中国科学院大学 2020 秋季《模式识别与机器学习》课程

随堂测 1

1. 朴素贝叶斯是一个（A， D）。
   （A） 低方差模型 （B） 高方差模型 （C）判别式模型 （D） 生成式模型

2. 对决策树模型，关于其超参数“树的最大深度”，下面说法哪些正确？ A，C
   （A） 如果验证准确率相同，值越低越好；
   （B） 如果验证准确率相同，值越高越好；
   （C） 值增加可能会导致过拟合；
   （D） 值增加可能会导致欠拟合。

3. 在 L1 正则的线性回归模型中，如果正则参数很大，会发生什么？ A
   （A）一些系数将变为零
   （B）一些系数将接近零，但不是绝对零

4. 对多项式回归而言，哪一项对过拟合和欠拟合影响最大？A
   （A） 多项式的阶数
   （B） 是否通过矩阵求逆/梯度下降学习权重
   （C） 高斯噪声方差
   （D） 每一次训练的输入个数固定

5. 使用梯度下降训练 Logistic 回归分类器后，您发现它对训练集欠拟合，在训练集或验证集上没有达到所需的性能。以下哪些项可能是有希望采取的步骤？ C、D
   （A）采用其他优化算法，因为梯度下降得到的可能是局部最小值
   （B）减少训练样本
   （C）增加多项式特征值
   （D）改用较多隐含结点的神经网络模型

6. 在 Logistic 回归中，关于一对其他（One vs. Rest）方法，以下哪个选项是正确? B（A 也可以）
   （A）我们需要在C类分类问题中拟合C个模型
   （B）我们需要拟合$C-1$个模型来分类C类
   （C）我们只需要拟合 1 个模型来分类C类
   （D）这些都不是

7. SVM 的有效性取决于：D
   （A）核函数选择
   （B）核函数的参数
   （C）软边距参数C
   （D）以上所有

8. 当出现下述哪种情况时，SVM 性能不佳？ C
   （A）数据线性可分

（B）数据干净
（C）数据有噪声

9. 假设在 SVM 中使用具有高γ（核函数宽度的倒数）值的 RBF 核函数。这意味着什么？
C
（A）模型将考虑离超平面很远的点；
（B）模型将只考虑离超平面很近的点；
（C）模型不受点到超平面距离的影响。
（D）以上都不对

10. 假设您采用一个线性 SVM 模型来处理某个任务，并且知道这个 SVM 模型是欠拟合的。下列哪些方法可以提升该模型性能？ C、E
（A）减少训练样本
（B）增加训练样本
（C）增加特征
（D）减少特征
（E）增加参数C
（F）减少参数C

11. 如果我使用数据集的所有特征，在训练集的准确率达到 100%，而验证集的准确率为 70%，那么我应该注意什么？ C
（A）模型是欠拟合的；
（B）模型是完美的；
（C）模型是过拟合的。

随堂测2

1. 假设某个地区细胞识别中正常（$y=1$）和异常（$y=0$）两类的先验概率分别为：正常状态：$P(y=1)=0.95$，异常状态$P(y=0)=0.05$。现有一待识别的细胞，其观察值为$\boldsymbol{x}$，已知$p(x|y=1)=0.2$，$p(x|y=0)=0.5$。基于最小错误率的贝叶斯决策，对该细胞进行分类，请写出必要的计算依据。

解答：$P(y|x)=\frac{P(y)p(x|y)}{p(x)}$

$P(y=0|x) \propto P(y=0)p(x|y=0)$=0.025

$P(y=1|x) \propto P(y=1)p(x|y=1)$=0.19

所以将 x 分为正常类。

2. 简述 Fisher 线性判别方法的基本思路，写出准则函数及对应的解。

3．对下列数据，分别采用 K-L 变换和 Fisher 线性判别将特征空间的维数降到一维（设$P(y=1)=0.2$，　$P(y=0)=0.8$）。

$y=1$ 类：$\{(-10,-10)^T,(-5,-4)^T,(-4,-5)^T,(-15,-16)^T,(-6,-5)^T\}$；

$y=0$ 类：$\{(2,2)^T,(2,3)^T,(3,2)^T,(1,2)^T,(2,1)^T\}$；

答：m_1=(-8 -8) ,m_2=(2,2) , m=$P(\omega_1)m_1$+ $P(\omega_2)m_2$=(0 0)计算协方差矩阵：C=$\begin{bmatrix}19.6 & 19.6\\ 19.6 & 20.4\end{bmatrix}$,

求得特征值 39.6 及特征向量[0.6999 0.7143]

投影后：-14.1420　-6.3567　-6.3711　-21.9273　-7.7709

2.8284　3.5427　3.5283　2.1141　2.1285

S_w=C;

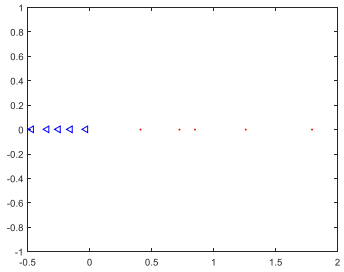
根据课件内容来进行 Fisher 变换：

$S_1=\begin{bmatrix}82 & 90\\ 90 & 102\end{bmatrix}, S_2=\begin{bmatrix}2 & 0\\ 0 & 2\end{bmatrix}, S_w=S_1+S_2=\begin{bmatrix}84 & 90\\ 90 & 104\end{bmatrix}$,

投影向量：$S_w^{-1}(m_1-m_2)=(-0.2201,0.0943)^t$

$y=1$ 类: 1.2579,　0.7233,　0.4088,　1.7925,　0.8491（在图中对应红色的点）

y = 0 类，-0.2516，　-0.1573，　-0.4717，　-0.0315，　-0.3459（在图中对应蓝色的点）



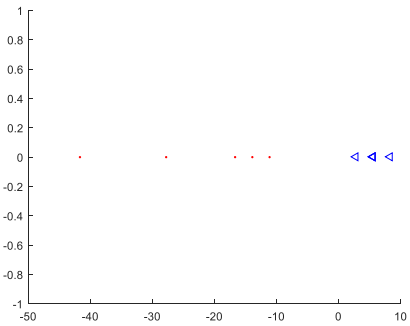
另一种方式是从平均类内距离(或者 S_i 与 C_i 的关系)来考虑：

$C_1=\begin{bmatrix}16.4 & 18\\ 18 & 20.4\end{bmatrix}, C_2=\begin{bmatrix}0.4 & 0\\ 0 & 0.4\end{bmatrix}$, $S_w=0.2C_1+0.8C_2=\begin{bmatrix}3.6 & 3.6\\ 3.6 & 4.4\end{bmatrix}$

