机器学习第一次作业

姓名：刘文晨

学号：202222280328

数据集包含1000个样本，其中500个正例、500个反例，将其划分为包含70%样本的训练集和30%样本的测试集用于留出法评估，试估算共有多少种划分方式?

留出法评估直接将数据集随机划分为两个互斥的集合。

按照题意，70%样本作为训练集，30%样本作为测试集。

于是，根据分层采样原则，我们从500个正例中拿出350个作为训练集，剩下150个作为测试集；对于500个反例，我们也这样划分。

划分方式的总数=$C_n^m \times C_n^m$，其中n=500，m=350，使用Matlab的nchoosek函数计算组合数，[nchoosek(500,350)]^2≈2.9859×10^{262}。

试述真正例率（TPR）、假正例率（FPR）与查准率（P）、查全率（R）之间的关系

首先我们要了解混淆矩阵，如下表所示：

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

然后得到真正例率、假正例率、查准率、查全率的概念和公式：

- **真正例率**：真实正例被预测为正例的比例
- **假正例率**：真实反例被预测为正例的比例
- **查准率**：又叫精准率，预测为正例的实例中真实正例的比例
- **查全率**：又叫召回率，真实正例被预测为正例的比例

指标	公式
真正例率	$TPR = \frac{TP}{TP+FN}$
假正例率	$FPR = \frac{FP}{FP+TN}$
查准率	$P = \frac{TP}{TP+FP}$
查全率	$R = \frac{TP}{TP+FN}$

显然，真正例率（TPR）和查全率（R）是相等的。而假正例率（FPR）和查准率（P）并没有直接的数值关系。

试述错误率与ROC曲线之间的关系

错误率：错分样本占全部样本的比例。

错误率的计算公式：$E = \frac{FN+FP}{TP+FN+FP+TN} = 1 - \frac{TP+TN}{TP+FN+FP+TN}$

ROC曲线以假正例率（FPR）为横轴，以真正例率（TPR）为纵轴，表示了模型在不同截断点取值下的泛化性能。

错误率是在阈值固定的情况下得出的，ROC曲线是在阈值随着样本预测值变化的情况下得出的。ROC曲线上的每一个点，都对应着一个错误率。

ROC中越接近(1,0)点的越完美，常常需要计算错误率实现查准率（P）和查全率（R）的折中，而P、R则反映了我们所侧重部分的错误率。