Homework1

1、 数据集包含 100 个样本，其中正、反例各一半，假定学习算法所产生的模型是将新样本预测为训练样本数较多的类别(训练样本数相同时进行随机猜测)，试给出用 10 折交叉验证法和留一法分别对错误率进行评估的结果。

2、 令码长为 9，类别数为 4，试给出海明距离意义下理论最优的 一种 ECOC 二元码。

3、 假设某机器学习模型的原始类别和预测类别如下表所示，求它的混淆矩阵、准确率、精确率、召回率、F1 score。

样本序号	1	2	3	4	5	6	7	8	9	10
原始类别	1	1	1	-1	-1	-1	1	1	-1	1
预测类别	1	1	-1	-1	-1	1	-1	1	-1	1

4、 对以下数据集，构造 ID3 决策树，判断是否买房：

用户 ID	**年龄**	**性别**	**收入**	**是否买房**
1	27	男	15W	否
2	47	女	30W	是
3	32	男	12W	否
4	24	男	45W	是
5	45	男	30W	否
6	56	男	32W	是
7	31	男	15W	否
8	23	女	30W	是

注：年龄分为 20-30，30-40，40+三个阶段，收入分为 10-20，20-40，40+三个级别。

5、 判断下面说法是否正确：

(i) If a learning algorithm is suffering from high bias, adding more training examples will improve the test error significantly.

(ii) We always prefer models with high variance (over those with high bias) as they will able to better fit the training set.

(iii) A model with more parameters is more prone to overfitting and typically has higher variance.

(iv) Introducing regularization to the model always results in equal or better performance on the training set.

(v) Using a very large value of regularization parameter λ cannot hurt the performance of your hypothesis.