作业 3：朴素贝叶斯分类、k 近邻分类

1. 根据顾客的年龄、收入、是否是学生来预测是否购买电脑。基于以下训练集请**先构造**朴素贝叶斯分类器，再给出测试集的预测结果。注意：年龄为连续属性，正态分布概率函数为$p(\mu,\sigma)=\frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi}\sigma}$，方差计算采用 $\sigma^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\mu)^2$，其中$\mu$为均值，其他两个类别属性取值分别为“收入”={高、中、低}，“学生”={是、否}。

训练集

编号	类别：是否买电脑	年龄	收入	学生
1	否	22	高	否
2	否	24	高	否
3	是	57	中	否
4	是	54	低	是
5	否	50	低	是
6	是	37	低	是
7	是	34	高	是
8	否	48	中	否

测试集

编号	类别：是否买电脑	年龄	收入	学生
9	是	30	低	是
10	是	53	中	是
11	是	24	中	是

a.分类器构造，即计算所有可能用到的概率：

注意：该题的预测目标是**“是否买电脑”，即这一列才是类别标签**，最后一列“学生”只是其中一个属性。

计算每个类的先验概率：

$P(C_i)=|C_{i,D}|/|D|$，$P(C_{是})=4/8=0.5$，$P(C_{否})=4/8=0.5$

计算离散属性“收入”、“是否是学生”的各个取值在每个类的概率：

$P(收入=低\ /\ C_{是})=2/4=0.5$，$P(收入=低\ /\ C_{否})=1/4=0.25$

$P(收入=中\ /\ C_{是})=1/4=0.25$，$P(收入=中\ /\ C_{否})=1/4=0.25$

$P(收入=高\ /\ C_{是})=1/4=0.25$，$P(收入=高\ /\ C_{否})=2/4=0.5$

$P(学生=否\ /\ C_{是})=1/4=0.25$，$P(学生=否\ /\ C_{否})=3/4=0.75$

$P(学生=是\ /\ C_{是})=3/4=0.75$，$P(学生=是\ /\ C_{否})=1/4=0.25$

计算连续属性“年龄”对每个类的均值和标准差

$\mu_{年龄,C_{是}}=\frac{1}{4}(57+54+37+34)=45.5$，

$\sigma^2_{年龄,C_{是}} = \frac{1}{4-1}((57-45.5)^2+(54-45.5)^2+(37-45.5)^2+(34-45.5)^2) = 136.33$，

$\sigma_{年龄,C_{是}} = 11.68$ 。

$\mu_{年龄,C_{否}} = \frac{1}{4}(22+24+50+48) = 36$，

$\sigma^2_{年龄,C_{否}} = \frac{1}{4-1}((22-36)^2+(24-36)^2+(50-36)^2+(48-36)^2) = 226.67$，

$\sigma_{年龄,C_{否}} = 15.06$。

构造完毕，得到以下表格：

		是否买电脑？=是（P=0.5）	是否买电脑？=否 (P=0.5)
年龄（均值、标准差）		（45.5， 11.68）	（36， 15.06）
收入	高	0.25	0.5
	中	0.25	0.25
	低	0.5	0.25
学生	是	0.75	0.25
	否	0.25	0.75

b. 基于以上各概率，预测顾客是否会买电脑（通过查询以上概率计算测试样本$\boldsymbol{P(c)P(x|c)}$并给出预测结果）。

测试：计算 $\boldsymbol{P(c)P(x|c) = P(c)\prod_i P(x_i|c)}$， 选择对应值最大的类。

样本 9：

根据$\mu_{年龄，C_{是}} = 45.5$, $\sigma_{年龄，C_{是}} = 11.68$ 计算得到 $P(年龄/C_{是}) = \frac{\exp\left(-\frac{(30-45.5)^2}{2\cdot 136.33}\right)}{\sqrt{2\pi}\cdot 11.68} = 0.0142$

根据$\mu_{年龄，C_{否}} = 36$, $\sigma_{年龄，C_{否}} = 15.06$ 计算得到 $P(年龄/C_{否}) = \frac{\exp\left(-\frac{(30-36)^2}{2\cdot 226.67}\right)}{\sqrt{2\pi}\cdot 15.06} = 0.0245$

根据 $P(收入=低/C_{是})=0.5$， $P(学生=是/C_{是})=0.75$ ，得到

$$P(X/C_{是}) = P(年龄/C_{是})P(收入=低/C_{是})P(学生=是/C_{是}) = 0.0142\cdot 0.5\cdot 0.75 = 0.0053$$