投影向量：$S_w^{-1}(m_1-m_2)=(-2.778,0)^t$

$y=1$ 类: -27.7780, -13.8890, -11.1112, -41.6670, -16.6668（在图中对应红色的点）

$y=0$ 类，5.5556, 5.5556, 8.3334, 2.7778, 5.5556 在图中对应蓝色的点）



第三种方式：$S_w=0.2S_1+0.8S_2$, $S_w=0.2S_1+0.8S_2=\begin{bmatrix}18 & 18\\ 18 & 22\end{bmatrix}$,

投影向量：$S_w^{-1}(m_1-m_2)=(-0.5556,0)^t$

学有余力的同学可以想想乘以先验概率的作用。根据先验概率，你觉得这 1 0 个样本是否是所有数据的随机采样还是有偏采样？

4 .给定下列两类模式的样本集合，定义损失函数$E(\boldsymbol{w})=\sum_{i=1}^{N}E_i(\boldsymbol{w})$，其中

$E_i(\boldsymbol{w})=\begin{cases}-\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i y_i & \text{如果}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i y_i<0\\ 0 & \text{如果}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i y_i\geq 0\end{cases}$（$\boldsymbol{x}_i$，$y_i$分别表示第$i$个样本的特征和类别），

利用梯度下降法（步长为 1），求解使损失函数取最小值的$\boldsymbol{w}$。

$y=1$ 类：$\{(0,0,0)^T,(1,0,0)^T,(1,0,1)^T,(1,1,0)^T\}$；

$y=-1$ 类：$\{(0,0,1)^T,(0,1,1)^T,(0,1,0)^T,(1,1,1)^T\}$；

$\omega_1:\{(0,0,0)^T,(1,0,0)^T,(1,0,1)^T,(1,1,0)^T\}$

$\omega_2:\{(0,0,1)^T,(0,1,1)^T,(0,1,0)^T,(1,1,1)^T\}$

答：（其实就是感知器算法）利用梯度下降求解权重更新规则。

1．将属于ω2 的训练样本乘以（-1），并写成增广向量的形式。

$x_1\sim x_4:\{(0,0,0,1)^T,(1,0,0,1)^T,(1,0,1,1)^T,(1,1,0,1)^T\}$

$x_5\sim x_8:\{(0,0,1,-1)^T,(0,1,1,-1)^T,(0,1,0,-1)^T,(1,1,1,-1)^T\}$

（0，0，-1，1，）（0，-1，-1，-1）（0，-1，0，-1）（-1，-1，-1，-1）

第一轮迭代：取 C=1，w(1)= (0 0 0 0)^T

因 $w^T(1)x_1$=(0 0 0 0)(0 0 0 1)^T=0≯0，故 w(2)=w(1)+C$x_①$=(0 0 0 1)^T

因 $w^T(2)x_2$=(0 0 0 1)(1 0 0 1)^T=1>0，故 w(3)=w(2)=(0 0 0 1)^T

因 $w^T(3)x_3$=(0 0 0 1)(1 0 1 1)^T=1>0，故 w(4)=w(3) =(0 0 0 1)^T

因 $w^T(4)x_4$=(0 0 0 1)(1 1 0 1)^T=1>0，故 w(5)=w(4) =(0 0 0 1)^T

因 $w^T(5)x_5$=(0 0 0 1)(0 0 −1 −1)^T=−1≯0，故 w(6)=w(5)+C x_5 =(0 0 −1 0)^T

因 $w^T(6)x_6$=(0 0 −1 0)(0 −1 −1 −1)^T=1>0，故 w(7)=w(6) =(0 0 −1 0)^T

因 $w^T(7)x_7$=(0 0 −1 0)(0 −1 0 −1)^T=0≯0，故 w(8)=w(7)+C x_7 =(0 −1 −1 −1)^T

因 $w^T(8)x_8$=(0 −1 −1 −1)^T (−1 −1 −1 −1)^T=3>0，故 w(9)=w(8)=(0 −1 −1 −1)^T

因为只有对全部模式都能正确判别的权向量才是正确的解，因此需进行第二轮迭代。

$w^T(9)x_1$=−1，w(10)=w(9)+C x_1 =(0 −1 −1 0)^T

$w^T(10)x_2$=0，w(11)=w(10)+C x_2 =(1 −1 −1 1)^T

$w^T(11)x_3$=1，w(12)=w(11)

$w^T(12)x_4$=1，w(13)=w(12)

$w^T(13)x_5$=0，w(14)=w(13) +C x_5 =(1 −1 −2 0)^T

$w^T(14)x_6$=3，w(15)=w(14)

$w^T(15)x_7$=1，w(16)=w(15)

$w^T(16)x_8$=2，w(17)=w(16)

第二轮迭代结束。

5.设有一维空间二次判别函数$g(x)=9x^2+7x+5$，试将其映射为广义齐次线性判别函数，并总结把高次函数（$g(x)=c_0+c_1x+c_2x^2+\cdots+c_nx^n$）映射成其次线性函数的方法。

6．假定给定如下数据集，其中x_1、x_2、x_3为二值随机变量，y为等预测的二值变量。

x_1	x_2	x_3	y
0	0	1	0
0	1	0	0

1	1	0	0
0	0	1	1
1	1	1	1
1	0	0	1
1	1	0	1

(1) 对一个新的输入$x_1 = 0$、$x_2 = 0$、$x_3 = 1$，如果采用朴素贝叶斯分类器，需要知道哪些参数，并采用拉普拉斯平滑估计这些参数。贝叶斯分类器将会怎么样预测y?

