# 机器学习第三章实验内容

## 一、实验内容

### 1.1 一元线性回归

使用Kaggle房价预测数据集：

1. 打乱数据顺序，取前70%的数据作为训练集，后30%的数据作为测试集
2. 分别以LotArea, BsmtUnfSF, GarageArea三种特征作为模型的输入，SalePrice作为模型的输出
3. 在训练集上，使用最小二乘法求解模型参数（需自己实现，不允许第三方库完成）
4. 计算三个模型在测试集上的MAE和RMSE这两种指标的大小（需自己实现，不允许第三方库完成）
5. 分别绘制模型的在训练集和测试集上的曲线
6. 选做：尝试去除训练集中的异常值或离群值后再次训练模型，绘制模型的预测曲线，观察模型在测试集上预测能力的变化

### 1.2 多元线性回归/对数线性回归（二选一）

1. 多元线性回归

使用多个特征作为输入完成房价预测问题，计算模型在十折交叉验证上MAE和RMSE的值，比较不同的特征组合在模型预测能力上的影响。

模型可使用sklearn.linear\_model.LinearRegression

选做：多项式回归（一元线性回归的扩展），尝试对部分特征进行变换，如将其二次幂，三次幂作为特征输入模型，观察模型在预测能力上的变化。

1. 对数线性回归

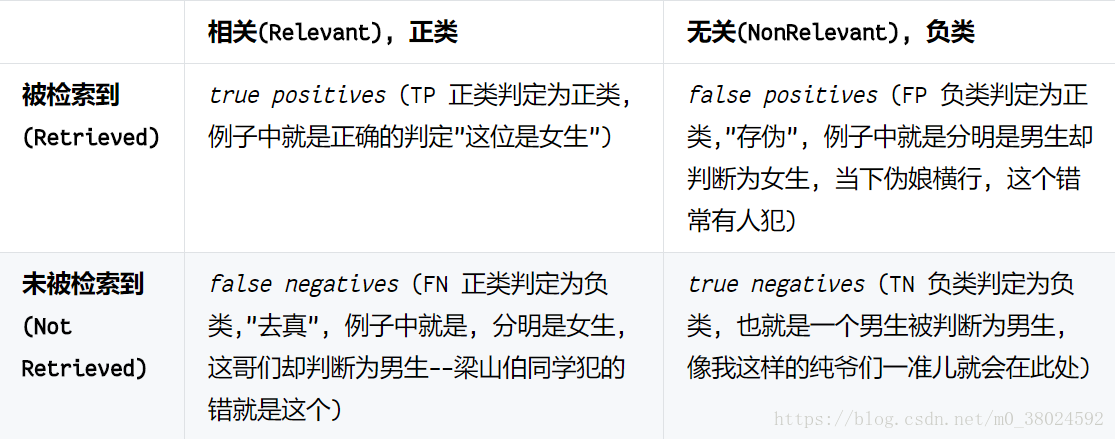
尝试使用对数线性回归完成房价预测问题，特征自选，比较模型十折交叉验证的MAE与RMSE。

模型可使用sklearn.linear\_model.LinearRegression

### 1.3 对数几率回归

1. 使用对数几率回归完成垃圾邮件分类问题和Dota2结果预测问题
2. 计算十折交叉验证下的精度(accuracy)，查准率(precision)，查全率(recall)，F1值。

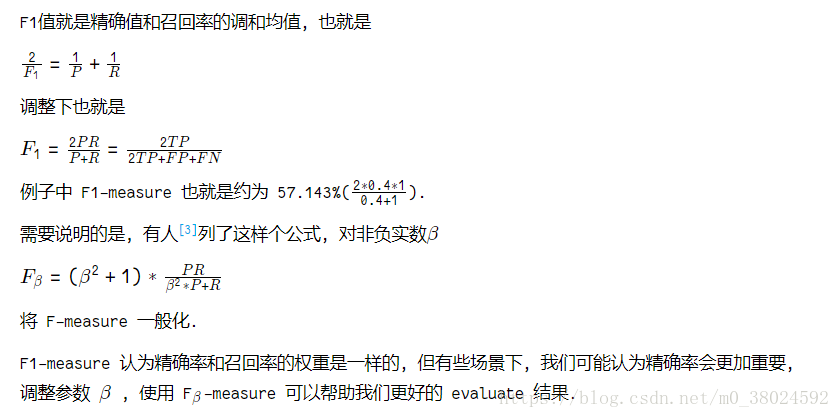
**精度(accuracy)**, 其定义是: 对于给定的测试数据集，分类器正确分类的样本数与总样本数之比.



**查准率(precision)**的公式是P =TP/(TP+FP)   **即 TP / 自选P， 这里自选P 是 自选出来的50个认为是“女生” -- 亦即 TP+FP** ​

**查全率(recall)**的公式是R = TP/(TP+FN)   **即 TP / 实际P， 这里实际P就是实际上女生的个数， -- 亦即 TP + FN，**它计算的是所有"正确被检索的结果(TP)"占所有"应该检索到的结果(TP+FN)"的比例.

**F1 = 查准率和查全率的调和均值**



1. 模型可使用sklearn.linear\_model.LogisticRegression

### 1.4 线性判别分析

1. 使用线性判别分析完成垃圾邮件分类问题和Dota2结果预测问题
2. 计算十折交叉验证下的精度(accuracy)，查准率(precision)，查全率(recall)，F1值。
3. 模型可使用sklearn.discriminant\_analysis.LinearDiscriminantAnalysis

### 1.5 综合题（选做）

任务要求：白葡萄酒质量共分为7个等级，对白葡萄酒质量进行预测。

1. 使用精度作为评价指标
2. 使用一元线性回归、多元线性回归、对数线性回归等线性回归模型对葡萄酒质量进行预测，计算其十折交叉验证的精度。
3. 使用对数几率回归、线性判别分析对葡萄酒质量进行预测，计算其十折交叉验证的精度。
4. 对比不同模型的预测能力。

## 二、数据介绍

### 2.1 kaggle房价预测数据集

文件名：

1. 原始数据：kaggle\_hourse\_price\_train.csv
2. 字段说明：kaggle房价预测字段说明.txt
3. 预处理脚本：preprocess.py
4. 处理后数据：kaggle房价预测处理后数据.csv

数据来源：<https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>

数据预处理步骤：

该数据集是真实数据集，含有大量的缺失值和噪声数据，提供的预处理脚本只对含缺失值的特征进行丢弃处理，学生可根据自己的需要进行其他方式的处理。

### 白葡萄酒质量数据集

文件名：

1. 原始数据：winequality-white.csv
2. 字段说明：葡萄酒字段说明.txt

数据来源：<http://archive.ics.uci.edu/ml/datasets/Wine+Quality>

数据无需预处理，读取即可使用

import pandas as pd

data = pd.read\_csv('data/wine\_quality/winequality-white.csv', delimiter=";")

### spambase垃圾邮件数据集

文件名：

1. 原始数据：spambase.data
2. 字段说明：spambase数据说明.txt

数据来源：<http://archive.ics.uci.edu/ml/datasets/spambase>

数据无需预处理，读取即可使用

import numpy as np

data = np.loadtxt('data/spambase/spambase.data', delimiter = ",")

### Dota2 Games Results Data Set Dota2游戏结果数据集

文件名：

1. 原始数据：dota2Train.csv
2. 字段说明：dota2比赛结果字段说明.txt

数据来源：<http://archive.ics.uci.edu/ml/datasets/Dota2+Games+Results>

数据无需预处理，读取即可使用

import numpy as np

data = np.loadtxt('data/dota2Dataset/dota2Train.csv', delimiter=',')