Záróvizsga tematika

Mérnökinformatikus alapképzési szak BSc F tanterv

Specializáció: **Big Data és üzleti intelligencia** *(Korszerű adatbázisok, Bevezetés az adattudományba, Adattárházak és üzleti intelligencia, Big Data és felhő alapú szolgáltatások, Haladó adatelemzés)*

1. **Adattudomány és üzleti intelligencia:** Mi az adattudomány, és melyek a legfontosabb céljai? Ismertesse az adattudomány interdiszciplináris jellegét és főbb alkalmazási területeit. Mutassa be az adattudomány, az üzleti intelligencia, a mesterséges intelligencia és a gépi tanulás kapcsolatát, valamint azok szerepét az adatvezérelt döntéshozatalban.
2. **Adatbányászat és tudásfeltárás:** Mi az adatbányászat, és miért van szükség rá? Ismertesse a tudásfeltárás folyamatának főbb lépéseit és azon belül az adatbányászat szerepét. Mutassa be az adatbányászati feladatok főbb típusait és jellemzőit, gyakorlati példákkal illusztrálva.
3. **Adattípusok és leíró statisztika:** Ismertesse az adattípusokat, valamint az attribútum típusokat. Mutassa be az adatok alapvető statisztikai jellemzőit átlag, medián, módusz, kvantlilisek, adatok eloszlása.
4. **Adatok előfeldolgozása:** Miért szükséges az adatok előfeldolgozása, illetve melyek az előfeldolgozás fő feladatai? Mit értünk adattisztításon, miért van rá szükség, milyen módszerei vannak? Mi az adatintegráció, kell ügyelni a megvalósítása során? Mi az adatredukció és milyen módszereket ismer annak megvalósítására?
5. **Adatok összehasonlítása, adatvizualizáció:** Hogyan tudja mérni az attribútumok és az adatok hasonlóságát, illetve különbségét? Mit mér és mire használható a korreláció? Miért és mire jó az adatvizualizáció? Milyen adatvizualizációs technikákat ismer?
6. **Osztályozás döntési fákkal:** Ismertesse az osztályozás feladatát és annak folyamatát. Mutassa be részletesen, példán keresztül a döntési fa készítésének folyamatát. Döntési fa, döntési fa algoritmusok, attribútum-választási stratégiák, túlilleszkedés és alulilleszkedés, szabályok készítése a döntési fa alapján.
7. **Az osztályozás pontossága:** Ismertesse az osztályozó modellek értékelésének főbb mérőszámait a keveredési mátrix (confusion matrix) alapján: accuracy, error rate, sensitivity, specificity, precision, F1-score. Mutassa be a ROC-görbe és az AUC értelmezését. Ismertesse az adatok tanuló- és teszthalmazra osztásának módszereit, valamint az adatszivárgás (data leakage) problémáját és elkerülésének lehetőségeit.
8. **Az osztályozás alternatív módszerei:** Ismertesse röviden az osztályozás alternatív módszereit (szabályalapú osztályozók, legközelebbi szomszéd osztályozók, Bayes osztályozási módszerek, neurális hálók, tartóvektor gépek, együttes módszerek) és egy tetszőlegesen kiválasztott módszert mutasson be részletesen, példával illusztrálva. Nevezzen meg az osztályozáshoz kapcsolódó néhány gyakorlati alkalmazást.
9. **Regresszió és a regressziós modellek értékelése:** Ismertesse a regresszió feladatát és szerepét a felügyelt tanulásban. Miben különbözik a regresszió az osztályozástól? Mutassa be az egyváltozós és többváltozós lineáris regresszió, a polinomiális regresszió és a regressziós fák alapelvét. Ismertesse a legkisebb négyzetek módszerének szerepét a regressziós modellek illesztésében. Mutassa be a

regressziós modellek teljesítményének értékelésére használt mérőszámokat: átlagos abszolút hiba (MAE), átlagos négyzetes hiba (MSE), négyzetes középhiba (RMSE).

