Homework 2

张思源　　*21110850018*

October 15, 2021

1 Ex1

某地成年人肥胖者 (A_1) 占 10%, 中等者 (A_2) 占 82%, 瘦小者 (A_3) 占 8%, 又有肥胖者, 中等者, 瘦小者高血压病的概率分别为 20%,10%,5%.
(1). 求该地成年人患高血压的概率.
(2). 若知某人患高血压, 则他最可能属于哪种体型.

Sol 1.1 记患高血压和不患高血压分别为 B_1, B_2.
(1) 由题意,$P(B_1|A_1) = 0.2, P(B_1|A_2) = 0.1, P(B_1|A_3) = 0.05$. 故由全概率公式:

$$P(B_1) = \sum_{i=1}^{3} P(B_1|A_i)P(A_i).$$

代入数据计算得 $P(B_1) = 0.106 = 10.6\%$.
即该地成年人患高血压的概率为 10.6%.
(2) 由题意及贝叶斯公式, 该人肥胖的概率为:

$$P(A_1|B_1) = \frac{P(B_1|A_1)P(A_1)}{P(B_1)} = \frac{0.2 \times 0.1}{0.106} = 0.1887,$$

该人中等的概率为:

$$P(A_2|B_1) = \frac{P(B_1|A_2)P(A_2)}{P(B_1)} = \frac{0.1 \times 0.82}{0.106} = 0.7736,$$

该人瘦小的概率为:

$$P(A_3|B_1) = \frac{P(B_1|A_3)P(A_3)}{P(B_1)} = \frac{0.05 \times 0.08}{0.106} = 0.0377,$$

所以该人最可能是中等体型.

2 Ex2

如教材 3.54 式和 3.56 式, 分别写出如下有向图和无向图对应的概率分布.

