1. 数值属性信息增益的计算公式如下：

其中是属性取值的区间中位点的候选集。如图数据集仅含属性“含糖率”，请计算这个属性的信息增益和划分值。请写明计算步骤。

答案：

(0.3437756635351099, 0.126)

具体步骤参见P84

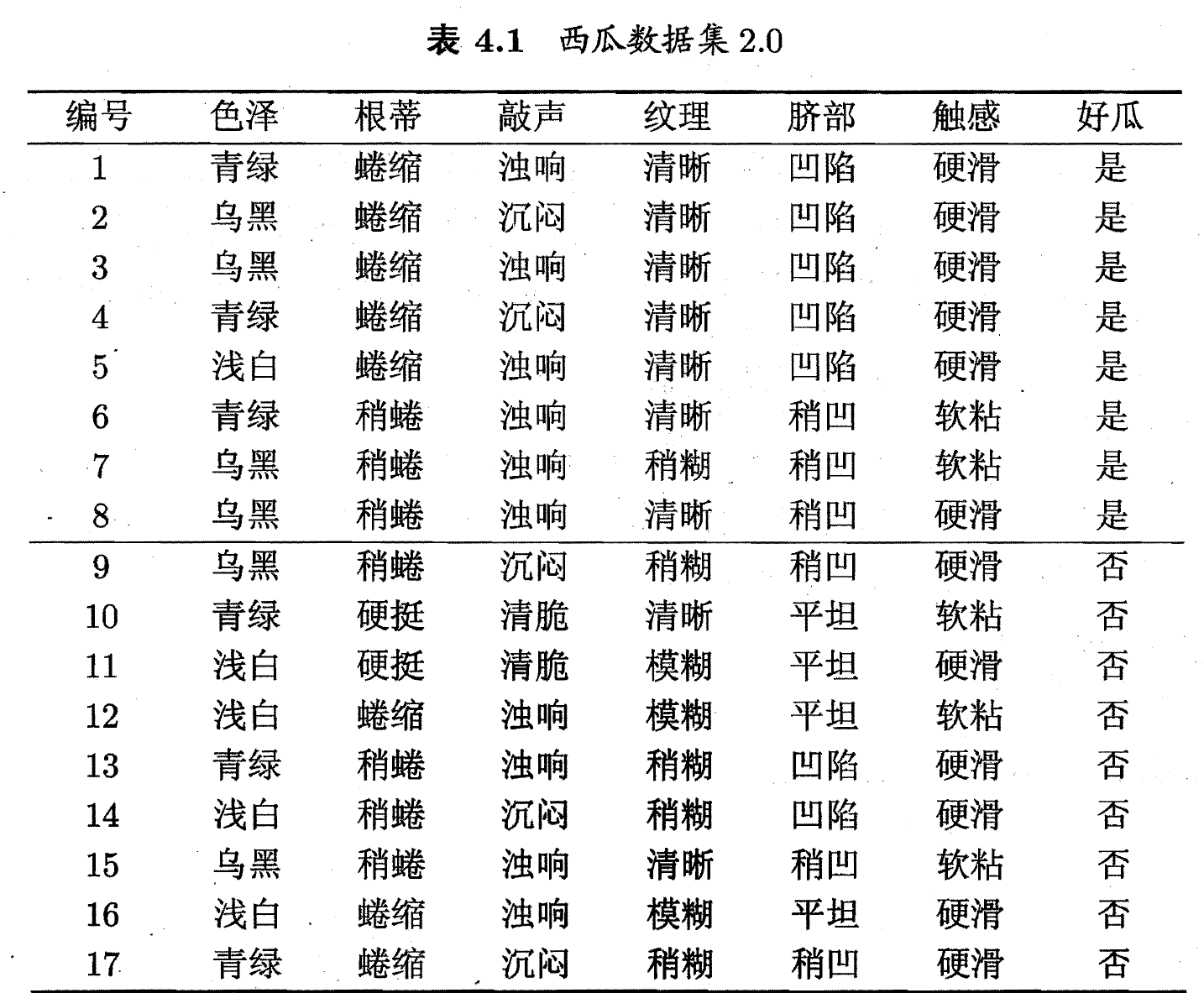
2. 如图所示是一个缺失值数据集，我们用缺失值处理方式构建决策树。已知属性“纹理”的信息增益最大，请用“纹理”属性作为树节点进行分支，写出各子树的样本集合和每个样本的权值。



答案：

"纹理"在所有属性中取得了最大的倍息增益，被用于对根结点进行划分.划分结果是使编号为{1,2,3,4,5,6,15} 的样本进入"纹理=清晰"分支，编号为{7,9, 13, 14, 17} 的样本边入"统理=稍糊"分支，而编号为{11,12,16} 的样本进入"纹理二模糊"分支，且样本在各子结点中的权重保持为1。编号为{8}和{10}的样本在属性"纹理"上出现了缺失值，因此它将同时进入三个分支中，由于原来的权重为1，权重在三个子结点中分别调整 {,,}.

