10:10
2	9:30-10:10	10:10-10:50
3	10:10-10:50	10:50-11:30
prestávka		
4	12:20-13:00	13:00-13:40
5	13:00-13:40	13:40-14:20
6	13:40-14:20	14:20-15:00
7	14:20-15:00	15:00-15:40
8	15:00-15:40	15:40-16:20

Medzi prihlásenými študentami si, prosím, dohodnite poradie. Ak sa nedohodnete, budem predpokladať poradie podľa abecedy. Zoznam prihlásených študentov vždy všetkým prihláseným pošlem mailom po tom, ako sa uzavrie prihlasovanie (teda doobeda v deň pred skúškou).

Odpoveď pozostáva z:

1. Vylosovanie dvoch otázok a príprava na odpoveď (40 minút),
 - 2-krát: (teoretická téma z prednášok + konkrétna otázka k nej), vid' poznámky nižšie.
 - Typicky počas odpovede predchádzajúceho študenta.
 - Príprava prebieha v tej istej miestnosti ako skúška.
2. Samotná odpoveď (40 minút)
 - (a) Krátka diskusia o projekte (vid' http://www.iam.fmph.uniba.sk/ospm/Rosa/PDV/PDV_projekt.pdf)
 - spýtam sa vás na nejakú konkrétnu časť projektu a na váš príspevok do projektu

- treba vedieť, čo ste ako tím v projekte spravili (a dobre tomu rozumieť)
- treba vedieť povedať, čo ste na projekte robili vy

(b) Odpoveď na vylosované otázky.

Poznámky:

- Základné otázky (témy) budú zo zoznamu nižšie, ale ku každej téme je navyše priradená jedna konkrétna otázka.
 - Príklad toho, ako približne môže vyzeráť jedna vylosovaná otázka:
Lineárna regresia (vrátane vektorového/maticového zápisu).
Máme “triviálny model” $wa_i = \beta_0 + \varepsilon_i$, kde dáta sú váhy 5 ľudí. Ako by podľa vás mal vyzeráť odhad $\hat{\beta}_0$? Vedeli by ste vzorček pre $\hat{\beta}_0$ aj odvodiť?
- Pri každej metóde je potrebné vedieť aj aký je jej cieľ/použitie a aké sú jej hlavné výhody (prípadne nevýhody, ak boli spomínané).
- Budem sa pýtať aj dopĺňujúce otázky.

Zoznam tém (pracovná verzia; finálna bude zverejnená na konci semestra):

Nezabudnite, že ku každej téme bude jedna konkrétna dopĺňujúca otázka v zadaní!

1. Exploratívna analýza: vykreslenia, deskriptívne štatistiky.
2. PCA.
3. Lineárna regresia (vrátane vektorového/maticového zápisu).
4. Lineárna regresia: riešenie “maticovo” a “gradientne”, transformácie premenných, kategorické x -ká.
5. Predikčná chyba, bias-variance tradeoff, meranie kvality predikčných modelov (validácia, krosvalidácia a ďalšie).
6. Špecifiká kategorických premenných: kategorické x -ká, kategorické y -ká, ordinálne dáta; miery kvality klasifikačných modelov.
7. Penalizovaná regresia (hrebeňová a lasso).
8. Princípy inferenčnej štatistiky, štatistika v lineárnej regresii.
9. Bootstrap.
10. k -NN.
11. Logistická regresia.
12. Neurónové siete
13. Rozhodovacie stromy a sady stromov.
14. Supervised a unsupervised learning. Zhlukovanie: k -means a k -medoids.

15. Hierarchické zhľukovanie.
16. Diskrétne náhodné premenné, ich základné rozdelenia (alternatívne, binomické, Poissonovo). Podmienená pravdepodobnosť, Bayesova veta.
17. Spojité náhodné premenné, hustota. Normálne a Paretovo rozdelenie. Centrálna limitná veta.
18. Bayesovské siete.
19. Naivný bayesovský klasifikátor, Markovovské modely.
20. Skryté markovovské modely, Viterbiho algoritmus.
21. Štatistické testy: o stredných hodnotách (stredoch rozdelení), robustné testy, permutačný test, testy závislosti (pre kvantitatívne a kategorické dáta).
22. Big data a práca s nimi.