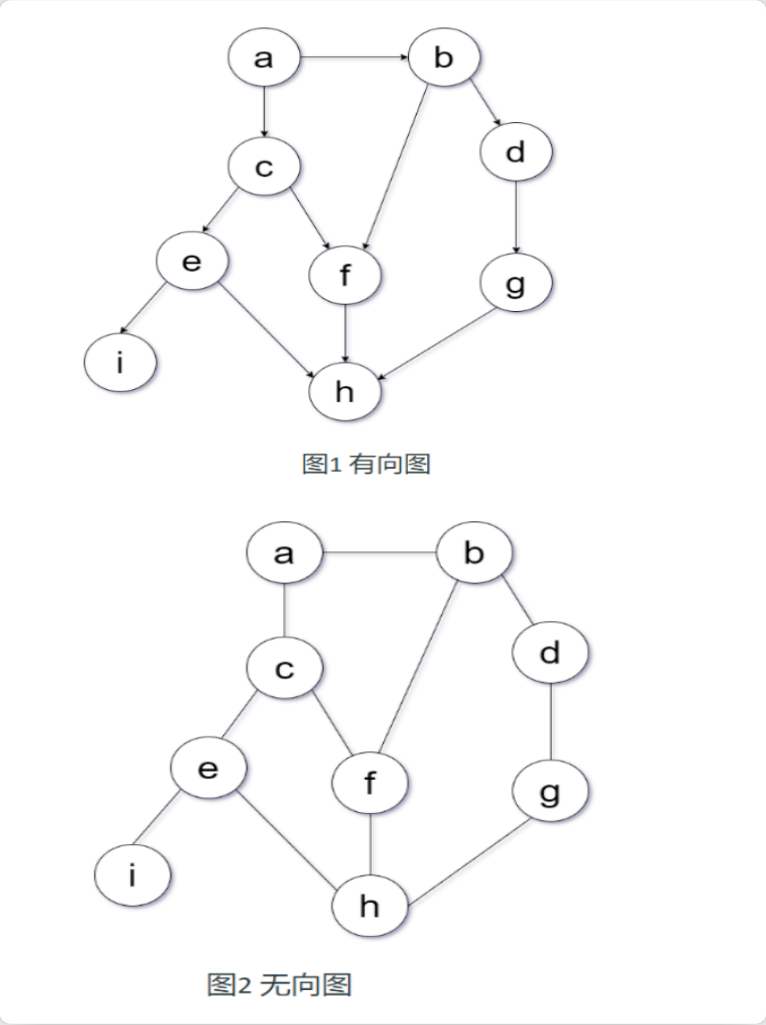

Figure 1: Ex2

Sol 2.1 对于有向图,

$$p(a,b,c,d,e,f,g,h,i) = p(a)p(b|a)p(c|a)p(d|b)p(e|c)p(f|b,c)p(g|d)p(h|e,f,g)p(i|e).$$

对于无向图, 可以看出对于本模型, 没有出现除了二阶子图之外的团, 故其联合概率分布可写作

$$\begin{aligned} p(a,b,c,d,e,f,g,h,i) =& \frac{1}{Z}\phi^{(1)}(a,b)\phi^{(2)}(a,c)\phi^{(3)}(b,f)\phi^{(4)}(b,d)\phi^{(5)}(c,e) \\ & \phi^{(6)}(e,h)\phi^{(7)}(g,h)\phi^{(8)}(d,g)\phi^{(9)}(e,i)\phi^{(10)}(c,f)\phi^{(11)}(f,h). \end{aligned}$$

3 Ex3

取值为 0,1,2,3,4,5 的概率分别为 1/2,1/4,1/8,1/16,1/32,1/32. 求其香农熵.

Sol 3.1 由香农信息熵的定义:

$$H(x) = -\sum_{i=0}^{5} P(x=i)\log_2(P(x=i)) = \frac{31}{16}.$$

4 Ex4

叙述 KL 散度和交叉熵定义, 并给出自己的理解.

Sol 4.1 *KL-散度: 设 p,q 是两个随机分布, 则对于离散情况,$D_{KL}(p||q)=\sum_{k=1}^{K}p_k\log_2\frac{p_k}{q_k}$, 对于连续情况,$D_{KL}(p||q)=\int_{\mathcal{X}}p(x)\log_2\frac{p(x)}{q(x)}dx$. KL-散度度量了使用基于 q 的分布来编码服从 p 的分布的样本所需的**额外的**平均比特数, 之所以是额外是因为 p 与 q 的不匹配.*
*交叉熵: 设 p,q 是两个随机分布, 则对于离散情况,$H(p||q)=-\sum_{k=1}^{K}p_k\log_2 q_k$, 对于连续情况,$H(p||q)=-\int_{\mathcal{X}}p(x)\log_2(q(x))dx$. 交叉熵度量了使用基于 q 的分布来编码服从 p 的分布的样本所需的平均比特数 (**这里不是额外的!!!**), 从公式来看可以写作 $D_{KL}(p||q)=H(p||q)-H(p)$, 这反映了上述"额外"二字.*

5 Ex5

说明每种分类器的含义以及区别并利用 sklearning 自带的数据集, 自行划分训练集和测试集 (无需验证集), 使用第一种分类器 (naive_bayes.GaussianNB), 并且可视化结果, 说明在不同数据类型下分类器效果的差别.

Sol 5.1 (1) *对于不同的朴素贝叶斯分类器, 其含义及区别分别为:*

- *naive_bayes.GaussianNB: 高斯朴素贝叶斯, 特征变量是连续变量且符合高斯分布:*

$$p(x_i|y)=\frac{1}{\sqrt{2\pi\sigma_y^2}}\exp(-\frac{(x_i-\mu_y)^2}{2\sigma_y^2)})$$

