一、 简答题

1.简述数据预处理方法。

数据预处理的方法：

（1）数据清洗 —— 去噪声和无关数据

（2）数据集成 —— 将多个数据源中的数据结合起来存放在一个一致的数据存储中

（3）数据变换 —— 把原始数据转换成为适合数据挖掘的形式

（4）数据规约 —— 主要方法包括：数据立方体聚集，维度归约，数据压缩，数值归约，离散化和概念分层等。

（5）图说事实

2.简述分类不平衡问题，以及一般的处理方法和评价指标。

简述：当样本在不同的类别上的数目分布（显著）不平衡时候，则称为类别不平衡（分类）问题，很多分类器在类别不平衡时难以取得满意的效果，常把小类样本分错。引起类别不平衡的原因往往由问题的本身或者采样引起的。

处理的方法：重采样方法，通过在数据上的重新采样，获得类别平衡的数据，在通过已有的分类算法训练。阈值调整，修改或重新设计训练算法，直接将 AUC 作为算法的评价指标。

3.简述过拟合和欠拟合问题，以及控制拟合程度的方法。

通俗一点地来说过拟合就是模型把数据学习的太彻底，以至于把噪声数据的特征也学习到了，这样就会导致在后期测试的时候不能够很好地识别数据，即不能正确的分类，模型泛化能力太差。欠拟合就是模型没有很好地捕捉到数据特征，不能够很好地拟合数据。常常采用交叉验证的方法来控制拟合程度。

4.简述基于内容的推荐系统和基于协同过滤方法的推荐系统的优缺点。

协同过滤方法只考虑了用户评分数据， 忽略了项目和用户本身的诸多特征， 如电影的导演、演员和发布时间等， 用户的地理位置、性别、年龄等. 如何充分、合理的利用这些特征， 获得更好的推荐效果， 是基于内容推荐策略所要解决的主要问题.

基于内容的推荐系统：根据历史信息(如评价、分享、收藏过的文档)构造用户偏好文档， 计算推荐项目与用户偏好文档的相似度， 将最相似的项目推荐给用户. 例如, 在电影推荐中, 基于内容的系统首先分析用户已经看过的打分比较高的电影的共性(演员、导演、风格等), 再推荐与这些用户感兴趣的电影内容相似度高的其他电影。

二、感知机

1.对于训练数据集，其中正例点是 x1=(4,3)T，x2=(3,4)T，负例点 x3=(1,1)T，用感知机学习算法的原始形式求感知机模型 f(x)=w 点乘 x+b。这里 w=(w(1),w(2))T，x=(x(1),x(2))T。

2.对上述实例，采用感知机学习的对偶形式求解感知机模型。

参考统计学习例题

五．朴素贝叶斯

（参考统计学习例题 4.1）

六、考点是 KNN 分类问题，用欧拉公式看不清~

七、聚类

1.说明 k-means 算法对初始点敏感的缺点

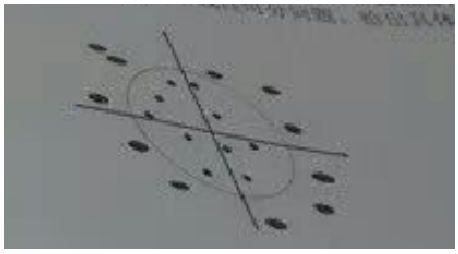
2.设计简单的方法克服 k-means 对初始点选取敏感的缺点

缺点：

代次数,n 是所有样本的个数.然而,传统的 k-means 算法对于初始聚类中心的选择是随机的,或者是由用户主观指定的.k-means 算法对初始聚类中心的依赖性很强,初始聚类中心选取的不同往往导致聚类结果相当不稳定,同时也会使迭代的次数大幅度增加;

克服：采用模糊 kmeans 算法是 kmeans 聚类模糊形式。与 kmeans 算法排他性聚类不同，模糊 kmeans 尝试从数据集中生成有重叠的簇。在研究领域，这也叫做模糊 c-means 算法，可以把模糊 kmeans 看作 kmeans 算法的扩展。kmeans 致力于寻找硬簇（一个数据集点只属于某一个簇）。在一个软聚类算法中，任何点都属于不止一个簇，而且该点到这些簇之间都有一定大小的吸引度。这种吸引度与该点到这个簇中心距离成比例。模糊 kmeans 有一个参数 m ，叫做模糊因子。与 kmeans 不同的是，模糊因子引入不是把向量分配到最近的中心，而是计算每个点到每个簇的关联度。

八、如下二维图所示的是一个线性不可分的分类问题。不同大小的点代表不同的数据。请问通过何种方式，可以将其转换成一个线性可分问题。给出具体的方案。（12 分）



将线性不可分的分类问题转换成线性可分的问题，可以通过将二维平面的数据映射到高维空间中去。在线性不可分的情况下，支持向量机首先在低维空间中完成计算，然后通过核函数将输入空间映射到高维特征空间，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。如图所示，一堆数据在二维空间无法划分，从而映射到三维空间里划分：