10. **Mesterséges neurális hálók:** Ismertesse a mesterséges neurális hálók felépítését és működését, a perceptron modelljét, valamint a súlyok, az eltolás (bias) és az aktivációs függvények szerepét. Mutassa be a leggyakoribb aktivációs függvényeket (sigmoid, softmax, ReLU, tanh), valamint az előrecsatolt és rekurrens neurális hálók jellemzőit. Ismertesse a mélytanulás fogalmát és a konvolúciós neurális hálók (CNN) működésének alapelvét.
11. **Neurális hálók tanítása és értékelése:** Ismertesse a neurális hálók tanításának folyamatát, a veszteségfüggvény, a gradienscsökkentés és a hiba-visszaterjesztés (backpropagation) szerepét. Mutassa be a tanulási ráta és a főbb optimalizáló algoritmusok (SGD, Adam, RMSprop) jelentőségét. Ismertesse az alulilleszkedés és túlilleszkedés problémáját, az ezek kezelésére alkalmazott módszereket, valamint a tanító-, validációs és teszthalmaz szerepét.
12. **Asszociációs elemzés:** Ismertesse az asszociációs elemzéshez használt alapfogalmakat (elemhalmaz, k-elemhalmaz, gyakori elemhalmaz, asszociációs szabály, asszociációs szabály gyakorisága és megbízhatósága). Ismertesse a gyakori elemhalmazok bányászatának folyamatát, azon belül az apriori algoritmus lényegét. Mutassa be az asszociációs szabályok feltárásának lépéseit. Nevezzen meg az asszociációs elemzéshez kapcsolódó néhány gyakorlati alkalmazást.
13. **Csoportosítás (klaszterezés):** Ismertesse a csoportosítás feladatát. Mutassa be röviden a csoportosítás módszereit (felosztó módszerek, hierarchikus módszerek, sűrűség-alapú módszerek) és egy tetszőlegesen kiválasztott módszert mutasson be részletesen, példával illusztrálva. Nevezzen meg a csoportosításhoz kapcsolódó néhány gyakorlati alkalmazást.
14. **Big Data rendszerek alapjai:** Ismertesse a Big Data fogalmát, kialakulásának okait és főbb alkalmazási területeit. Mutassa be a Big Data rendszerek legfontosabb jellemzőit a 4V modell (Volume, Velocity, Variety, Veracity) alapján. Ismertesse a Big Data főbb adatforrásait (gépek, emberek, szervezetek), valamint a nagy adatmennyiség tárolásával és feldolgozásával kapcsolatos kihívásokat. Mutassa be a felhőalapú számítástechnika (cloud computing) és a skálázhatóság szerepét a Big Data rendszerekben.
15. **Hadoop keretrendszer és MapReduce:** Ismertesse az elosztott Big Data rendszerek működésének alapelveit és a Hadoop keretrendszer főbb komponenseit. Mutassa be a Hadoop elosztott fájlrendszer (HDFS) felépítését és működését, a NameNode és DataNode szerepét, valamint az adattöbbszörözés jelentőségét a hibatűrés biztosításában. Ismertesse a YARN erőforrás-kezelési feladatait és a MapReduce programozási modell működését. Mutassa be a Map, Shuffle és Reduce lépések szerepét egy egyszerű példán keresztül.
16. **Big Data rendszerek elemzése és gépi tanulás:** Ismertesse a gépi tanulási feladatok megvalósításának főbb lépéseit Apache Spark és Databricks környezetben, az adatok előkészítésétől a modellek értékeléséig. Mutassa be az osztályozás és a klaszterezés megvalósítását döntési fa és k-közép (K-means) algoritmus alkalmazásával. Ismertesse az MLflow szerepét a gépi tanulási kísérletek, hiperparaméterek és modellek nyomon követésében és összehasonlításában.
17. **Batch processing Apache Spark keretrendszerrel:** Ismertesse a kötegelt adatfeldolgozás (batch processing) fogalmát, jellemzőit és alkalmazási területeit. Mutassa be az Apache Spark keretrendszer felépítését, működésének alapelvét és előnyeit a Hadoop MapReduce megoldáshoz képest. Ismertesse az RDD (Resilient

Distributed Dataset) fogalmát és tulajdonságait, a transzformációk (transformations) és akciók (actions) közötti különbséget, valamint a lusta kiértékelés (lazy evaluation) és a DAG szerepét. Mutassa be az RDD, DataFrame és Dataset adatstruktúrák jellemzőit és alkalmazásukat.

