# Midterm of Data Mining

Jason He

October 2019

## 1 题型

1. 判断正误，尽力解释
2. 像第一题的作业题
3. 给数据集, 问 measures or coefficient to use
4.gini value and cross-entropy index
5.overfiting
6. 给 dataset，问用什么方法 approach, 优缺点
7. 计算
8.association analysis

## 2 Ch1

1.**P36** 5 steps: Input data -> preprocessing -> data mining -> post-processing ->information
2. **P37** Differentiate whether an example is data mining or not. 给个例子区分
3. **P55** What is not a cluster Analysis?

## 3 Ch2

**1.P6-7** 4 types of attributes: nominal, ordinal, interval, ratio
**2.P14** definition of asymmetric attribute.
**3.P17** types of data set.

**4.** Noise, Outliers and how to handle missing value

**5.** Preprocessing.

**6.P72 Similarity measure** 欧式，闵氏距离（会给公式）, Manhalanobis distance 会给协方差矩阵

**7.P85** SMC, Jaccard, cosine, extended Jaccard

**8.P89** correlation, 考试会给数据 (standardize before calculate) , drawback: only consider the linear relationship.

## 4 Ch3

**1. Mean Value** harmonic mean <= geometric mean <= arithmetic mean <= quadratic mean. 区别和使用场合

## 5 Ch4

**1.Decision Tree** How to split subtrees: binary vs.multi-way

**2.GINI value and cross-entropy** gini value for a multi-way split is always more than a binary split on the same attribute. **P29** 会计算数值，自带计算器 **P44**advantage of decision tree

## 6 Ch5

**1.P6 Rule Based** calculate coverage and accuracy of a rule 计算

**2.P8** exclusive v.s. exhaustive rules

**3.P11** more than 1 rule, on rule -> default class

**4.P16-17** 理解

**5.algo for rule growing** direct method: ripper 25-27

**6.indirect method**

**7.P32** rule-based 特点

**8.P44** KNN 的特点 + 理解 KNN 的算法

**9.ensemble method** 理解基本原理： 为什么要用？why it works？ 不考 bagging boosting,adaboost

**10.P73 Imbalanced problem**: generate a confusion matrix, calculate accuracy rate, precision, recall, F-measure **11.ROC** 只需要理解如何读 roc 图 **12.P82-83** handling class imbalance problem, cost matrix 如何计算 cost

## 7 Ch6

**1.P11** 朴素贝叶斯计算
**2.P14** 区分 3 种 error
**3.P26,28-29** calculate optimistic and pessimistic error
**4. validation set** drawback

**5.P59 SVM** characteristic of SVM

## 8 Ch7

**1.P4-5 association analysis** 如何计算 support count, support, confidence
**2.P7, 9 computational complexity** two-step. P14, 理解 P23-24 ，找 closed itemset， maximal frequent, P37 理解 cross-support patterns, P39 H-confidence, P44 理解 statistical independence

## 9 Ch8

**1.P6 categorical attribute**