简答题

1. 监督学习和无监督学习的区别，举例；分类和回归的区别?

(1) 监督学习：初始训练样本有标记信息。如分类和回归，如支持向量机、决策树等；

(2) 无监督学习：初始训练样本无标记信息。如聚类，密度估计，异常检测等。

分类：属性值是离散值，预测值是有限标记，通常用准确率来评估

回归：属性值是连续值，预测值是一系列连续的值，通常用均方误差来评估

2. 简述支持向量机的原理和为什么核函数有用?

(1) 支持向量机的原理：寻找一个超平面，使正负样例划分开，同时使得正负样例到超平面的间隔最大，数学表达式略。

(2) 为什么核函数有用：核函数能够将初始样本空间转化为一个更高维的特征空间，即“再生核希尔伯特空间”，使样本线性可分。如果样本空间是有限维，即属性值是有限的，则必然存在一个更高维的特征空间，在这里，样本是线性可分的。

3. 决策树连续值如何处理?

将连续属性值升序排序，取相邻值的平均值，高于作为正例，低的作为负例，然后再用那个信息增益公式寻找最大信息增益

4. 随机森林的生成过程，随机性体现在哪里?

①随机森林是在 bagging 的基础上，以决策树作为基学习器进行生成，在决策树结点寻找划分属性时，是从该结点的所有属性中选取若干属性进行划分，而不是全部。

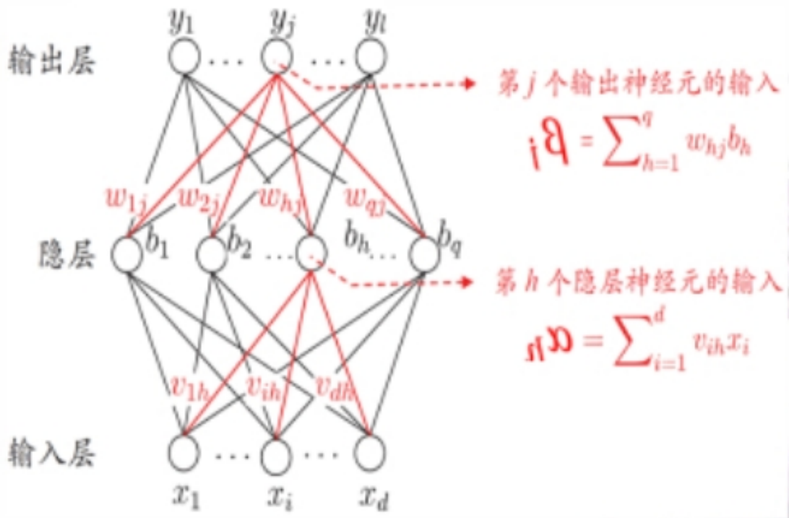
②bagging 的生成过程：基于自助采样采集出 T 个含 m 个样本的训练数据集，然后用这 T 个数据集训练出 T 个基学习器，利用结合策略(如分类用简单投票法，回归用简单平均法)在集成出最终的学习器。

③随机性体现在随机选取其中一部分属性进行划分，而不是该结点的全部属性。

计算题

1. 神经网络

权重：v_{ih}、w_{hj}　阈值：θ_j、γ_h　$(i=1,\cdots,d, h=1,,\cdots,q, j=1,\cdots,l)$



step1:　　α_h　　α_h

step2:　$\hat{y}_j^k = f(\beta_j - \theta_j)$, $\beta_j = \sum_{h=1}^{q} w_{hj} b_h$

step3:　$E_k = \frac{1}{2}\sum_{j=1}^{l}(\hat{y}_j^k - y_j^k)^2$

f(x)= $\text{sigmoid}(x) = \frac{1}{1+e^{-x}}$

2. 计算概率(贝叶斯)

例子：假定某种疾病很稀少，每100万人只有一人患病。还假定有一种化验很有效，如果一个人患此病，则化验结果为阳性的可能性为99%。然而，这种化验是不完美的，在健康人身上化验结果为阳性的可能性是1/1000。假定来了一位新患者，其化验结果为阳性。利用贝叶斯法则计算该患者患此疾病的概率有多大?

解：设该疾病用 d 表示，化验结果用 t 表示。我们有：$P(d=1)=10^{-6}$，$P(t=1|d=1)=0.99$，$P(t=1|d=0)=10^{-3}$。我们要求出 $P(d=1|t=1)$。

使用贝叶斯规则：

$$P(d=1|t=1) = \frac{P(t=1|d=1)P(d=1)}{P(t=1)}$$
$$= \frac{P(t=1|d=1)P(d=1)}{P(t=1|d=1)P(d=1)+P(t=1|d=0)P(d=0)}$$
$$= \frac{0.99\times10^{-6}}{0.99\times10^{-6}+10^{-3}\times(1-10^{-6})}$$
$$= 0.00098902$$

也就是说，知道化验结果为正把患病概率从 1/1000000 提高到 1/1000。

3. K 均值计算：略

思考题

1. 过拟合欠拟合概念；结合偏差和方差，解释造成过拟合欠拟合的原因；神经网络中怎么解决过拟合，给出至少 2 种方法

(1)过拟合：学习器的学习能力太强，将训练样本自身的一些特性而不是所有潜在样本都具有的性质到了，使泛化能力降低的情况。

(2)欠拟合：学习器的学习能力不足，样本的一般的性质尚未学好，泛化能力不强。

在偏差、方差、泛化误差曲线中，当模型的学习程度较低时，数据样本的扰动对模型的影响较小，此时偏差占主导地位，偏差较小，发生欠拟合现象。

随着模型的学习程度增强时，数据样本的扰动渐渐被模型学习到。当模型学习程度过强，以及样本自身的特性被学习到时，偏差降低，方差占主导地位，发生过拟合现象。

(1) 神经网络缓解过拟合：

①早停：数据集划分为训练集和验证集，训练集用于计算梯度，更新连接权和阈值，验证集用来计算泛化误差。当训练集误差降低，验证集误差升高时，停止训练。

②加正则化项：在目标误差中加入描述神经网络复杂度的部分，例如阈值或连接权的平方和

2. 一些数据有标记，一些数据没有标记；设计一种方法，能够尽可能充分利用这些数据。

略。