18. **Adatfolyam-feldolgozás (Stream Processing):** Ismertesse az adatfolyam-feldolgozás fogalmát, jellemzőit és alkalmazási területeit. Hasonlítsa össze a kötegelt, adatfolyam-alapú és mikro-kötegelt feldolgozást. Mutassa be az ablakalapú feldolgozás (windowing) lényegét és főbb típusait (tumbling, hopping, sliding), valamint az állapotkezelés (state management) és az ellenőrzőpontok (checkpointing) szerepét.
19. **Databricks platform és Delta Lake:** Ismertesse a Databricks platform szerepét a Big Data rendszerekben, valamint a Lakehouse architektúra alapelvét és előnyeit. Mutassa be a Unity Catalog szerepét, objektumhierarchiáját (Catalog, Schema, Table/Volume) és az adatok kezelésében betöltött feladatait. Ismertesse a Delta Lake tárolási megoldás működését, a Delta táblák felépítését, a táblatörténet (table history) és az időutazás (time travel) lehetőségeit, valamint a Parquet fájlok és a tranzakciós napló (Delta Log) szerepét.
20. **Idősorok elemzése és előrejelzése:** Ismertesse az idősoros adatok fogalmát, jellemzőit és legfontosabb alkalmazási területeit. Mutassa be az idősorelemzés főbb feladatait, valamint az idősorok komponenseit: trend, szezonalitás, ciklikus és véletlen ingadozások. Ismertesse az idősorok leíró elemzésének és előrejelzésének alapelveit, valamint az egyedi és több idősor elemzésének lehetőségeit. Mutassa be az idősor-előrejelzés módszereit (egyszerű módszerek, exponenciális simítás, ARIMA, Prophet) és alkalmazási lehetőségeit.
21. **Természetes nyelvfeldolgozás (NLP):** Ismertesse a természetes nyelvfeldolgozás fogalmát és alkalmazási területeit. Mutassa be a szöveges adatok előfeldolgozásának főbb lépéseit (tokenizálás, írásjelek és stopszavak eltávolítása), valamint a szövegek numerikus reprezentációjának módszereit (Bag of Words, dokumentum–kifejezés mátrix, TF-IDF). Ismertesse a szövegosztályozás folyamatát és a modellek értékelésének lehetőségeit.
22. **Gráfadatbázisok:** Ismertesse a gráfadatbázisok fogalmát, jellemzőit és főbb alkalmazási területeit. Mutassa be a gráfalapú adatmodell felépítését, a csomópontok (nodes), kapcsolatok (relationships) és tulajdonságok (properties) szerepét. Hasonlítsa össze a gráfadatbázisokat a relációs adatbázisokkal, különös tekintettel az adatok tárolására és a kapcsolatok kezelésére. Ismertesse a tulajdonsággráf (property graph) modell alapelveit, valamint a Neo4j gráfadatbázis-kezelő rendszer főbb jellemzőit. Ismertesse a gráfbejárás és az útvonalkeresés alapelveit, valamint a gráfadatbázisok alkalmazását például közösségi hálózatok, ajánlórendszerek és kapcsolatelemzési feladatok esetén.
23. **Adattárházi alapok, rendszerek, adatkockák és fizikai tárolás:** Hasonlítsa össze az OLTP és OLAP rendszereket, különös tekintettel a sor- és oszlopalapú tárolási modellekre és azok teljesítménybeli hatásaira. Ismertesse az adattárház fő jellemzőit Inmon definíciója szerint (témaközpontú, integrált, nem illékony, időbélyegzett). Ismertesse az adatkocka lényegét és legfontosabb műveleteit (Drill-down, Roll-up, Slice, Dice, Pivot). Egy cég a lassuló ügyfélkiszolgálás ellenére a komplex vezetői elemzéseket és adatkocka-műveleteket közvetlenül a soralapú éles (OLTP) adatbázison futtatja. Mutasson rá, miért hibás ez a megközelítés a soralapú tárolás és a rendszercélok miatt. Vázolja fel, hogyan oldaná meg a problémát egy különálló OLAP réteggel!