根据 $P(收入=低/C_{否})=0.25$， $P(学生=是/C_{否})=0.25$ ，得到

$$P(X/C_{否}) = P(年龄/C_{否})P(收入=低/C_{否})P(学生=是/C_{否}) = 0.0245\cdot 0.25\cdot 0.25 = 0.0015$$

$$P(C_{是})P(X/C_{是}) = 0.5\cdot 0.0053 = 0.00265$$

$$P(C_{否})P(X/C_{否}) = 0.5\cdot 0.0015 = 0.00075$$

$$P(C_{是})P(X/C_{是}) > P(C_{否})P(X/C_{否})$$

预测“是”，实际“是”，正确。

样本 10：

$$P\left(年龄\ /\ C_{是}\right)=\frac{\exp\left(-\frac{(53-45.5)^2}{2\cdot 136.33}\right)}{\sqrt{2\pi}\cdot 11.68}=0.0278$$

$$P\left(年龄\ /\ C_{否}\right)=\frac{\exp\left(-\frac{(53-36)^2}{2\cdot 226.67}\right)}{\sqrt{2\pi}\cdot 15.06}=0.0140$$

$P\left(收入=中\ /\ C_{是}\right)=0.25$，$P\left(学生=是\ /\ C_{是}\right)=0.75$ ，得到

$P\left(X\ /\ C_{是}\right)=0.0278\cdot 0.25\cdot 0.75=0.0052$，

$P\left(收入=中\ /\ C_{否}\right)=0.25$，$P\left(学生=是\ /\ C_{否}\right)=0.25$，得到

$P\left(X\ /\ C_{否}\right)=0.0140\cdot 0.25\cdot 0.25=0.0009$

$P\left(C_{是}\right)P\left(X\ /\ C_{是}\right)=0.5\cdot\ 0.0052=0.0026$，$P\left(C_{否}\right)P\left(X\ /\ C_{否}\right)=0.5\cdot 0.0009=0.00045$，

$$P\left(C_{是}\right)P\left(X\ /\ C_{是}\right)>P\left(C_{否}\right)P\left(X\ /\ C_{否}\right)$$

预测“是”，实际“是”，正确。

样本 11：

$$P\left(年龄\ /\ C_{是}\right)=\frac{\exp\left(-\frac{(24-45.5)^2}{2\cdot 136.33}\right)}{\sqrt{2\pi}\cdot 11.68}=0.0063,$$

$$P\left(年龄\ /\ C_{否}\right)=\frac{\exp\left(-\frac{(24-36)^2}{2\cdot 226.67}\right)}{\sqrt{2\pi}\cdot 15.06}=0.0193$$

$P\left(收入=中\ /\ C_{是}\right)=0.25$，$P\left(学生=是\ /\ C_{是}\right)=0.75$，得到

$$P\left(X\ /\ C_{是}\right)=0.0063\cdot 0.25\cdot 0.75=0.0012$$

$P\left(收入=中\ /\ C_{否}\right)=0.25, P\left(学生=是\ /\ C_{否}\right)=0.25$ ，得到

$P\left(X\ /\ C_{否}\right)=0.0193\cdot 0.25\cdot 0.25=0.00121$

$P\left(C_{是}\right)P\left(X\ /\ C_{是}\right)=0.5\cdot\ 0.0012=0.0006$，$P\left(C_{否}\right)P\left(X\ /\ C_{否}\right)=0.5\cdot 0.00121=0.00061$

$$P\left(C_{否}\right)P\left(X\ /\ C_{否}\right)>P\left(C_{是}\right)P\left(X\ /\ C_{是}\right)$$

预测“否”，实际“是”，错误。

2. 如何用 k 近邻算法对以上数据集进行分类？给出实现步骤（包括选用哪种距离度量，k 的设定等），但不用计算。

   答：K 近邻算法需要计算样本之间的距离/相似度，这里总共有三个特征，其中“年龄”为连续型特征，“收入”为等级型特征，“学生”是类别型特征。
   第一步是把“收入”和“学生”转换成连续特征。“收入”的高、中、低可以分别用 1，0.5，0.1 来替换，“学生”的“是”和“否”可以用长度为 2 的独热编码[1, 0] 和 [0, 1] 来代替，比如把测试例 9 变换成[30, 0.1, 1, 0]。
   由于距离的计算与特征取值范围相关，考虑把所有特征线性变化到[0, 1]区间。通过前一步操作，只有年龄的取值范围不在[0,1]区间，所以只要把这个特征进行 0-1 规范化。min-max 规范化需要确定原数据集的最小值和最大值。由于测试样本（新样本）的“年龄”可能大于/小于训练集中看到的“年龄”范围，所以统一把“最大年龄”设置成 60，大于 60 的用 60 替代，“最小年龄”设成 10，小于 10 的用 10 代替，然后进行 0-1 规范化。利用以上方法，测试例 9 中的年龄 30 被映射到 0.4。
   给定一个测试样本，进行以上两步变换后，计算其到所有训练样本的距离，这里没有特殊要求，考虑通用的欧式距离。选择 K 个距离最小的样本作为近邻，把该样本标记为近邻中出现最多的类别。由于该数据集很小，K 的取值可以从 2，3，4 中尝试，最后选择验证集上准确率最高的 K。

3. 朴素贝叶斯的核心假设是什么？有什么优缺点？
   答：朴素贝叶斯假设给定类别情况下所有属性之间相互独立。
   这个假设的优点是简化问题，降低“维度灾难”导致的过拟合风险。缺点是不能描述属性之间的相关性，当实际问题中存在相互关联的属性时，模拟准确率不够理想。