3. 如图数据集，计算色泽属性的信息增益、增益率和基尼值。



答案：参见教材P77-79

4. 设有逻辑回归分类学习器对8个样本的测试，测试值由小到大排序结果是：

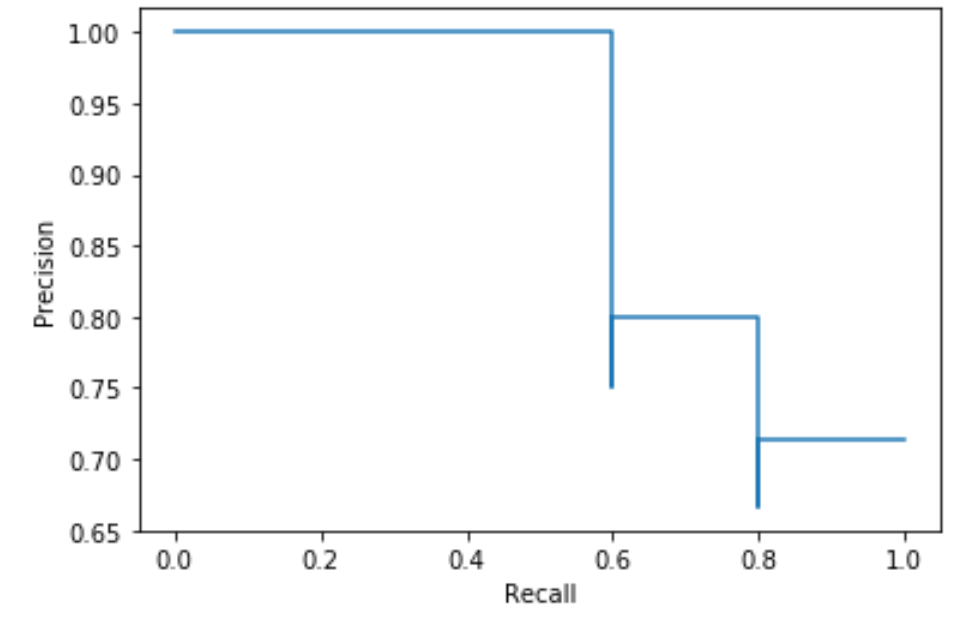
[0.1, 0.4, 0.45, 0.5,0.55,0.7,0.8,0.9]

这8个样本的标记是：

[FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRUE]

请绘制P-R曲线

答案：



5. 设有逻辑回归分类学习器对8个样本的测试，测试值由小到大排序结果是：

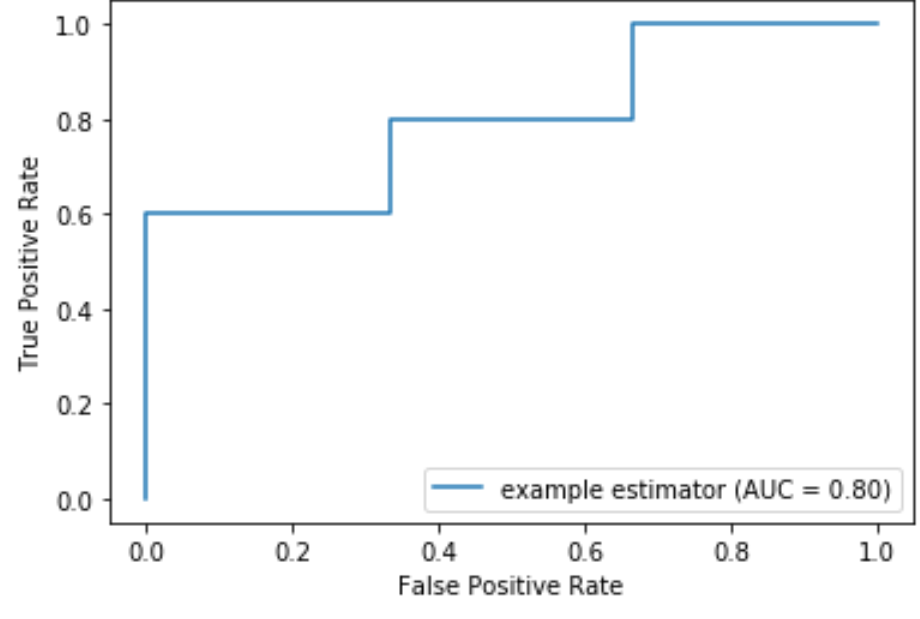
[0.1, 0.4, 0.45, 0.5,0.55,0.7,0.8,0.9]