- *naive_bayes.BernoulliNB: 多元伯努利朴素贝叶斯, 模型适用于多元伯努利分布, 即每个特征都是二值变量, 如果不是二值变量, 该模型可以先对变量进行二值化, 特征变量满足分布:*

$$p(x_i|y)=p(i|y)x_i+(1-p(i|y))(1-x_i)$$

- *naive_bayes.MultinomialNB: 多项朴素贝叶斯, 特征变量是离散变量, 符合多项分布:*

$$\hat{\theta}_{yi}=\frac{N_{yi}+\alpha}{N_y+\alpha n}$$

- *naive_bayes.ComplementNB: 是 MultinomialNB 模型的一个变种, 实现了补码朴素贝叶斯 (CNB) 算法.CNB 是标准多项式朴素贝叶斯 (MNB) 算法的一种改进, 比较适用于不平衡的数据集.CNB 使用来自每个类的补数的统计数据来计算模型的权重:*

$$\hat{\theta}_{ci}=\frac{\alpha_i+\sum_{j:y_j\neq c}d_{ij}}{\alpha+\sum_{j:y_j\neq c}\sum_k d_{kj}}$$

$$w_{ci}=\log\hat{\theta_{ci}}$$

$$\hat{w}_{ci}=\frac{w_{ci}}{\sum_j|w_{cj}|}$$

(2)

