2025 春 数据分析及实践 期末考试

一、单选题（共 30 分）

1. 网络爬虫不包括（ ）
   A. 载入过程
   B. 解析过程
   C. 检索过程
   D. 存储过程

2. 某时间段抽取 10 位顾客的购买记录：购买物品 A、B、C 的分别有 2、3、5 人。求信息熵（ ）
   A. 5.06
   B. 18.36
   C. 1.03
   D. 1.49

3. 哪一种数据变换方法把数据缩放到 $[0,1]$ 区间（ ）
   A. Z - score 标准化
   B. 最小 - 最大规范化
   C. 独热编码
   D. 数据离散化

4. 关于假设检验，下列说法正确的是（ ）
   A. 假设检验的目的是接受原假设
   B. 第一类错误指拒绝了错误的原假设
   C. 假设检验中，犯第一类错误的错误率即为置信度
   D. 大数据分析不涉及假设检验

5. 下列哪项不属于 NoSQL（数据库）的特点（ ）
   A. 数据模型简单
   B. 数据有高度一致性
   C. 灵活性强
   D. 高性能

6. 关于 TF - IDF 说法正确的是（ ）
   A. TF 是逆文档频率
   B. IDF 用于衡量词语在单个文档中的重要性
   C. TF - IDF 可用于提取文档关键词
   D. TF - IDF 算法复杂，效率低

7. 关于数据分布，说法错误的是（ ）

A. 集中趋势反映了一组数据中心点位置，以及该组数据向中间靠拢或聚集水平。变异系数是常用指标

B. 数据离散程度增大，集中趋势的测度值对该组数据的代表性越差，反之亦然

C. 在数值型数据中，刻画数据围绕其中心位置附近分布数字特征时，常用方差和标准差

D. 若极差或四分位差较大，建模时需考虑数据是否有长尾现象

8. 某医院进行病症诊断，某病诊断出 120 例病例。后续确诊过程中，发现只有 80 例真正患病，其余 40 例是误诊（假阳性），则该诊断方法的正确率（Precision）为（ ）；假设样本中仍有 120 例未被诊断（漏诊，假阴性），则该诊断方法的查全率（Recall）为（ ）

A. 66.7%　40%

B. 33.3%　60%

C. 66.7%　60%

D. 33.3%　40%

9. 哪些指标属于不确定性时序预测评价指标（ ）

A. CRPS

B. MSE

C. RMSE

D. MAE

10. 哪个数据挖掘算法是最为代表的符号主义流派（ ）

A. 感知机

B. 支持向量机

C. 决策树

D. 关联规则

二、简答题（共 20 分）

1. 请简述 3 种数据预处理方法，并说明为什么要进行数据预处理。

2. 请简述特征工程的意义以及主要流程（步骤）。

3. 请简述如何进行假设检验，并说明假设检验和参数估计的区别。

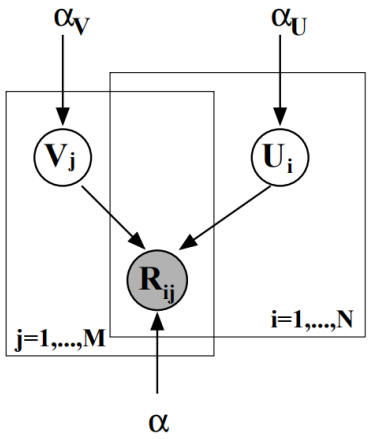
4. 请简述 ROC 的绘制方法，并说明当 AUC 为 0.5 和 1 时分别代表什么。

5. 请简述 K 近邻（KNN）算法的基本原理，并说明其为什么被称为“非参数方法”。

三、计算题（共 50 分）

1. 给定两个 5 维数据点 $x_1 = (1\ 1\ 0\ 1\ 0)$，$x_2 = (0\ 1\ 1\ 0\ 1)$，请依次计算 Jaccard 相似度、Cosine 相似度、Euclidean 距离、Pearson Correlation 值。

2. 概率矩阵分解方法可用于预测用户对未评分项目的评分。对每个用户 i 和每个项目 j 都可以通过潜在因子矩阵 U 与 V 表示，用户项目评分矩阵 R 可以近似建模为 $R_{ij} \approx U_i^T V_j$，如图所示。



假设 (R_{ij} 服从高斯分布，方差为 σ^2，每个评分相互独立。用户潜在因子 U_i 和项目潜在因子 V_j 服从均值为 0、方差分别为 σ_U^2 和 σ_V^2 的高斯分布。用最大后验概率估计参数 U 和 V，即最大化 $p(U, V|R, \sigma^2, \sigma_U^2, \sigma_V^2)$，写出优化目标公式即可，不用求导计算。

3. 以下是某商店的交易清单。请使用 Apriori 算法，以支持值阈值 33.34%、置信度阈值 60%，详细记录算法的执行过程。列出每次数据库扫描的候选项集和频繁项集，列出所有最终的频繁项集，生成所有的关联规则，标出其中的强关联规则，并且按照置信度排序。

交易 ID	物品
T1	中性笔、笔记本、荧光笔
T2	中性笔、笔记本
T3	中性笔、矿泉水、巧克力
T4	巧克力、矿泉水
T5	巧克力、荧光笔
T6	中性笔、矿泉水、巧克力

4. 滑雪是指利用滑雪板在雪地滑行的一种活动，最初是为了便于在冬季的雪地中出行，后来逐渐演变成一种冬季运动项目。已知两种属性：天气（晴天、雨天、雪天）和降雪量（≥ 50、< 50）。小明 8 天的训练集如下：

天气	晴天	雨天	雨天	雪天	雪天	晴天	雪天	雨天
降雪量	< 50	< 50	≥ 50	≥ 50	< 50	≥ 50	≥ 50	< 50
滑雪	否	否	否	是	否	是	是	是

(1) 计算训练集中“滑雪 = 是”和“滑雪 = 否”的先验概率；
(2) 计算每个属性在两类别下的条件分布；
(3) 请你帮助小明做出决策，使用贝叶斯分类器决策。

序号	天气	降雪量
A	晴天	≥ 50
B	雨天	< 50
C	雪天	< 50

5. 某电网过去 6 小时的负荷（单位：MW）如下图：

小时	1	2	3	4	5	6
负荷	500	504	509	515	520	528

使用 ARIMA(1, 2, 1) 预测第 7 个小时的负荷。其中，AR(1) 系数为 $\Phi_1 = 0.5$，MA(1) 系数为 $\theta_1 = -0.4$，初始残差 $\epsilon_1 = 0$，$\epsilon_2 = 0$，$\epsilon_3 = 0$。