机器学习 Perceptron 作业

姓名： 张泽群 学号： 19049100002 班级：1班

Perf-1

答：此处划分的方法显然是**留出法**,直接将数据集D划分为两个互斥的集合,其中一个集合作为训练集S,另一个作为测试集T,即D= S ∪ T, S ∩ T=0.在S上训练出模型后,用T来评估其测试误差,作为对泛化误差的估计.

以二分类任务为例,假定D包含1000个样本,将其划分为S包含700个样本，T包含300个样本,需注意的是，训练/测试集的划分要尽可能保持数据分布的一致性, 避免因数据划分过程引入额外的偏差而对最终结果产生影响，例如在分类任务中至少要保持样本的类别比例相似.如果从采样的角度来看待数据集的划分过程,则保留类别比例的采样方式通常称为“分层采样”.

通过对D进行分层采样而获得含70%样本的训练集s和含30%样本的测试集T,若D包含500个正例、500个反例，则分层采样得到的S应包含350个正例、350个反例，而T则包含150个正例和150个反例。

则共有 $(C_{500}^{150})^2$ 种划分方式。

Perf-2

答：(1). 当概率阈值比所有预测值小的时候,此时我们假定所有样本均为正例，并且设n为样本总数

即 $TP = x$， $FP = y$， $x + y = n, FN = 0, TN = 0$,

则有 $TPR = \frac{TP}{TP+FN} = 1$, $FPR = \frac{FP}{TN+FP} = 1$

因此必过点（1，1）

(2). 当概率阈值比所有预测值小的时候,此时我们假定所有样本均为反例，并且设n为样本总数

即 $TN = x$， $FN = y$， $x + y = n, FP = 0, TP = 0$,

则有 $TPR = \frac{TP}{TP+FN} = 0$, $FPR = \frac{FP}{TN+FP} = 0$

因此必过点（0，0）

Perf-3

答:

(a). 在实际分类问题的建模过程中，经常会遇到数据集正负样本的样本量差距很大的情况，如果直接使用这种数据进行建模，将会对样本数据量大的的样本造成过拟合，也就是说预测偏向样本数较多的分类。这样就会大大降低模型的泛化能力。

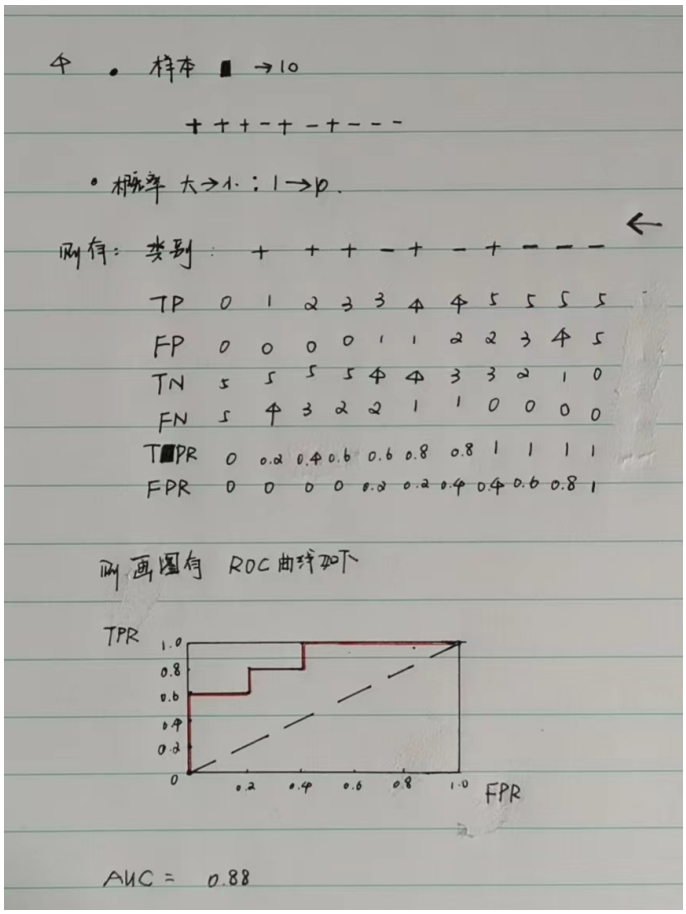
该学习机选用的正负样本非常不均衡，Accuracy=99.5%可能是因为对负样本的预测比较好，但对更感兴趣的正样本预测可能不好。例如，TP=0,TN=995,FP=5,FN=0,此时也有Accuracy=99.5%，但是正样本的精度为0%，因此学习性能不一定好。

因此该学习性能不一定准确，无法用于疾病筛查。

(b). 我认为用**RUC曲线和AUC** 或 **P-R曲线和AUC以及F1度量**能够反映出这个学习机在疾病筛查上的筛查性能。

Perf-4

答： AUC = 0.88



Perf-5

$Accuary = \frac{TP+TN}{TP+TN+FP+FN} = \frac{160}{200} = 80\%$

$Error\ rate = \frac{FP+FN}{TP+TN+FP+FN} = \frac{40}{200} = 20\%$

$Precision = \frac{TP}{TP+FP} = \frac{75}{90} = \frac{5}{6}$

$Recall = \frac{TP}{TP+FN} = \frac{75}{100} = 0.75$

$F1 = \frac{2PR}{P+R} = \frac{15}{19}$

Perf-6

答:

(1). 偏差-方差困境 (**biase-variance dilemma**) 指无法同时降低偏差和方差，其中一个的下降会导致另一个的上升，只能在两者之间取得均衡。

给定学习任务，假定我们能控制学习算法的训练程度，则在训练不足时，学习器的拟合能力不够强，训练数据的扰动不足以使学习器产生显著变化，此时**偏差主导了泛化错误率**；随着训练程度的加深，学习器的拟合能力逐渐加强，训练数据发生的扰动渐渐能被学习器学到，**方差逐渐主导了泛化错误率**；在训练程度充足后，学习器的拟合能力已经非常强，训练数据发生的轻微扰动都会导致学习器发生显著变化，若训练数据自身的、非全局的特性被学习器学习到了，则将发生过拟合。

简单来说，在模型中，若想降低偏差，便会增加模型的复杂度，防止欠拟合;但同时又不能让模型太复杂而导致方差增加，造成过拟合 。因此在模型的复杂度上，需要找到一个平衡点。

(2).

bias（偏倚） 度量了某种学习算法的平均估计结果所能逼近学习目标的程度。（独立于训练样本的误差，刻画了匹配的准确性和质量：一个高的偏差意味着一个坏的匹配） variance （方差）则度量了在面对同样规模的不同训练集时，学习算法的估计结果发生变动的程度。（相关于观测样本的误差，刻画了一个学习算法的精确性和特定性：一个高的方差意味着一个弱的匹配）

高偏差对应了欠拟合、高方差对应了过拟合。

Perf-7

答：为了debugging学习机，我们将训练数据更进一步分为训练集和验证集，训练集用于训练模型，验证集用于验证模型。根据模型我们可以算出训练误差，验证误差和测试误差。

对于没有足够先验经验的情况时，调整超参数，修改模型结构总是必要的，这是就是(1). 选择改变学习机的结构（调整超参数及增加或者删减特征），直到达到验证误差的最小值。

通过达到验证误差的最小值，我们即可获得具有最佳泛化能力的模型，如果此时模型的学习性能仍然不满足要求，我们就需要考虑是否为数据的噪声过大或者数据量不足，此时我们就需要选择(2). 选择增加、调整训练数据。