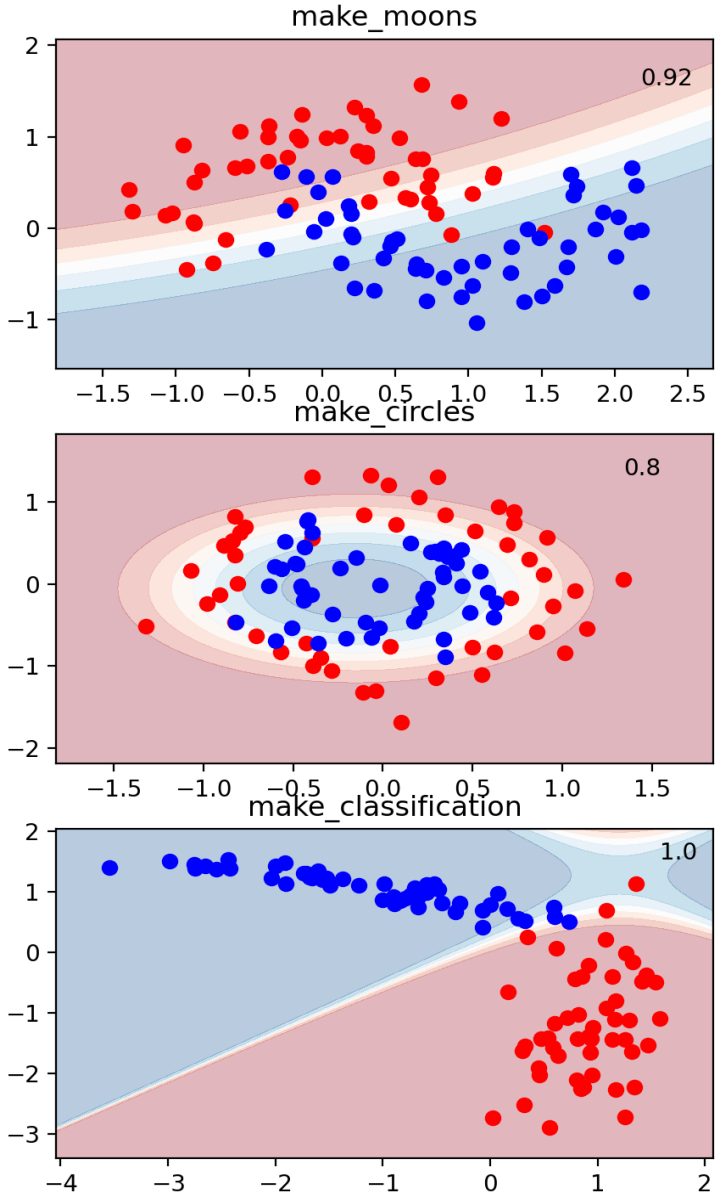

Figure 2: numerical experiment 1

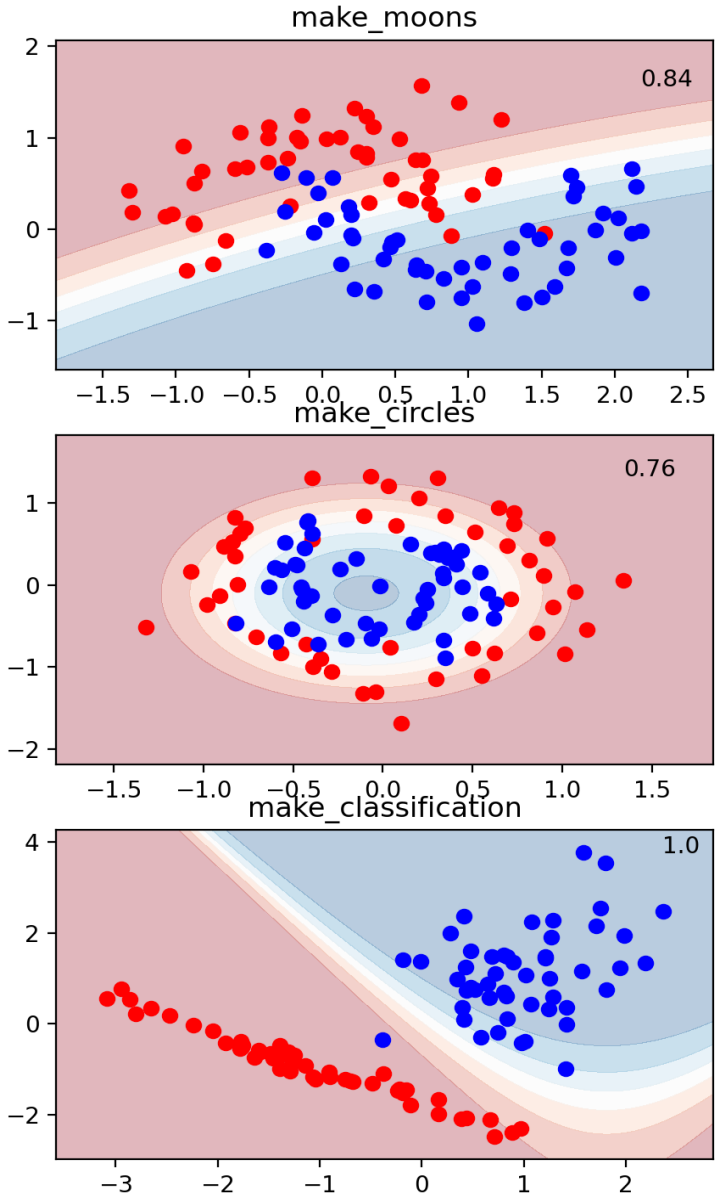

Figure 3: numerical experiment 2

可以看出，在 *naive_bayes.GaussianNB* 分类器下，对于近似线性可分的 *make_moons* 数据集和线性可分的 *make_classification* 数据集的分类效果较好，对非线性可分的 *make_circles* 数据集的分类效果较差. 月牙形数据集的分类边界大概为正弦函数 $\frac{1}{4}$ 的曲线，这与其正弦形的形状可能有关；圆环形数据集的分类边界大概为圆环，且内部为蓝色，外部为红色，这与数据的分布基本一致；线性可分数据集的分类边界大概为线性函数的组合，这可能是与其数据的生成方式有关，同时由于该数据集的线性可分性，其具有最好的可分性质，换言之，分类器在测试集上的准确率

最高.

一方面，这可能是因为对于线性可分的数据集，其先验概率的计算 (或者说估计) 更为接近高斯分布；另一方面，这可能是因为对于 *make_moons* 和 *make_classification* 数据集其每一特征重要性接近，或者说权重相当.

References

[1] 李航. 统计学习方法 [M]. 清华大学出版社, 2012.

[2] Goodfellow, Ian, et al. Deep Learning[M]. MIT Press, 2016.