统计学知识：均值，中值，方差

居中趋势测量：均值、中值、众数

众数：在统计分布上具有明显集中趋势点的数值，代表数据的一般水平（众数可以不存在或多于一个）。 修正定义：是一组数据中出现次数最多的数值，叫众数，有时众数在一组数中有好几个。

数据的离散性：四分位距法、异常值、标准偏差、贝塞尔修正。

不能将测试数据用于训练。

混淆矩阵：

|  |  |  |
| --- | --- | --- |
|  | 识别为阳 | 识别为阴 |
| 阳 | True positive | False Negative |
| 阴 | False positive | True Negative |



ROC曲线：

|  |  |  |  |
| --- | --- | --- | --- |
|  | Predict Positive | Predict Negative |  |
| Positive | True Positive(TP) | False Negative(FN) | Actual Positive(TP + FN) |
| Negative | Fasle Positive(FP) | True Negative(TN) | Actual Negative(FP + TN) |
|  |  |  |  |

真正类率(True Positive Rate)：TPR = TP / (TP + FN)，也等于Positive的召回率。

负正类率(False Positive Rate)：FPR = FP / (FP + TN)。

真负类率(True Negative Rate)：TNR = TN / (FP + TN)，也等于Negative的召回率。

TPR又称为敏感度Sensitivity

TNR又称为特异度Specificity

FPR就是1-Specificity。

ROC曲线，横轴为FPR，竖轴为TPR。

ROC曲线的绘制：对于一个分类器，对于二分类问题，可以预测出一个样本为TP的概率，设置不同的阈值，就会得到不同的(FPR, TPR)。

<https://www.cnblogs.com/wuchuanying/p/6243987.html>

理想的分类器，FRP = 0，TPR = 1

AUC：ROC曲线下方的面积，理想的为1，随机情况（最坏的情形）为0.5，面积也可能小于0.5，即TPR的值一直小于FPR，识别正确的真阳性TP一直小于被误识别为阳性FN，但如果进行分类翻转，则会好于AUC为0.5的情形。

R2系数：

SST：总离差平方和，

SSE：残差平方和（误差平方和），



好的模型，R2的值应该接近于1，坏的模型，R2的值接近0.