集成学习

1.假设硬币正面朝上的概率为 p，反面朝上的概率为 1-p。令 H(n)代表抛 n 次硬币所得正面朝上的次数，则最多 k 次正面朝上的概率为

取$p-\delta=\frac{1}{2}$，则$\delta=p-\frac{1}{2}=\frac{1}{2}-\varepsilon$

$$P(H(n)\leqslant\frac{n}{2})=\sum_{i=0}^{\frac{n}{2}}\binom{n}{i}p^i(1-p)^{n-i}\leqslant e^{-2(\frac{1}{2}-\varepsilon)^2n}=e^{-\frac{1}{2}(1-2\varepsilon)^2n}$$

2.对于 0/1 损失函数来说，指数损失函数并非仅有的一致替代函数。考虑式(8.5)，试证明:任意随机函数

总损失$L=l(-H(x)f(x))P(f(x)|x)=l(-H(x))P(f(x)=1|x)+l(H(x))P(f(x)=0|x)$
$H(x)\in-1,1$，要使L最小，当$P(f(x)=1|x)>P(f(x)=0|x)$时，会希望$l(-H(x))<l(H(x))$,由于l是递减的，得$H(x)>-H(x)$，的$H(x)=1$。同理当$P(f(x)=1|x)<P(f(x)=0|x)$时,$H(x)=-1$。

$l(-H(x)f(x))$是对$H(X)$的单调递减函数，那么可以认为$l(-H(x)f(x))$是对$(-H(X))$的单调递增函数
此时$H(x)=argmax_{y\in0,1}P(f(x)=y|x)$,即达到了贝叶斯最优错误率，说明$l$是0/1损失函数的一致替代函数。

4.GradientBoosting 是一种常用的 Boosting 算法，是分析其与 AdaBoost 的异同。
答：GradientBoosting 与 AdaBoost 相同的地方在于要生成多个分类器以及每个分类器都有一个权值，最后将所有分类器加权累加起来
不同在于：
AdaBoost 通过每个分类器的分类结果改变每个样本的权值用于新的分类器和生成权值，但不改变每个样本不会改变。
GradientBoosting 将每个分类器对样本的预测值与真实值的差值传入下一个分类器来生成新的分类器和权值(这个差值就是下降方向)，而每个样本的权值不变。

6.试述为什么 Bagging 难以提升朴素贝叶斯分类器的性能。
答：Bagging 主要是降低分类器的方差，而朴素贝叶斯分类器没有方差可以减小。对全训练样本生成的朴素贝叶斯分类器是最优的分类器，不能用随机抽样来提高泛化性能。

7.试述随即森林为什么比决策树 Bagging 集成的训练速度快。
答：随机森林不仅会随机样本，还会在所有样本属性中随机几种出来计算。这样每次生成分

类器时都是对部分属性计算最优，速度会比 Bagging 计算全属性要快。

8.MultiBoosting 算法与 Iterative Bagging 的优缺点。
答：MultiBoosting 由于集合了 Bagging，Wagging，AdaBoost，可以有效的降低误差和方差，特别是误差。但是训练成本和预测成本都会显著增加。
Iterative Bagging 相比 Bagging 会降低误差，但是方差上升。由于 Bagging 本身就是一种降低方差的算法，所以 Iterative Bagging 相当于 Bagging 与单分类器的折中。

10.试设计一种能提升 k 近邻分类器性能的集成学习算法。
答：可以使用 Bagging 来提升 k 近邻分类器的性能，每次随机抽样出一个子样本，并训练一个 k 近邻分类器，对测试样本进行分类。最终取最多的一种分类。

绪论

3.若数据包含噪声，则假设空间中可能不存在与所有训练样本都一致的假设。在此情形下，试设计一种归纳偏好用于假设选择
通常认为两个数据的属性越相近，则更倾向于将他们分为同一类。若相同属性出现了两种不同的分类，则认为它属于与他最临近几个数据的属性。也可以考虑同时去掉所有具有相同属性而不同分类的数据，留下的数据就是没误差的数据，但是可能会丢失部分信息。

5.试述机器学习在互联网搜索的哪些环节起什么作用
1.最常见的，消息推送，比如某东经常说某些商品我可能会感兴趣，然而并没有。
2.网站相关度排行，通过点击量，网页内容进行综合分析。
3.图片搜索，现在大部分还是通过标签来搜索，不过基于像素的搜索也总会有的吧。

支持向量机

5.试述高斯核 SVM 与 RBF 神经网络的联系
RBF 网络的径向基函数与 SVM 都可以采用高斯核，也就分别得到了高斯核 RBF 网络与高斯核 SVM。
神经网络是最小化累计误差，将参数作为惩罚项，而 SVM 相反，主要是最小化参数，将误差作为惩罚项。
在二分类问题中，如果将 RBF 中隐层数为样本个数，且每个样本中心就是样本参数，得出的 RBF 网络与核 SVM 基本等价，非支持向量将得到很小的 w
.使用 LIBSVM 对异或问题训练一个高斯核 SVM 得到 α 修改第 5 章 RBF 网络的代码，固定 β 参数为高斯核 SVM 的参数，修改每个隐层神经元的中心为各个输入参数，得到结果 w,w 与 α 各项成正比例。

6.试析 SVM 对噪声敏感的原因。
SVM 的目的是求出与支持向量有最大化距离的直线，以每个样本为圆心，该距离为半径做圆，可以近似认为圆内的点与该样本属于相同分类。如果出现了噪声，那么这个噪声所带来的错误分类也将最大化，所以 SVM 对噪声是很敏感的。