这8个样本的标记是：

[FALSE, TRUE, FALSE, TRUE, FALSE, TRUE, TRUE, TRUE]

请绘制ROC曲线

答案：



6. 用朴素贝叶斯分类器，测试以下样本类别，要求写明计算步骤：

（1）['乌黑', '硬挺', '浊响']

（2）['青绿', '蜷缩', '沉闷']



答案：（1）否，（2）是

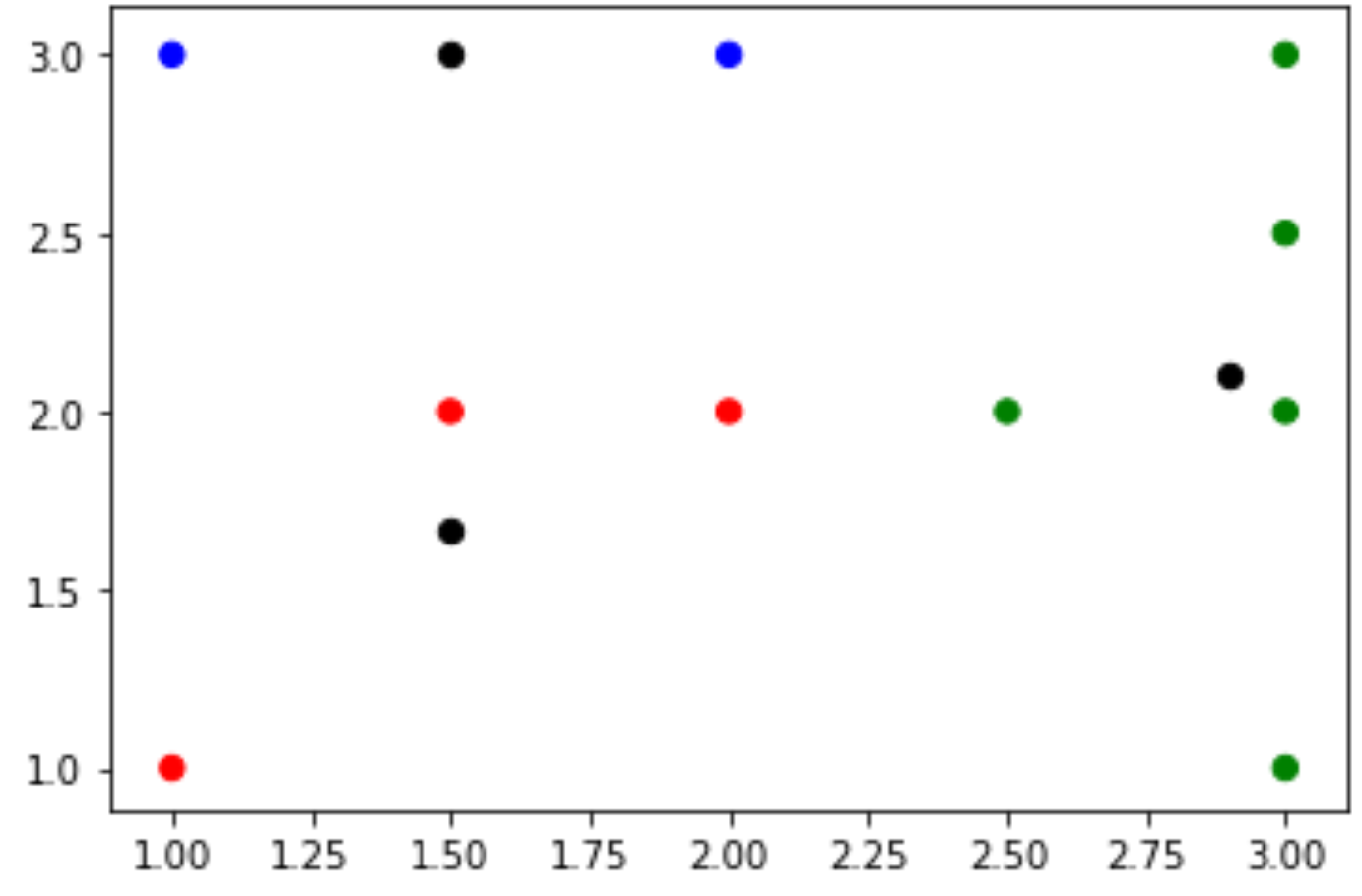
具体步骤参考教材P150

7. 有数据集如下：

X=np.array([[1,1],[1.5,2.0],[2,2],[3,3],[1,3],[2,3],[3,1],[2.5,2],[3.0,2.0],[3,2.5]])