答：需要知道 P(y=0),**及** $p(x_i|y=0)$

P(y=0)=（3+1）/（7+2）=4/9， P(y=1)=（4+1）/（7+2）=5/9

P(x1=0|y=0)=(2+1)/（3+2）=3/5，P(x1=1|y=0)=(1+1)/（3+2）=2/5

P(x1=0|y=1)=(1+1)/（4+2）=1/3，P(x1=1|y=0)=(3+1)/（4+2）=2/3

P(x2=0|y=0)=(1+1)/（3+2）=2/5，P(x1=1|y=0)=(2+1)/（3+2）=3/5

P(x2=0|y=1)=(2+1)/（4+2）=1/2，P(x1=1|y=1)=(2+1)/（4+2）=1/2

P(x3=0|y=0)=(2+1)/（3+2）=3/5，P(x3=1|y=0)=(1+1)/（3+2）=2/5

P(x3=0|y=1)=(2+1)/（4+2）=1/2，P(x3=1|y=1)=(2+1)/（4+2）=1/2

可得

$P(y = 1|x_1 = 0\ ,\ \ x_2 = 0\ ,\ \ x_3 = 1\ ) \propto P(y = 1)P(x_1 = 0|y = 1)P(x_2 = 0|y = 1)P(x_3 = 1|y = 1)$

$=\frac{5}{9}\times\frac{1}{3}\times\frac{1}{2}\times\frac{1}{2} = \frac{5}{128}$

$P(y = 0|x_1 = 0\ ,\ \ x_2 = 0\ ,\ \ x_3 = 1\ ) \propto P(y = 0)P(x_1 = 0|y = 0)P(x_2 = 0|y = 0)P(x_3 = 1|y = 0)$

$=\frac{4}{9}\times\frac{3}{5}\times\frac{2}{5}\times\frac{2}{5} = \frac{16}{125*3} > \frac{5}{128}$

所以将预测 y=0.

(2) 假设在给定类别的情况下，x_1、x_2、x_3是独立的随机变量，那么其他分类器如（logistic回归、SVM分类器等）会比朴素贝叶斯分类器表现更好吗？为什么？

答：不会。因为已知独立同分布的前提下，估计$p(x_i|y=0)$**或**$p(x_i|y=1)$，朴素贝叶斯只用3个参数，不用朴素贝叶斯则需要2^3-1=7个参数。若不独立，则其它基于数据本身的判别式分类器效果较好。

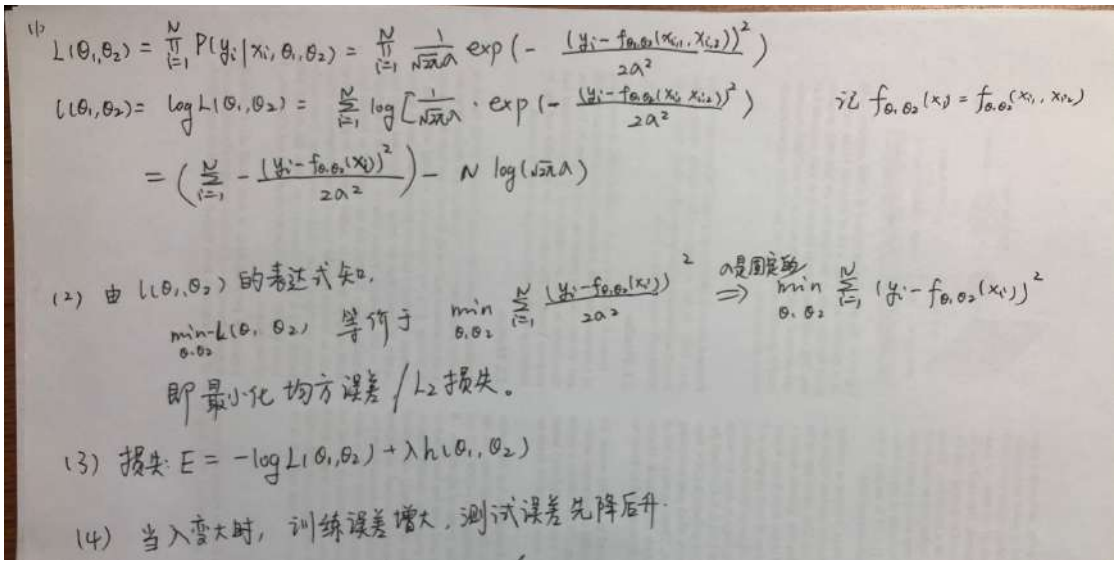
7. 给定数据集$\mathcal{D}=(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_N,y_N)$，其中输入特征$\boldsymbol{x}_i=(x_{i,1},x_{i,2})^{\mathrm{T}}\in\mathbb{R}^2$，输出变量$y_i\in\mathbb{R}$为连续值。考虑模型：$y_i=f_{\theta_1,\theta_2}(x_{i,1},x_{i,2})+e_i$，其中$e_i\sim N(0,\sigma^2)$为独立的高斯噪声。

（1）给出数据产生过程的**对数似然函数**。

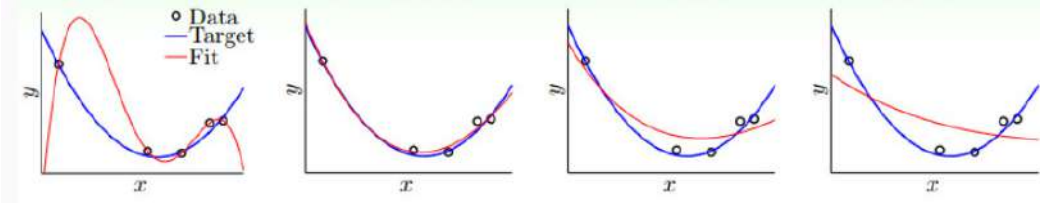
（2）证明负对数似然损失与 L2 损失函数是等价。

（3）假设给定一个函数当f比较复杂时，$h(\theta_1,\theta_2)$较大；当f_{θ_1,θ_2}比较简单时，$h(\theta_1,\theta_2)$较小。采用$h(\theta_1,\theta_2)$作为正则项，用负对数似然函数和正则参数λ的形式给出正则化损失函数的表达式。

（4）对(3)中的损失函数，分析当λ变化时对应训练误差和测试误差的变化情况。



8. 给定训练数据集如下图所示.



