Howework #2

一、分类算法（5 分）

我们提供了真实的数据挖掘场景(个贷违约预测)的数据，见

train_100000.csv，要求根据该题目。请提交相应代码，可以用 scikit-learn 实现。

1. 首先应用第一次作业的缺省值填充方法。使用互信息对特征进行选择，选出与目标变量（is_default）最相关的前 20 个特征，并绘制每个特征的互信息。（1 分）
   互信息定义如下：

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \end{aligned}$$

   其中 H(X)表示 X 的熵，H(X|Y)表示条件熵，H(X, Y)表示联合熵。
2. 使用决策树、朴素贝叶斯和 AdaBoost 三种分类算法对数据集进行训练。使用 5 折交叉验证（5-fold cross-validation）进行模型评估。在每次交验证中，训练并评估模型，计算并报告 AUC 值。（3 分）
3. 比较三种模型的表现，找出最适合本数据集的算法。（1 分）

二、聚类算法（5 分）

1. 给定下列 12 个数据点：(1,4)； (2,5)； (3,3)； (4,4)； (2,3)； (3,2)； (6,6)； (7,5)； (6,4)； (8,6)； (7,7)； (8,5)
   使用 K-means 算法对它们聚类。令 k=2，初始中心点为 (2,4) 和 (7,6)，写出聚类过程。（2 分）

2. 我们提供了 twitter 的语料，在 twitter.txt 文件中。每一行表示一个 twitter 的推文。请使用任意一种编程语言，对该语料进行 K-means 聚类。请在聚类后给出每类的关键词，尝试不同的 k 值（k=2,3,4）进行分析。（3 分）

   提示:
   a.对语料进行去除停用词、分词等预处理，将每个推文表示成 tf-idf 向

量， 将 tf-idf 向量作为推文的表示进行聚类。
b.tf-idf 和 K-means 算法可以调用直接调用第三方的库。提交说明:需要提交源代码与报告。报告中简单说明 2)的实现思路，结果与分析。