第二次课后作业

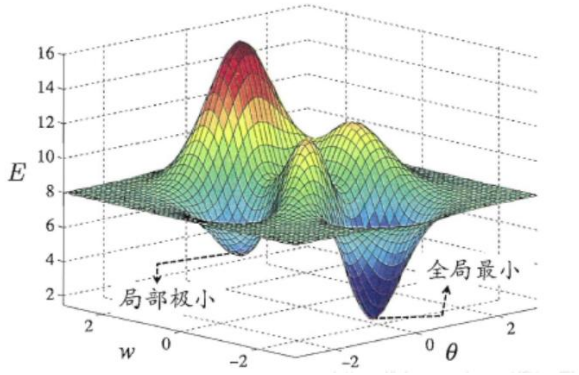
1、朴素贝叶斯分类器解决了什么障碍？ 它的关键假设是什么？

贝叶斯分类器可以根据给出的待分类项求解出各类别概率，选取概率最大的作为当前分类。而朴素贝叶斯分类器基于条件独立性假设，假设导致事情发生的各属性之间相互独立，通过独立事件乘法展开，将复杂的多属性的联合概率转化为多个简单的单一属性概率的乘积。朴素贝叶斯分类器解决了多属性联合概率难以计算的障碍。

它的关键假设就是条件独立性假设，即属性之间相互独立。

2、请简述局部极小与全局极小。

局部极小指的是空间中某一点比与其相邻的所有点的取值都要小，而全局极小指的是在整个空间中某点比所有点的取值都小（或相等）。因此一个函数空间中可以有多个局部极小值，但只能有唯一的全局极小（允许多个全局极小点，但取值必须相同且为最小）。全局极小和局部极小如图所示：



3、什么是监督学习和非监督学习，请说明它们的区别并各举一个例子；请说明分类和回归问题的区别。

监督学习是指利用已知分类值/回归值的样本数据进行训练，得到最优模型。通过性能测试后即可对未知类别/预测值的数据进行分类或回归。例如决策树、逻辑对率回归模型、神经网络模型等，都是在样本数据有明确标签或回归值的情况下进行训练，然后对测试样本进行预测分类值或回归值。

非监督学习则是利用没有标签的样本，对数据集进行建模，通过属性相似性或关联性进行分类或者预测。例如聚类算法，将相似度高的样本分为同一类。

分类问题指的是根据样本属性对样本类别进行划分，有明确的的目标类别，如西瓜数据集利用条纹、响度等属性来判别西瓜是“好瓜”还是“坏瓜”。这些类别是离散且固定的。

回归问题是根据样本属性预测其他属性，没有明确指定的值可供选择或参考。如房价数据集根据经纬度、楼层等属性预测价格。其预测出来的价格是连续的，

没有事先给定的集合元素可供选择。

4.请简述随机森林的生成方法以及其随机性体现在哪里？

假设样本容量为 N，随机森林的生成方法如下：

（1）在容量为 N 的样本集中**有放回**的抽取 N 次，每次只取 1 个样本，形成了一个包含 N 个样本的数据集；

（2）假设样本特征数为 a，在其中**随机选择 k（k<a）个特征**，用这 N 个样本的 k 个特征建立决策树；

（3）重复上述操作 m 次，得到 m 棵决策树；

（4）采用如投片机制等集成方法进行集成，最终得到一个总的生成模型。

随机森林的随机性主要体现在两个方面，即：

• Bootstrap 取样，也就是每次建立决策树前在样本中有放回随机抽取样本，这里面有可能有的样本重复出现，也有样本可能不出现；

• 在样本的属性中随机抽取 k 个建立决策树，这会使得决策树考虑的属性具有随机性。

5、请为以下决策树算法的步骤 3，6，8，12 填写为代码

输入： 训练集 $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}$;
属性集 $A = \{a_1, a_2, \ldots, a_d\}$.

过程： 函数 TreeGenerate(D, A)

1: 生成结点 node;
2: **if** D 中样本全属于同一类别 C **then**
3: 该node为叶结点，类别为C；return
4: **end if**
5: **if** $A = \varnothing$ **OR** D 中样本在 A 上取值相同 **then**
6: 该node为叶结点，类别为D中频数最大的类；return
7: **end if**
8: 选取最优划分属性a_*
9: **for** a_* 的每一个值 a_*^v **do**
10: 为 node 生成一个分支；令 D_v 表示 D 中在 a_* 上取值为 a_*^v 的样本子集；
11: **if** D_v 为空 **then**
12: 该node为叶结点，类别为D中频数最大的类；return
13: **else**
14: 以 TreeGenerate(D_v, $A \setminus \{a_*\}$)为分支结点
15: **end if**
16: **end for**

输出： 以 node 为根结点的一棵决策树

决策树学习基本算法

6、请阐述机器学习中欠拟合和过拟合现象，并结合偏差(bias)和方差(variance)解释其出现的原因。以人工神经网路学习为例，请给出至少两种解决其过拟合

的方法。

欠拟合指的是对样本的普遍性特征没有学习到位，导致模型性能欠佳。产生欠拟合的原因可能是训练次数过少、样本数较少等。欠拟合表现的数学特征就是偏差过大，因为模型学习不到位，模型欠拟合，所以对测试样本的预测值离真实值较远，即偏差过大。

过拟合指的是训练过度，导致模型将一些样本的个性化特征训练成了所有样本的特征。产生的原因可能是训练次数过多。过拟合表现的数学特征是方差过大，因为模型训练过度，误将训练样本的一些个性化特征学习为群体特征，这导致了大部分预测值都会有一定的偏离，哪怕预测值总体与真值相近，但大量的偏移导致了方差过大。

在 ANN 中，防止欠拟合可以采用增加迭代次数、增加样本数等方法，防止过拟合可以采用添加正则项、Dropout、提前停止训练等方法。