(a) (b) (c) (d)

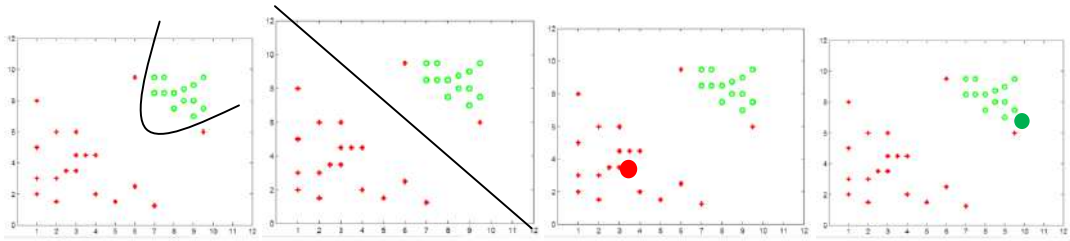
（1） 图上给出了用多项式拟合（红线）的结果，请问此时哪个模型训练误差最小？如果需要选择最佳预测模型的话，应该选择哪个模型？

答：(a)训练 误差最小；最佳预测模型为（b）

（2）考虑L2 正则项约束的线性回归方法，其中正则项系数为λ。则上图四个图中哪个图对应的λ值最大？上面哪个图对应的λ值最小？为什么？

答：λ越大，越倾向于选择简单的模型，训练 误差就越大，因此（d）对应的λ最大；（a）对应的λ对小。

9. 考虑如图给出的训练样本，我们采用二次多项式为核函数，松弛因子为C。请对下列问题做出定性分析，并用一两句话给出原因。



(a) （b） （c） （d）

图1

(1) 当$C \to \infty$时，决策边界会变成什么样？请在图1(a)中画出。

(2) 当$C \to 0$时，决策边界会变成什么样？请在图1(b)中画出。

(3) 你认为上述两种情况，哪个在实际测试时效果会好些？ （b）

(4) 在图中增加一个点，使得当$C \to \infty$时，决策边界会不变。请在图1(c)中画出。

图（c）中的大红点

(5) 在图中增加一个点，使得当$C \to \infty$时，该点会显著影响决策边界。请在图1(d)中画出。大绿点

10．已知正例样本$\boldsymbol{x}_1=(1,2)^{\mathrm{T}},\boldsymbol{x}_2=(2,3)^{\mathrm{T}},\boldsymbol{x}_3=(3,3)^{\mathrm{T}}$，负例样本$\boldsymbol{x}_4=(1,1)^{\mathrm{T}}$, 试用线性支持向量机的对偶算法求最大间隔分离超平面和分类决策函数，并在图中画出分离超平面、间隔边界及支持向量。

答：SVM 的对偶问题为：

$$\max_{\alpha}\sum_{i=1}^{N}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{N}y^iy^j\alpha_i\alpha_j\left(\boldsymbol{x}^i\right)^T\boldsymbol{x}^j$$
$$s.t.\ \alpha_i\geq 0,i=1,..,N,$$
$$\sum_{i=1}^{N}\alpha_iy^i=0.$$

将训练样本代入可得：

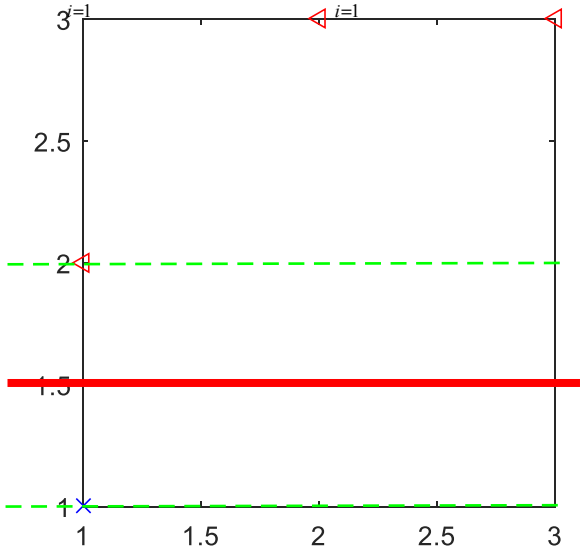
$$\max_{\alpha}\sum_{i=1}^{4}\alpha_i-\frac{5}{2}\alpha_1^2-\frac{13}{2}\alpha_2^2-9\alpha_3^2-\alpha_4^2-8\alpha_1\alpha_2-9\alpha_1\alpha_3+3\alpha_1\alpha_4-15\alpha_2\alpha_3+5\alpha_2\alpha_4+6\alpha_3\alpha_4$$
$$s.t.\ \alpha_i\geq 0,i=1,..,N,$$
$$\sum_{i=1}^{N}\alpha_iy^i=0.$$

求解方法不限（简单的做法就是先通过观察确定支持向量）

可得$\alpha_1^*=2,\alpha_2^*=0,\alpha_3^*=0,\alpha_4^*=2$

由于：

$$\boldsymbol{w}^*=\sum_{i=1}^{N}\alpha_i^*y^i\boldsymbol{x}^i,\ b^*=y^j-\sum_{i=1}^{N}\alpha_i^*y^i\left(\boldsymbol{x}^i\right)^T\boldsymbol{x}^j.$$



所以 $w=(0,2)^T$, $b=3$

分离超平面为红色实线： *x_2-1.5=0;* ***或*** *$2x_2$-3=0*

间隔边界为绿色虚线： *x_2-2=0; x_2-1=0;* ***或*** *$2x_2$-3=1 ， $2x_2$-3=-1*

支持向量：（1，2）（1，1）

11．SVM 可以借助核函数在特征空间学习一个具有最大间隔的超平面。对于两类的分类问题，任意输入x的分类结果取决于下式：

$$< \boldsymbol{w}, \phi(x) > + b = f(x; \boldsymbol{\alpha}, b),$$

其中，$\boldsymbol{w}$和b是分类超平面的参数，$\boldsymbol{\alpha} = [\alpha_1, \dots \alpha_{|SV|}]$表示支持向量（support vector）的系数，SV 表示支持向量集合。我们使用径向基函数定义核函数$K(\cdot,\cdot)$。假设训练数据在特征空间线性可分，SVM 可以完全正确地划分这些训练数据。给定一个测试样本x_{far}，它距离所有训练样本都非常远。试写出$f(x; \boldsymbol{\alpha}, b)$在核特征空间的表达形式，进而证明：$f(x_{far}; \boldsymbol{\alpha}, b) \approx b$。