请用k均值聚类算法（聚类簇数k=3）进行聚类，计算3个核心点的坐标，要求写明计算步骤。

答案：参见P203算法



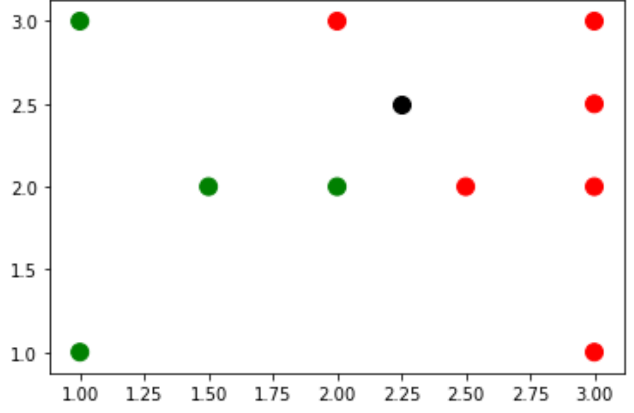
8. 有数据集如下：

X=np.array([[1,1],[1.5,2.0],[2,2],[3,3],[1,3],[2,3],[3,1],[2.5,2],[3.0,2.0],[3,2.5]])

y=np.array([1,1, 1, 0, 1, 0, 0, 0, 0,0])

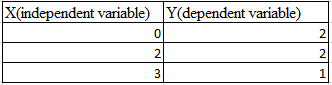
请用K近邻分类器（k=3）分类算法预测样本[2.25,2.5]的分类，要求写明算法步骤。

答案：



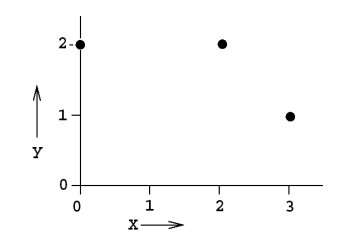
1. 计算[2.25,2.5]到各点的距离。
2. 对各距离排序
3. 选择前3个点构成子集
4. 子集中类别数最多的即为[2.25,2.5]的类别

9. 假设你有以下数据：输入和输出都只有一个变量。使用线性回归模型（y=wx+b）来拟合数据。那么使用留一法（Leave-One Out）交叉验证得到的均方误差是多少？



答案：留一法，简单来说就是假设有 N 个样本，将每一个样本作为测试样本，其它 N-1 个样本作为训练样本。这样得到 N 个分类器，N 个测试结果。用这 N个结果的平均值来衡量模型的性能。

对于该题，我们先画出 3 个样本点的坐标：



使用两个点进行线性拟合，分成三种情况:

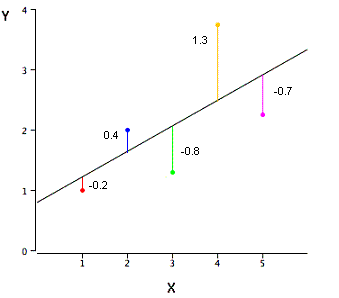
第一种情况下，回归模型是 y = 2，误差 E1 = 1。

第二种情况下，回归模型是 y = -x + 4，误差 E2 = 2。

第三种情况下，回归模型是 y = -1/3x + 2,误差 E3 = 2/3。

则总的均方误差为：(2^2 +(2/3)^2 +1^2) /3 = 49/27

10. 下面这张图是一个简单的线性回归模型,图中标注了每个样本点预测值与真实值的残差。计算 误差平方和 （SSE）为多少？



 答案：SSE 是平方误差之和（Sum of Squared Error），SSE = (-0.2)^2 + (0.4)^2 + (-0.8)^2 + (1.3)^2 + (-0.7)^2 = 3.02

