Otázky k magisterské státní závěrečné zkoušce z předmětu
Vytěžování znalostí z dat
Obor: Datová věda
Akademický rok 2024/2025

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, vytváření datového týmu v organizaci, měření úspešnosti datových projektů)
2. **Data** strukturovaná, nestrukturovaná, semi-strukturovaná**. Big Data** (typické vlastnosti analytické a programátorské, Volume, Velocity, Variety). Typy dat, rozlišení v oblasti datové vědy (kvalitativní a kvantitativní, nominální, ordinální, data spojitá a nespojitá, alfanumerická). Reprezentace různých typů v datovém souboru a počítači (příklady).
3. **Statistická inference jako rozhodovací problém** (Statistické zobecňování, hypotézy, postup klasických testů hypotéz, rozdělení testových kritérií, rozhodovací pravidla, chyba I. a II. druhu). Příklady statistických hypotéz a testů. Rizika současného testování více než dvou hypotéz (False Discovery Rate, Family-Wise Error Rate) a specifika (big data).
4. **Závislosti a vztahy dvou nebo 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. K**roky a komponenty průzkumové analýzy dat**. Uspořádání dat (file: flat, wide, long). Popis dat při jedné nebo více dimenzích (statistiky, grafy). Asociace a závislosti, veličina vysvětlující a vysvětlovaná, náhodná a nenáhodná, transformace dat, seskupování hodnot, segmentace. **Explorace dat** (míry polohy, variability, tvaru rozdělení, četnosti, kvantily, vizualizace), míry heterogenity (entropie, Gini), typy úloh, příklady užití.
6. **Statistické přístupy k analýze dat a vytváření modelů na základě dat** (statistical learning), metodologie CRISP. Modely klasifikace (Lineární diskriminační analýza, rozhodovací stromy CART nebo CHAID, rozhodovací pravidla, křížová validace, klasifikační tabulka). Data v časové řadě, popisné charakteristiky, vlastnosti, model a validace.
7. **Modelování a simulace** (Náhodná čísla a číslice, očekávané statistické vlastnosti generovaných pseudonáhodných čísel. Metody simulace náhodných veličin (rozdělení spojité rovnoměrné, normální, empirické nespojité). Metoda bootstrap (princip, užití). Dynamický proces s diskrétními stavy, model Markovovým řetězcem (popis, aplikace). Význam a užití.
8. **Vizualizace dat** (Cíle a význam vizualizace dat, vizualizační metody, oblasti užití, příklady nasazení vizualizačních metod, řetězec kroků vedoucích k vizualizaci dat). Formy a typy vizualizací a grafů, volba typu grafu, náležitosti(komponenty) vizualizací, soustavy souřadnic, statické a dynamické vizualizace, výhody a nevýhody vizualizace.
9. **Vizualizační nástroje** (SW prostředky pro vizualizace, příklady a srovnání SW nástrojů, programovací jazyky pro zpracování a vizualizaci dat, knihovny, příklady řešených úloh).

10. **Vizuální atributy a percepce** (vnímání usuzování)**, rizika** (Typy a vlastnosti vizuálních atributů, způsoby mapování dat na vizuální atributy, zavedené standardy, problémy a omezení vizualizace), rizika vizualizace (vizualizační chyby a lži, vznik, metriky hodnocení, eliminace, příklady).
11. **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).
12. **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).
13. **Skryté Markovovy řetězce** (Koncept skrytého Markovova řetězce, předpoklady tohoto modelu a jeho aplikace, základní úlohy pro skryté Markovovy řetězce, metody jejich řešení).
14. **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).