答：这里将支持向量记为$(\boldsymbol{x}_1, y_1) \dots (\boldsymbol{x}_{|SV|}, y_{|SV|})$，

由于$(\boldsymbol{x}_1, y_1)$是支持向量，所以$b = y_1 - \sum_{i=1}^{|SV|} \alpha_i y_i K(\boldsymbol{x}_1, \boldsymbol{x}_i)$

$f(\boldsymbol{x}; \boldsymbol{\alpha}, b) = \sum_1^{|SV|} \alpha_i y_i K(\boldsymbol{x}_i, \boldsymbol{x}) + b^*$

若测试样本x_{far}**离所有样本都非常远**，则$K(\boldsymbol{x}_i, \boldsymbol{x})$趋于0，所以$f(\boldsymbol{x_{far}}; \boldsymbol{\alpha}, b) = \sum_1^{|SV|} 0 + b \approx 0.$

12、我们用Logistic回归模型解决K类分类问题，假设每个输入样本$\boldsymbol{x} \in \mathbb{R}^d$的后验概率可以表示为：

$$P(Y = k|\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}_k^{\mathrm{T}} \boldsymbol{x})}{1 + \sum_{l=1}^{K-1} \exp(\boldsymbol{w}_l^{\mathrm{T}} \boldsymbol{x})}, \quad k = 1,2, \dots, K-1$$

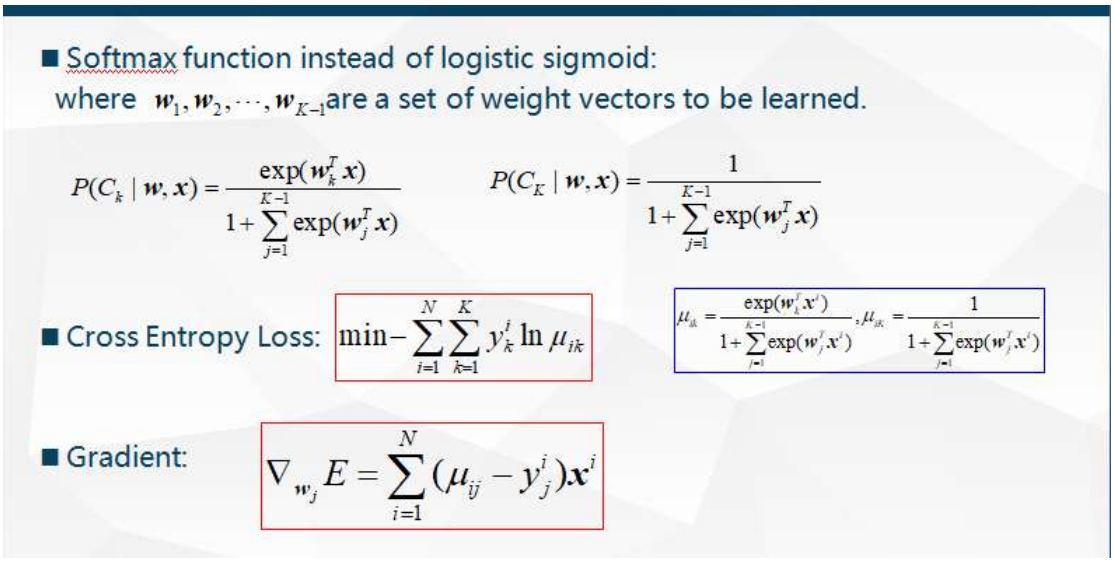
$$P(Y=K|\boldsymbol{x})=\frac{\exp(\boldsymbol{w}_K^{\mathrm{T}}\boldsymbol{x})}{1+\sum_{l=1}^{K-1}\exp(\boldsymbol{w}_l^{\mathrm{T}}\boldsymbol{x})}。$$

通过引入$\boldsymbol{w}_K=\vec{0}$，上式也可以合并为一个表达式。

(1) 该模型的参数是什么？数量有多少？

答：由给定的式子可知，参数是 $\boldsymbol{w}_1,\ldots,\boldsymbol{w}_{K-1}$，每个 $\boldsymbol{w}_i$ 为 d 维，因此参数量共有(K-1)*d 个。

(2) 给定N个训练样本$\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\ldots,(\boldsymbol{x}_N,y_N)\}$，请写出对数似然函数$L(w_1,\ldots,w_{K-1})=\sum_{i=1}^{N}\ln P(Y=y_i|\boldsymbol{x}_i)$的表达形式，并尽量化简。



答：可参考上图。这里假定y_i是独热(one hot)表示，即y_i是 K 维列向量，若$\boldsymbol{x}_i$属于第 k 类，则y_i的第 k 个分量$y_{i,k}$为 1；否则$y_{i,k}$=0.

令：$\mu_{ik}=\frac{\exp(\boldsymbol{w}_k^T\boldsymbol{x}_i)}{1+\sum_{j=1}^{K-1}\exp(\boldsymbol{w}_j^T\boldsymbol{x}_i)}, \mu_{iK}=\frac{1}{1+\sum_{j=1}^{K-1}\exp(\boldsymbol{w}_j^T\boldsymbol{x}_i)}$，则$P(y_i|\boldsymbol{x}_i)=\prod_{j=1}^{K}(\mu_{ik})^{y_{i,k}}$

$$L(w_1,\ldots,w_{K-1})=\sum_{i=1}^{N}\ln P(y_i|\boldsymbol{x}_i)=\sum_{i=1}^{N}\sum_{k=1}^{K}y_{i,k}\ln\mu_{ik}$$

$$=\sum_{i=1}^{N}\sum_{k=1}^{K}y_{i,k}\ln\mu_{ik}$$

(3) 请计算L相对于每个$\boldsymbol{w}_k$的梯度。

答:

$$\ln \mu_{ij} = \boldsymbol{w}_j^T \boldsymbol{x}^i - \ln\left(1+\sum_{m=1}^{K-1}\exp(\boldsymbol{w}_m^T \boldsymbol{x}^i)\right)$$

$$if\ k=j, \frac{\partial \ln \mu_{ij}}{\partial \boldsymbol{w}_k} = \boldsymbol{x}^i - \frac{\boldsymbol{x}^i \exp(\boldsymbol{w}_k^T \boldsymbol{x}^i)}{1+\sum_{m=1}^{K-1}\exp(\boldsymbol{w}_m^T \boldsymbol{x}^i)} = \boldsymbol{x}^i - \boldsymbol{x}^i \mu_{ik}$$

$$if\ k\neq j, \frac{\partial \ln \mu_{ij}}{\partial \boldsymbol{w}_k} = \frac{\boldsymbol{x}^i \exp(\boldsymbol{w}_k^T \boldsymbol{x}^i)}{1+\sum_{m=1}^{K-1}\exp(\boldsymbol{w}_m^T \boldsymbol{x}^i)} = -\boldsymbol{x}^i \mu_{ik}$$

$$\frac{\partial \mathrm{L}}{\partial \mathbf{w_k}} = \sum_{i=1}^{N}\sum_{j=1}^{K} y_{i,j}\frac{\partial \ln \mu_{ij}}{\partial \mathbf{w_k}} = \sum_{i=1}^{N} y_{i,k}\boldsymbol{x}_i - \left(\sum_{j=1}^{K} y_{i,j}(\boldsymbol{x}_i \mu_{ik})\right)$$

由于$\sum_{m=1}^{K} y_{i,j}=1$,

所以$\frac{\partial \mathrm{L}}{\partial \mathbf{w_k}} = \sum_{i=1}^{N}\left(y_{i,k}\boldsymbol{x}_i - \boldsymbol{x}_i \mu_{ik}\right) = \sum_{i=1}^{N}\boldsymbol{x}_i\left(y_{i,k} - \mu_{ik}\right)$

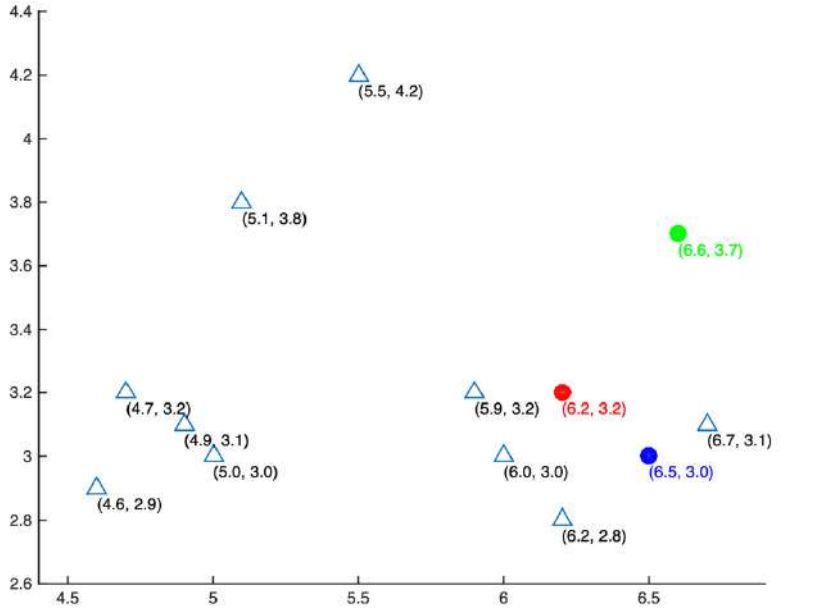
(4) 如果我们加入正则化项，定义新的目标函数为：

$$J(\boldsymbol{w}_1,\dots,\boldsymbol{w}_{K-1}) = L(\boldsymbol{w}_1,\dots,\boldsymbol{w}_{K-1}) - \lambda\sum_{l=1}^{K}\|\boldsymbol{w}_l\|_2^2$$

请计算J相对于每个$\boldsymbol{w}_k$的梯度。

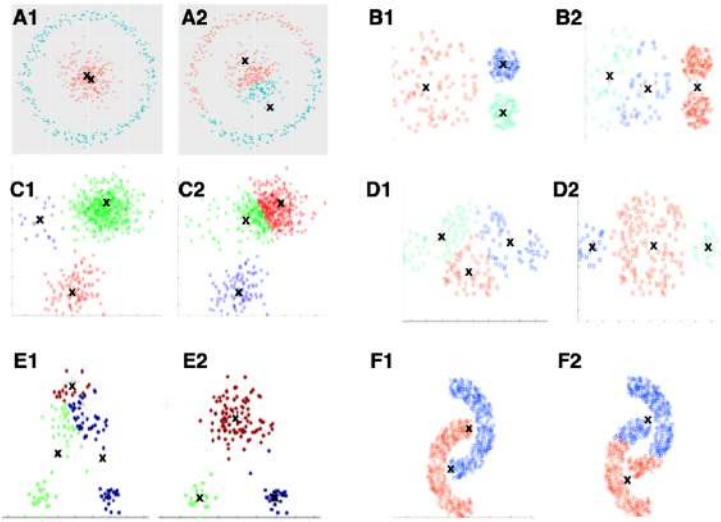
答：$\frac{\partial \mathrm{J}}{\partial \mathbf{w_k}} = \frac{\partial \mathrm{L}}{\partial \mathbf{w_k}} - 2\lambda\mathbf{w_k}$

13. 对如图所示的数据集，采用 K 均值聚类。设$K=3$，3 个聚类中心分别为$\mu_1=(6.2,3.2)^T$(红色)，$\mu_2=(6.6,3.7)^T$(绿色)，$\mu_3=(6.5,3.0)^T$（蓝色）。请给出一次迭代后属于第一簇的样本及更新后的簇中心（保留两位小数）。



$$\mathbf{X} = \begin{bmatrix} 5.9 & 3.2 \\ 4.6 & 2.9 \\ 6.2 & 2.8 \\ 4.7 & 3.2 \\ 5.5 & 4.2 \\ 5.0 & 3.0 \\ 4.9 & 3.1 \\ 6.7 & 3.1 \\ 5.1 & 3.8 \\ 6.0 & 3.0 \end{bmatrix}$$

