machine learning 笔记

徐世桐

1 基础定义

二元分类：输出分类个数为 2

多元分类：输出分类个数不限

$one-versus-the-rest$ OvR：计算属于每一分类的可能性，取可能性最大的分类为输出分类

$one-versus-one$ OvO：对所有分类两两使用二元分类，每一分类器训练只需一部分数据

multilabel 多标签分类：目标检测，对一图像中的物体加 label

multioutput 多类分类：多标签分类，每一标签可包含多种信息

MSE $= \frac{1}{m}\sum_{i=1}^{m}(x^{(i)}-y^{(i)})^2$

learning schedule：根据迭代次数更新学习率

learning curves：观察模型是否有 over underfit

x 轴为使用的训练集大小，y 轴为 root MSE。

画出训练集 测试集在使用不同训练集大小后的 root MSE。

形状：

训练集的 root MSE 从 0 开始，使用的训练集增多后曲线平缓

测试集的 root MSE 从一高值开始，训练集增多后曲线平缓

分析：

当 2 曲线平缓值差值较大，测试集平缓值教低，则过拟合

当 2 曲线平缓值较高，则欠拟合

2 数学计算

pseudo inverse:

对矩阵 $X = USV^T$，pseudo inverse $X^+ = VS^+U^T$。S^+ 求法：

1. 对所有 S 元素，接近 0 的值赋为 0
2. 对所有非零元素取倒数
3. 取矩阵转置，得到 S^+

3 分析结果

confusion matrix 困惑矩阵：分析二元/多元分类

$$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$$

一行对应同一期望输出，一列对应同一计算输出

T/F: 此位置的计算输出是否和预计输出一致

P/N: 此位置的预计输出是否为真

precision $= \frac{TP}{TP+FP}$

即 $P($ 计算结果匹配 | 计算结果为正 $)$

recall $= \frac{TP}{TP+FN}$

即 $P($ 计算结果匹配 | 预计结果为正 $)$

$F_1 = \frac{2}{\frac{1}{precision}+\frac{1}{recall}}$

precision 和 recall 的调和平均值

specificity $= \frac{TN}{TN+FN}$

ROC curve: 分析二元/多元分类

y 轴 recall 值，x 轴 false positive rate$FPR = \frac{FN}{FN+TN} = \frac{FN}{1-specificity}$

期望的 ROC curve 为 recall 从 0 快速增长到 1。并保持直到 FPR 为 1。

即期望曲线下方面积接近 1