# 机器学习导论

## 习题一

## Problem 1

若数据包含噪声，则假设空间中有可能不存在与所有训练样本都一致的假设，此时的版本空间是什么？在此情形下，试设计一种归纳偏好用于假设选择。

## Problem 2

对于有限样例，请证明

$$\mathrm{AUC} = \frac{1}{m^+m^-}\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\mathbb{I}(f(x^+) > f(x^-)) + \frac{1}{2}\mathbb{I}(f(x^+) = f(x^-))\right)$$

## Problem 3

在某个西瓜分类任务的验证集中，共有10个示例，其中有3个类别标记为“1”，表示该示例是好瓜；有7个类别标记为“0”，表示该示例不是好瓜。由于学习方法能力有限，我们只能产生在验证集上精度(accuracy)为0.8的分类器。

(a) 如果想要在验证集上得到最佳查准率(precision)，该分类器应该作出何种预测？

此时的查全率(recall)和F1分别是多少？

(b) 如果想要在验证集上得到最佳查全率(recall)，该分类器应该作出何种预测？

此时的查准率(precision)和F1分别是多少？

## Problem 4

在数据集$D_1, D_2, D_3, D_4, D_5$运行了$A, B, C, D, E$五种算法，算法比较序值表如表1所示:

表 1: 算法比较序值表

| 数据集 | 算法$A$ | 算法$B$ | 算法$C$ | 算法$D$ | 算法$E$ |
|---|---|---|---|---|---|
| $D_1$ | 2 | 3 | 1 | 5 | 4 |
| $D_2$ | 5 | 4 | 2 | 3 | 1 |
| $D_3$ | 4 | 5 | 1 | 2 | 3 |
| $D_4$ | 2 | 3 | 1 | 5 | 4 |
| $D_5$ | 3 | 4 | 1 | 5 | 2 |
| 平均序值 | 3.2 | 3.8 | 1.2 | 4 | 2.8 |

使用Friedman检验($\alpha = 0.05$)判断这些算法是否性能都相同。若不相同，进行Nemenyi后续检验($\alpha = 0.05$)，并说明性能最好的算法与哪些算法有显著差别。