11. 根据下表完成一下工作： （16分）

（1）简述决策树学习的目标，并列举两种度量样本集合纯度最常用的指标。

（2）“信息熵”是度量样本集合纯度最常用的一种指标，假定当前样本集合D中第k类样本所占的比例为pk，k=1，2，…∣Y∣，写出D的信息熵Ent(D)的公式表达。

（3）以属性“根蒂”为例，其对应的3个数据子集分别为 D1(根蒂=蜷缩)，D2(根蒂=稍蜷)，D3(根蒂=硬挺)，分别写出计算，Ent(D1)、Ent(D2)和Ent(D3)的过程。



答案：

（1）决策树学习的目标：提升结点纯度，即随着划分过程不断进行，我们希望决策树的分支结点所包含的样本，尽可能属于同一类别，即结点的“纯度”越来越高

度量样本集合纯度最常用的指标如信息熵、基尼值。

（2）

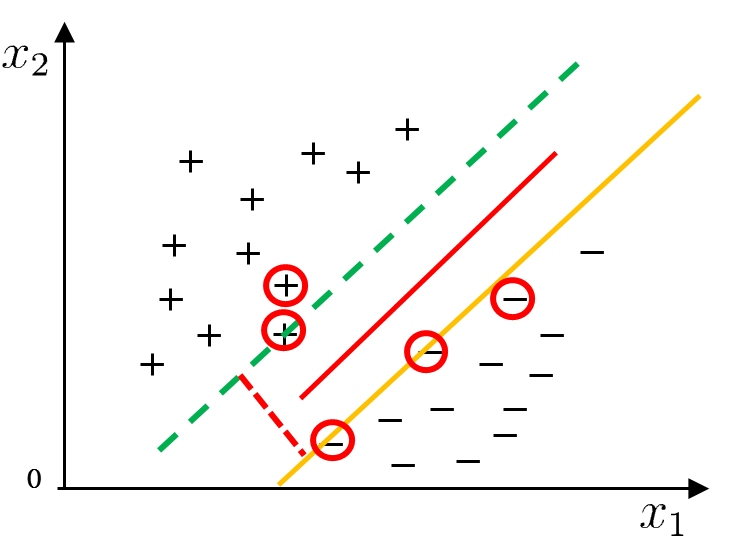
（3）D1(根蒂=蜷缩)有8个，其中正例5个，反例3个，D2(根蒂=稍蜷)有7个，其中正例3个，反例4个，D3(根蒂=硬挺)有2个，正例0个，反例2个，由此得出：

