1 简答部分

1. 应用贝叶斯决策需要满足的三个前提条件是什么?

答：若要使用贝叶斯判定准则来最小化决策风险 R，首先要获得后验概率 P(c|x)，基于贝叶斯定理：

$$P(c|x) = \frac{P(c) \cdot P(x|c)}{P(x)}$$

其中，P(c) 是类先验概率，P(x|c) 是样本 x 相对于类标记 c 的类条件概率(似然)。

2. 试简述对先验概率和后验概率的理解

答：先验概率表达了样本空间中各类样本的比例，根据大数定律，当训练集包含充足的独立同分布样本时，先验概率就可以通过各类样本出现的频率来进行估计。

后验概率是通过贝叶斯公式对先验概率进行修正，计算而得出的概率。

3. 试简述 Fisher 线性判别的基本思想

答：对于给定的训练样本，需要设法将这些样本投射到一条直线上，使得同一类样本在该直线上的投影点之间的距离尽可能的小，非同一类样本在该直线上的投影点之间的距离尽可能大；在对新样本进行分类时，将它同样投影到该直线上，根据其投影点的位置来判断它的类别。

4. 试简述何为 K-近邻法

答：K-近邻学习是一种常用的监督学习方法，其工作机制非常简单：给定测试样本，基于某种距离度量（比如欧几里得距离）找出训练集中与其最靠近的 K 个训练样本，然后基于这 K 个『邻居』的信息来进行预测。

5. 试简述对非线性支持向量机（SVM）的理解

答：对于线性支持向量机，选择一个合适的惩罚参数，并构造凸二次函数线性规划问题，求得原始问题的对偶问题的最优解 α^*，由此可以求出原始问题的最优解；在处理非线性行问题时，通过选择合适的核函数来将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。

6. 试简述何为度量学习

答：在机器学习中，对高维数据进行降维的主要目的是希望找到一个合适的低维空间,在此空间中学习比原始空间更好。事实上,每个空间对应了在样本属性上定义的一个距离度量。度量学习就是通过『学习』来找出一个合适的距离度量。

7. 简述何为半监督学习

答：半监督学习 (Semi-Supervised Learning, SSL) 是模式识别和机器学习领域研究的重点问题，是监督学习与无监督学习相结合的一种学习方法。它主要考虑如何利用少量的标注样本和大量的未标注样本进行训练和分类的问题。主要分为半监督分类，半监督回归，半监督聚类和半监督降维算法。

8. 试简述何为聚类

答：聚类试图将数据集中的样本划分为若干个通常是不相交的子集称为一个『簇』，通过这样的划分，每个簇可能对应于一些潜在的概念（类别），并且这些概念对于聚类算法而言事先是未知的，聚类过程仅能自动地形成簇结构，簇所对应的概念语义需要使用者来把握和定义。

9. 简述对稀疏表达的理解

答：假设一个样本数据 D，D 对应的矩阵中存在很多零元素，并且它们不是以整行整列的形式出现的，这样的稀疏表达形式对学习任务会有不少好处（例如，SVM 在文本上有很好的性能）；若给定数据集 D 是稠密的，即普通非稀疏数据, 我们可以通过『字典学习』(『稀疏编码』) 来将样本转化为合适的稀疏表示。

10. 简述对流型学习的理解

答：流型学习是一类借鉴了拓扑流形概念的降维方法，它可以容易的在局部建立降维映射关系，然后再设法将局部关系拓广到全局；流型学习也通常被用于可视化，因为当维数被降至二维或三维时，能进行可视化。等度量映射和局部线性嵌入是两种著名的流型学习方法。

11. 简述对同分布问题的理解

答：通常假设样本空间中的样本全部都服从一个未知的分布 D，我们获得的所有样本都是独立同分布的从这个分布上采样获得的，即『独立同分布』，一般而言，训练样本越多，我们得到的对于 D 的信息越多。

12. 试简述对模型泛化能力的理解

答：机器学习的目标是使学得的模型能很好地应用于『新样本』，而不是仅仅在训练样本上工作得很好，学得模型适用于新样本的能力称为『模型泛化能力』。具有抢饭化能力的模型能很好地适用于整个样本空间。

13. 为什么机器学习此时才热起来?

答：两个基本原因：数据大了，计算能力强了。

数据小，样本少，容易产生过拟合问题，面对复杂模型，计算能力不够，则无法求解。

2 计算问题

设在某个局部地区细胞识别中正常 w1 和异常 w2 两类的先验概率分别为：

正常状态：$\mathrm{P(w1)} = 0.9$

异常状态：$\mathrm{P(w2)} = 0.1$

现有一待识别的细胞，其观察值为 x，从类条件概率密度分布曲线上查得：

$\mathrm{P(x|w1)} = 0.2, \mathrm{P(x|w2)} = 0.4$

试使用贝叶斯决策对该细胞 x 进行分类

答：根据贝叶斯公式：

$$P\left(\omega_1|x\right) = \frac{P\left(x|\omega_1\right) \cdot P\left(\omega_1\right)}{P\left(x|\omega_1\right) \cdot P\left(\omega_1\right) + P\left(x|\omega_2\right) \cdot P\left(\omega_2\right)} \approx 0.818$$

又有，

$$P(\omega_2|x) = 1 - P(\omega_1|x) = 0.182$$

因为，

$$P(w_1|x) > P(w_2|x)$$

所以，x 归类于正常状态。

3 论述部分

如果让您设计与实现一个模式识别系统，用于实现齐鲁软件学院男、女士的分类（二分类问题），您将如何考虑？其中有哪些需要注意的问题？请就您的理解，尽可能全面，深入地描述，以此展示您对《模式识别技术》这门课程的理解。如果您觉得必要，必要之处也可以画图辅助表达。

答：假设我们所拥有的训练样本有如下属性：

身高	是否喜欢网购	出行次数（月）	生活费（月）	性别
155	是	12	2500	女
159	是	11	2200	女
182	否	8	1800	男
.....	.....	.....	.....	.....

我们可以采用 ID3 决策树算法来用于对学生性别的分类。

在建立决策树的过程中，首先需要对属性进行划分，为了选择出最优划分属性，我们需要计算出用每个属性对样本集进行划分所获得的信息增益，选择信息增益大的属性划分，我们可以得到一棵决策树。

可能存在的问题：

过拟合：我们可以通过剪枝来解决过拟合的问题，使得决策树不会出现分支过多的问题。连续值处理：对上述例子中的身高属性，即为连续值的属性，因为连续属性的可取值的数目不是无限的，所以不能根据属性值来划分，因此要计算，找出划分点。缺失值处理：如果某些样本的某些属性缺失，我们也不能浪费这些样本，C4.5 算法提供了解决方案。