14. 下图给出了 6 个数据集 A、B、C、D、E、F 用两种聚类算法得到的聚类结果，已知其中一种聚类算法是 K 均值聚类。请问对每个数据集，哪个最可能是 K 均值聚类的结果。如果 K 均值聚类结果不够理解，对每个数据集，你建议采用哪种聚类算法？



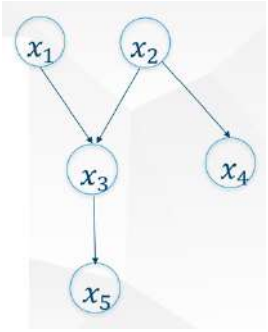
答：K 均值聚类的结果：A2、B2、C2、D1、E2、F2

15. 请给出半监督学习的假设。

16. 根据如图所示的概率图模型，判断下述表述是否成立（用概率分布表示或贝叶斯球的运动说明均可）：

(1) 在给定x_3的条件下，x_1和x_4独立。

(2) 在给定x_2的条件下，x_4和x_5独立。



答：根据上述概率图模型：

$p(x_1,x_2,x_3,x_4,x_5) = p(x_1)p(x_2)p(x_3|x_1,x_2)p(x_4|x_2)p(x_5|x_3)$

（1）在给定x_3的条件下，从x_1出发的贝叶斯球，到达x_3时被反弹到x_2，再从x_2到x_4，所以在给定x_3的条件下，x_1和x_4不独立。

$$p(x_1,x_3,x_4) = p(x_1)\sum_{x_2} p(x_2)p(x_3|x_1,x_2)p(x_4|x_2)\sum_{x_5} p(x_5|x_3)$$

$$= p(x_1)p(x_3|x_1)p(x_4) = p(x_1,x_3)p(x_4)$$

由上述表达式不能得到$p(x_4|x_3,x_1) = \frac{p(x_1,x_3,x_4)}{(x_1,x_3)} = p(x_4) \neq p(x_4,|x_3)$，因此证明在给定$x_3$的条件下，$x_1$和$x_4$不独立。

（2）在给定x_2的条件下，从x_4出发的贝叶斯球，被x_2阻断，不能到达x_3和x_5，所以在给定x_2的条件下，x_4和x_5独立。

$p(x_2,x_4,x_5)$

$$= p(x_4|x_2)p(x_2)\sum_{x_1} p(x_1)\sum_{x_3} p(x_5|x_3)\,p(x_3|x_1,x_2)$$

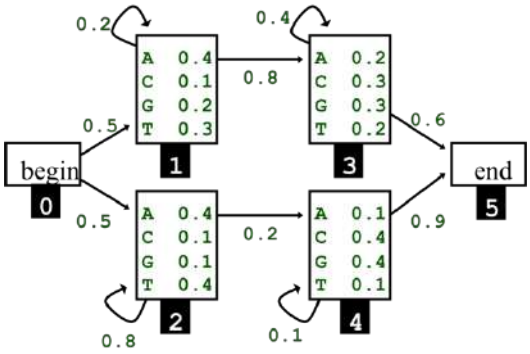
$$= p(x_4|x_2)p(x_2)\sum_{x_1} p(x_1)p\,(x_5|x_1,x_2)$$

$$= p(x_4|x_2)p(x_2)\,p(x_5|x_2)$$

$$p(x_4,x_5|x_2) = \frac{p(x_2,x_4,x_5)}{p(x_2)} = p(x_4|x_2)\,p(x_5|x_2)$$

所以在给定x_2的条件下，x_4和x_5独立。

17. **给定如图所示** HMM：



(1) 采用前向算法计算序列 AGTT 出现的概率。

(2) 计算观测 TATA 最可能的状态序列（步骤、每一步的概率和最佳路径）。

答：初始概率$\boldsymbol{\pi} = (1,0,0,0,0,0)$，

转移概率矩阵$\mathbf{A} = \begin{bmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0.2 & 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0 & 0.2 & 06 \\ 0 & 0 & 0 & 0.4 & 0 & 0.6 \\ 0 & 0 & 0 & 0 & 0.1 & 0.9 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$，

（去掉状态 0/begin 和状态 4/end）

发放概率矩阵$\mathbf{B} = \begin{bmatrix} 0.4 & 0.1 & 0.2 & 0.3 \\ 0.4 & 0.1 & 0.1 & 0.4 \\ 0.2 & 0.3 & 0.3 & 0.2 \\ 0.1 & 0.4 & 0.4 & 0.4 \end{bmatrix}$。

（1） 前向算法计算观测序列 begin AGTT end 的概率

1）初始化：$\alpha_1(y_1) = p(x_1|y_1)\pi_i$，$x_1 = begin$

$\alpha_1(0) = 1$，$\alpha_1(1) = \alpha_1(2) = \alpha_1(3) = \alpha_1(4) = \alpha_1(5) = 0$；

2）根据$\alpha_{t+1}(y_{t+1}) = \left(\sum_{y_t} \alpha_t(y_t) a_{y_t,y_{t+1}}\right) p(x_{t+1}|y_{t+1})$，

当$t = 1$时，$\alpha_2(y_2) = \left(\sum_{y_1} \alpha_1(y_1) a_{y_1,y_2}\right) p(x_2|y_2)$，$x_2 = A$，

$\alpha_2(0) = 0$

$\alpha_2(1) = \left(\sum_{y_1} \alpha_1(y_1) a_{y_1,1}\right) p(A|1) = \alpha_1(0) \times a_{0,1} \times p(A|1) = 1 \times 0.5 \times 0.4 = 0.2$；

$\alpha_2(2) = \left(\sum_{y_1} \alpha_1(y_1) a_{y_1,2}\right) p(A|2) = \alpha_1(0) \times a_{0,2} \times p(A|2) = 1 \times 0.5 \times 0.4 = 0.2$；

$\alpha_2(3) = \alpha_3(4) = \alpha_3(5) = 0$

当$t = 2$时，$\alpha_3(y_3) = \left(\sum_{y_2} \alpha_2(y_2) a_{y_2,y_3}\right) p(x_3|y_3)$，$x_3 = G$，

