**24-25（1）《数据科学导论》复习提纲**

## 一、准备

1、数据科学

2、两个核心任务：从数据中洞见真知，基于数据驱动的决策

3、数据科学的基本流程

## 二、探索式数据分析

1、数值型、类别型变量

2、数据清洗：缺失值处理、离群值发现与处理、离散化、编码和转换、归一化（Min-max 归一化（规范化）、Z-score 归一化（标准化）、十位数归一化）、等深直方图与等宽直方图

3、数据集成：实体相似度、基于集合的相似度、编辑距离

4、可视化与数据探索的基本方法

## 三、机器学习

1、一些基本概念：输入输出、模型、实际值预测值，估计预测

2、数据分析建模 4 大步骤：模型选择、目标函数选择、优化目标函数、评估模型

常数模型、平均绝对误差、平均平方误差

3、分类 KNN、决策树，熵、条件熵、信息增益，决策树构建算法流程（特征选择）

4、分类算法评价 accuracy, confusion matrix , precision, recall, f1, train set, test set, K-fold cross validation

5、聚类算法 K-means 迭代过程，利用肘法选择参数 K

6、聚类算法 GMM、EM 算法

Expectation Step 的目的（参数估计→更新软分配）、输入、输出、处理过程

Maximization step 的目的（软分配→更新参数估计）、输入、输出、处理过程

7、多元线性回归，设计矩阵、平均平方误差 MSE

线性回归的矩阵形式、张成空间、矩阵形式的解析解

线性回归的代数形式、针对各个变量的梯度计算、梯度下降算法

目标函数的矩阵形式、矩阵形式的梯度计算、梯度下降算法

梯度下降、随机梯度下降、小批量梯度下降的区别

8、分类算法 SVM、硬间隔软间隔、~~核函数技巧、~~铰链损失函数 hinge loss 与梯度下降算法

~~9、线性分类器（感知机线性分类器、logistic 回归、交叉熵损失函数）~~

## 四、文本分析

1、中文文本分词：基于规则的分词、HMM 算法

2、文本表示：独热编码、绝对词频、相对词频、TF-IDF

3、文本的降维表示：LSI（基于 SVD 的矩阵分解、降维、词项

的表示、文档的表示）

4、文本分类、独立假设位置无关假设、朴素贝叶斯算法（平滑技术）

5、~~文本相似度、向量夹角余弦相似度~~

## 五、图数据分析

1、基本概念：图、节点、边、无向图、有向图、weighted graph、unweighted graph

2、图的节点的重要度

Degree centrality、Closeness centrality、Betweenness centrality

Page rank 算法（A 矩阵、M 矩阵、迭代过程）, damping factor 的引入, dangling node 的处理

3、社区检测

模块度 Q 计算：基本公式，简化公式

模块度变化量ΔQ 计算

Louvain 算法

phase1：ΔQ（old community →i）, ΔQ（i→new community）

phase 2：每个社区缩减为一个超级节点，以及正确标注节点 degree 以及 edge weight

4、影响力最大化 Influence Maximization

IC 传播模型

Degree discount 算法