Ent(D1)=-(（5/8）\*log2(5/8)+ （3/8）\*log2(3/8)）

Ent(D2)= -(（3/7）\*log2(3/7)+ （4/7）\*log2(4/7)）

Ent(D3)= -(（0/2）\*log2(0/2)+ （2/2）\*log2(2/2)）

12. 结合下图，请回答问题：



（1）给定包含m个样本的数据集，

，写出支持向量机中划分超平面的方程，并解释其参数的含义。

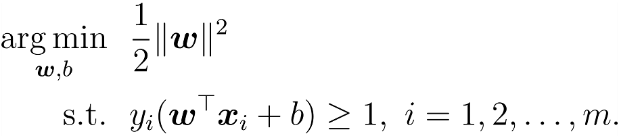
（2）写出基于该数据集的SVM的基本型（原问题）及其拉格朗日函数。

（3）解释SVM其解的稀疏性。

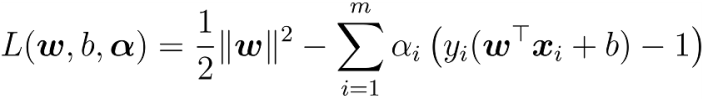
答案：

（1）支持向量机划分超平面，即 f(x) = ωTx+b 。ω为法向量，决定了超平面的方向；b为位移项，决定了超平面与原点之间的距离。

（2）SVM的基本型（原问题）为：



引入拉格朗日乘子 αi ≥ 0 得到拉格朗日函数：



（3）支持向量机解的稀疏性： 训练完成后，大部分的训练样本都不需保留， 最终模型仅与支持向量有关。

13. 请回答以下问题： （18分）

（1）简述k均值聚类算法的主要思想。

（2）简述k均值聚类算法的目标函数及其含义。

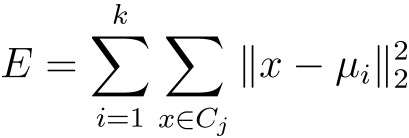
（3）简述k均值聚类算法的算法流程。

答案：（1）k均值算法算法思想：给定样本集 D ={x1， x2，…， xm}，将样本集划分为 k 个类。

以空间中 k 个点为中心进行聚类，对最靠近他们的对象归类。

通过迭代的方法，逐次更新各聚类中心的值，直至得到最好的聚类结果(簇划分) C ={C1， C2，…， Ck}

（2）k均值算法目标函数



其中，μi 是簇 Ci 的均值向量。

E值在一定程度上刻画了簇内样本围绕簇均值向量的紧密程度，E值越小，则簇内样本相似度越高。

（3）算法流程（迭代优化）：

随机选择k个点作为初始的聚类中心

repeat

对于所有的 k 个聚类中心

1）固定簇中心{μ1， μ2，…， μi}， 更新簇划分{C1， C2，…， Ck}

对任意一个样本，求其到k个中心的距离，将该样本归到距离最短的中心所在的类

2）固定簇划分{C1， C2，…， Ck}， 求解每个簇的均值向量{μ1， μ2，…， μi}

对每个簇，计算所有点的均值作为新的聚类中心

until 聚类中心不再发生改变

14. 预测传送带送过来的鱼是鲈鱼还是鲑鱼，其中鲈鱼c1和鲑鱼c2两类的先验概率分别为：0.8和0.2。现有一待识别的鱼，其鱼的光泽度指标x，从类条件概率密度分布曲线上查得鲈鱼：P(x|c1) = 0.15；鲑鱼：P(x|c2) = 0.5。试回答以下问题：

（1）简述朴素贝叶斯分类器原理。

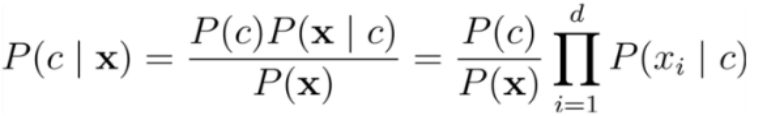
（2）请给出后验概率P(c|x)和贝叶斯判定准则的具体式子。

（3）使用贝叶斯决策对鱼的类别进行预测（要求给出具体计算过程及计算结果）

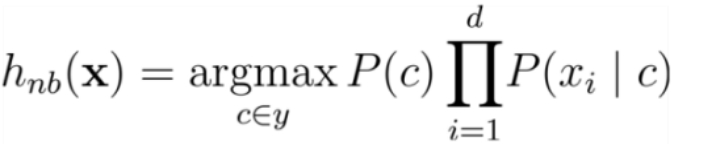
答案：

（1）朴素贝叶斯分类器采用“属性条件独立性假设”，对已知类别，假设所有属性相互独立，换言之，假设每个属性独立地对分类结果发生影响。

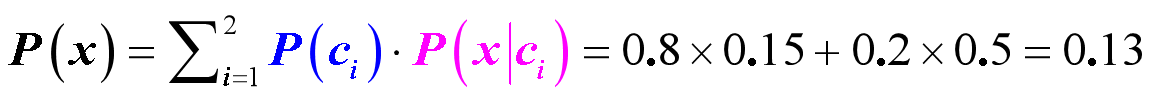
（2）基于属性条件独立性假设，P(cïx)为：

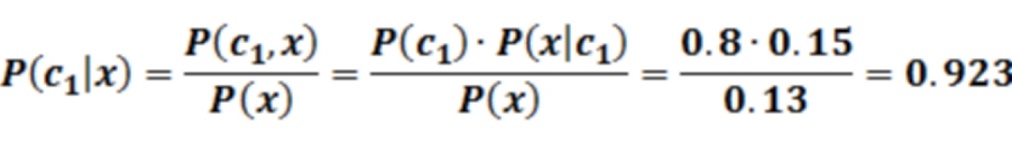


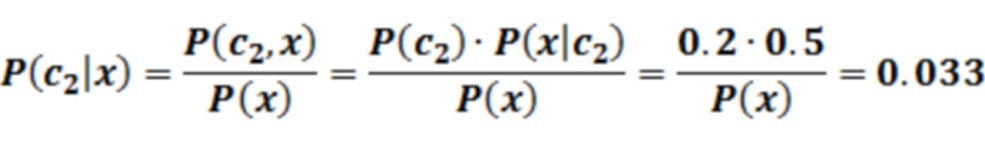
由于对所有类别来说，P(x)相同，因此，贝叶斯判定准则有：



（3）由已知条件可知 鲈鱼：P(c1) = 0.8；鲑鱼： P(c2) = 0.2







因此：

所以，鱼的类别判定为鲈鱼.