Otázky k magisterské státní závěrečné zkoušce z předmětu Vytěžování znalostí z dat

Obor: Datová věda

Akademický rok 2023/2024

1. **Principy datové vědy** (Data-driven vs. Data-informed přístup, taxonomie datové vědy, slovník datové vědy, životní cyklus datového projektu; Aplikace data science v obchodním rozhodování, VVVV, vytváření datového týmu v organizaci, měření úspešnosti datových projektů)
2. **Data nominální, ordinální a kvantitativní. Popisné charakteristiky** (polohy, variability, tvaru rozdělení, kvantily), typy úloh a příklady užití. Míry heterogenity (entropie, Gini) a užití. Modely rozdělení pravděpodobností, parametry, příklady (rozdělení binomické, Poissonovo, rovnoměrné diskrétní, rovnoměrné spojité, normální).
3. **Statistická inference jako rozhodovací problém** (Vysvětlení principu statistické inference, statistické hypotézy, postupu klasických testů hypotéz (rozdělení testových kritérií, rozhodovací pravidla, chyba I. a II. druhu). Příklady statistických hypotéz a testů. Rozhodovací rizika při současném testování více než dvou hypotéz, FDR (False Discovery Rate), FWER (Family-Wise Error Rate)).
4. **Závislosti a vztahy dvou a více kvalitativních nebo kvantitativních veličin** (Asociace dvou kvalitativních znaků, sdružená a marginální pravděpodobnost, kontingenční tabulka, hypotézy. Vícerozměrný lineární regresní model, metoda MNČ, předpoklady, charakteristiky kvality modelu, hypotézy v regresním modelu, předpovědní interval, rizika modelu).
5. **Kroky a komponenty průzkumové analýzy dat (explorace)** (Data, příprava pro statistickou analýzu. Typy proměnných, veličina vysvětlující a vysvětlovaná, náhodná a nenáhodná, měrné stupnice, transformace dat, seskupování hodnot, segmentace. Možnosti statistického popisu dat při jedné nebo více dimenzích (statistiky, grafy). Asociace a závislosti pro vícerozměrná data).
6. **Statistické přístupy k analýze dat a vytváření modelů na základě dat (statistical learning), metodologie CRISP** (Modely klasifikace: Lineární diskriminační analýza (předpoklady, princip řešení) a rozhodovací strom (metoda CART, případně CHAID). Popis výsledků obou typů modelů, rozhodovací pravidla, křížová validace, klasifikační tabulka).
7. **Modelování a simulace** (Náhodná čísla a náhodné číslice. Očekávané statistické vlastnosti generátorů pseudonáhodných čísel. Metody simulace náhodných veličin (rozdělení spojité rovnoměrné, normální, empirické nespojité). Metoda bootstrap (princip a příklad užití). Model dynamického procesu s diskrétními stavy Markovovým řetězcem (popis, aplikace). Význam a užití statistických modelů).
8. **Vizualizace dat** (Cíle a důvody vizualizace dat, metodiky postupu při vizualizací, oblasti použití a příklady nasazení vizualizačních metod, řetězec zpracování dat vedoucí k jejich vizualizaci).
9. **Formy vizualizace dat** (Typy vizualizací a grafů, srovnat výhody a nevýhody, volba typu grafu, náležitosti(komponenty) vizualizací, soustavy souřadnic, statické a dynamické vizualizace).
10. **Vizualizační nástroje** (Sw prostředky pro tvorbu vizualizací, příklady a srovnání sw nástrojů, programovací jazyky pro zpracování a vizualizaci dat, sw knihovny, příklady řešených úloh).
11. **Vizuální atributy a percepce** (Typy vizuálních a vlastnosti vizuálních atributů, způsoby mapování dat na vizuální atributy, zavedené způsoby a standardy znázornění, problémy a omezení při vizualizaci).
12. **Vizualizační chyby a lži** (Vysvětlení pojmu, důvody vzniku vizualizačních lží, metriky hodnocení, zásady eliminace, konkrétní příklady chyb).

13. **Klasifikace dokumentů** (význam klasifikace, algoritmy, vyhodnocení výsledku klasifikace, nástroje) a jejich **shlukování** (význam shlukování, druhy podobnosti textových dokumentů, způsoby určování podobnosti dokumentů, kategorizace shlukovačích metod a jejich algoritmy, vyhodnocení výsledku shlukování, nástroje).
14. **Metoda Monte Carlo a MCMC algoritmus** (Podstata metody Monte Carlo, výpočet plošného obsahu, výpočet střední hodnoty, výpočet pravděpodobnosti, odhad chyb výpočtu, simulace Markovových řetězců, algoritmus MCMC a jeho použití).
15. **Skryté Markovovy řetězce** (Koncept skrytého Markovova řetězce, předpoklady tohoto modelu a jeho aplikace, základní úlohy pro skryté Markovovy řetězce, metody jejich řešení).
16. **Bayesovská statistika** (Četnostní a Bayesovký přístup k pravděpodobnosti, podmíněná pravděpodobnost, Bayesova věta a její použití, apriorní a posteriorní pravděpodobnost, apriorní a posteriiorní rozdělení, věrohodnostní funkce, konjugované třídy rozdělení, Bayesovské odhady, Bayesovské testování hypotéz).