1. 表1.1中若只包含编号为1和4的两个样例，试给出相应的版本空间。

数据集：

	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	稍蜷	沉闷	否
	2	2	2	

假设空间

共 3×3×3+1 种假设。

与样例1匹配的假设有

1.	*	*	*	2.	*	蜷	*
3.	*	*	浊	4.	青	*	*
5.	*	蜷	浊	6.	青	蜷	*
7.	青	*	浊	8.	青	蜷	浊.

除去与样例2匹配的结果：1.

∴ 版本空间有7个，如上所示 2-8

2. 试述机器学习能在互联网搜索的哪些环节起什么作用。

图片搜索：通过标签检索图片，需先对图片进行分类标注。

推送服务：基于用户的点击，分析出用户可能需求，从而推送广告。

自然语言处理：音频输入，翻译功能。

3. 数据集1000个样本，正：反＝1:1，划分为70%样本的训练集和30%的测试集，试估算共有多少种划分方式。

对于测试集应有150个正例和150个反例。

∴ 共有 $(C_{500}^{150})^2$ 种划分。

4. 100个样本，正：反＝1:1，假定学习算法所产生的模型是将新样本预测为训练样本较多的类别，试给出用10折交叉验证法和留一法分别对错误率进行评估的结果。

10折交叉验证：因为每个子集都需保证数据一致性 ∴训练集中有正例＝反例＝45.

测试集中随机猜测，错误率为50%

留一法：训练集中正:反或反:正 ＝ 49:50

对每种情况，预测都错误，总错误率为100%

5. 试述真正例率(TPR)、假正例率(FPR)与查准率(P)、查全率(R)之间的联系。

	正	反
正	TP	FN
反	FP	TN

$P = \frac{TP}{TP+FP}$ 查准率 表示预测为正的样本有多少为真的正样本。

$R = \frac{TP}{TP+FN}$ 查全率，表示所有正例中，有多少正例被预测出来。

两者一般是制约关系，不能同时增加。

$TPR = \frac{TP}{TP+FN}$ 正样本预测数占总正例的比例

$FPR = \frac{FP}{FP+TN}$ 预测为正的负样本占总负例的比例

6. 线性判别分析在线性可分数据上能获得理想结果，设计一个改进方法，使其也能够很好用于非线性可分数据。

采用SVM，选择一个核函数，通过将数据映射到高维空间，最终在高维特征空间中构造出最优分离超平面，从而把平面上本身不好分的非线性数据分开。