24. **Adattárházi architektúrák:** Ismertesse és vesse össze az Inmon-, Kimball- és hibrid adattárház-építési stratégiákat. Mutassa be a hagyományos adattárházi rétegeket (Staging, Core DWH, Data Mart) és a modern Medallion architektúrát (Bronze, Silver, Gold), kitérve azok párhuzamaira. Vázoljon fel egy ERP és CRM adatokat integráló Medallion-alapú architektúrát! Ismertesse az adat útját az operatív tehermentesítéstől az egységes riportolásig (Single Source of Truth).
25. **Dimenziós modellezés I. – Tények, dimenziók és sémák:** Ismertesse a Kimball-féle négylépéses dimenziós tervezést, a szemcsézettség (granularitás) jelentőségét és a helyettesítő kulcsok szerepét! Hasonlítsa össze a csillag-, hópehely- és galaxissémát! Mutassa be a tények additivitási típusait (különös tekintettel a félig additív mutatókra). Ismertesse egy hópehelyséma csillagsémává alakításának lépéseit Power Query-ben! Magyarázza el, miért hibás a havi zárókészlet egyszerű összegzése (SUM) az idődimenzió mentén, és hogyan kell helyesen kezelni.
26. **Dimenziós modellezés II. – Ténytáblatípusok és változó dimenziók:** Mutassa be a tranzakciós, periodikus és kumulatív pillanatkép-ténytáblákat, valamint a degenerált, junk és szerepjáték dimenziókat (valamint a hídtábla és ténymentes ténytábla szerepét)! Ismertesse az SCD 0, 1, 2, 3 és 6 típusú kezelési stratégiáit! Tervezzen meg egy olyan modellt, amely az értékesítéseket az üzletkötők korábbi és aktuális régióbeosztása szerint is elemzi (SCD2 + szerepjáték). Mutassa be az aktív/inaktív dátumkapcsolatok működését és a USERELATIONSHIP függvény alkalmazását Power BI-ban!
27. **Adatintegráció, adattisztítás és inkrementális betöltés:** Hasonlítsa össze az ETL és ELT megközelítéseket, a teljes és inkrementális betöltés stratégiáit, valamint a CDC (Change Data Capture) működését! Mutassa be az adatminőség-ellenőrzés és tisztítás lépéseit Python és Power Query környezetben! Egy webshop nagy forgalmú tranzakciós tábláját (napi ~1 millió sor) kell inkrementális módon betölteni a céges adatbázisba. A forrásból érkező delta adatokban időnként duplikációk, formátumhibás dátumok, valamint hiányzó értékek fordulnak elő. Válasszon architektúrát, határozza meg a CDC alapú változáskövetés logikáját, és vázolja fel, hogyan szűrné ki és tisztítaná meg ezeket a hibás rekordokat Power Query vagy Python (Pandas) segítségével a célbázisra való betöltés előtt!
28. **Power BI adatmodellezés, DAX és üzleti vizualizáció:** Mutassa be a Power BI fő komponenseit, a Power Query és a csillagséma jelentőségét az adatmodellben! Hasonlítsa össze a kalkulált oszlopokat és a DAX-mértékeket, valamint a sor- és szűrőkontextust! Magyarázza el, mikor kell SUM vagy SUMX függvényt használni, és hogyan optimalizálható egy túl sok kalkulált oszlopot tartalmazó, nehéz modell. Ismertesse a dátumtábla feltételeit és a helyes időszaki összehasonlítások elveit. Vázolja fel a hatékony vezetői dashboardok tervezési alapelveit!