$\alpha_3(0) = 0$

$\alpha_3(1) = \left(\sum_{y_2} \alpha_2(y_2) a_{y_2,1}\right) p(G|1) = \left(\alpha_2(1) \times a_{1,1}\right) \times 0.2 = (0.2 \times 0.2) \times 0.2 = 0.008;$

$\alpha_3(2) = \left(\sum_{y_2} \alpha_2(y_2) a_{y_2,2}\right) p(G|2) = \left(\alpha_2(2) \times a_{2,2}\right) \times 0.1 = (0.2 \times 0.8) \times 0.1 = 0.016;$

$\alpha_3(3) = \left(\sum_{y_2} \alpha_2(y_2) a_{y_2,3}\right) p(G|3) = \left(\alpha_2(1) \times a_{1,3}\right) \times 0.3 = (0.2 \times 0.8) \times 0.3 = 0.048;$

$\alpha_3(4) = \left(\sum_{y_2} \alpha_2(y_2) a_{y_2,4}\right) p(G|4) = \left(\alpha_2(2) \times a_{2,4}\right) \times 0.4 = (0.2 \times 0.2) \times 0.4 = 0.016;$

$\alpha_3(5) = 0$

当$t = 3$时，$\alpha_4(y_4) = \left(\sum_{y_3} \alpha_3(y_3) a_{y_3,y_4}\right) p(x_4|y_4)$，$x_4 = T$，

$\alpha_4(0) = 0$

$\alpha_4(1) = \left(\sum_{y_3} \alpha_3(y_3) a_{y_3,1}\right) p(T|1) = \left(\alpha_3(1) \times a_{1,1}\right) \times p(T|1) = (0.008 \times 0.2) \times 0.3 = 0.00048;$

$\alpha_4(2) = \left(\sum_{y_3} \alpha_3(y_3) a_{y_3,2}\right) p(T|2) = \left(\alpha_3(2) \times a_{2,2}\right) \times p(T|2) = (0.016 \times 0.8) \times 0.4 = 0.00512;$

$\alpha_4(3) = \left(\sum_{y_3} \alpha_3(y_3) a_{y_3,3}\right) p(T|3) = \left(\alpha_3(1) \times a_{1,3} + \alpha_3(3) \times a_{3,3}\right) \times p(T|3) = (0.008 \times 0.8 + 0.048 \times 0.4) \times 0.2 = 0.00512;$

$\alpha_4(4) = \left(\sum_{y_3} \alpha_3(y_3) a_{y_3,4}\right) p(T|4) = \left(\alpha_3(2) \times a_{2,4} + \alpha_3(4) \times a_{4,4}\right) \times p(T|4) = (0.016 \times 0.2 + 0.016 \times 0.1) \times 0.1 = 0.00048;$

$\alpha_4(5) = 0$

当$t = 4$时，$\alpha_5(y_5) = \left(\sum_{y_4} \alpha_4(y_4) a_{y_4,y_5}\right) p(x_5|y_5)$，$x_5 = T$，

$\alpha_5(0) = 0$

$\alpha_5(1) = \left(\sum_{y_4} \alpha_4(y_4) a_{y_4,1}\right) p(T|1) = \left(\alpha_4(1) \times a_{1,1}\right) \times p(T|1) = (0.00048 \times 0.2) \times 0.3 = 0.0000288;$

$\alpha_5(2) = \left(\sum_{y_4} \alpha_4(y_4) a_{y_4,2}\right) p(T|2) = \left(\alpha_4(2) \times a_{2,2}\right) \times p(T|2) = (0.00512 \times 0.8) \times 0.4 = 0.00164;$

$\alpha_5(3) = \left(\sum_{y_4} \alpha_4(y_4) a_{y_4,3}\right) p(T|3) = \left(\alpha_4(1) \times a_{1,3} + \alpha_4(3) \times a_{3,3}\right) \times p(T|3) = (0.00048 \times 0.8 + 0.00512 \times 0.4) \times 0.2 = 0.000486;$

$\alpha_5(4) = \left(\sum_{y_4} \alpha_4(y_4) a_{y_4,4}\right) p(T|4) = \left(\alpha_4(2) \times a_{2,4} + \alpha_4(4) \times a_{4,4}\right) \times p(T|4) = (0.00512 \times 0.2 + 0.00048 \times 0.1) \times 0.1 = 0.000107;$

$\alpha_5(5) = 0$

当$t = 5$时，$\alpha_6(y_6) = \left(\sum_{y_5} \alpha_5(y_5) a_{y_5,y_6}\right) p(6|y_6)$，$x_6 = end$，

$\alpha_6(0) = 0$

$\alpha_6(5) = \left(\sum_{y_5} \alpha_5(y_5) a_{y_5,5}\right) p(end|5) = \left(\alpha_5(3) a_{3,5} + \alpha_5(4) a_{4,5}\right) = 0.6 \times 0.000486 + 0.9 \times 0.000107 = 0.000388$。

总结如下：

状态	0	1	2	3	4	5
0	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
A 1	0.00000	0.20000	0.20000	0.00000	0.00000	0.00000
G 2	0.00000	0.00800	0.01600	0.04800	0.01600	0.00000
T 3	0.00000	0.00048	0.00512	0.00512	0.00048	0.00000
T 4	0.0000000	0.0000288	0.0016384	0.0004864	0.000107	0.000388

（2）

2. Viterbi values for AGTT

状态	0	1	2	3	4	5
0	1.00000	0.00000	0.00000	0.00000	0.00000	0.00000
T 1	0.00000	0.15000	0.20000	0.00000	0.00000	0.00000
A 2	0.00000	0.01200	0.06400	0.02400	0.00400	0.00000
T 3	0.00000	0.00072	0.02048	0.00192	0.00128	0.00000
A 4	0.00000	0.0000576	0.00655	0.000154	0.000410	0.0003686

φ	0	1	2	3	4	5
0						
T 1	0	0				
A 2	1	2	1	2		
T 3	1	2	1,3	2		
A 4	1	2	3	2	4	

反向回溯：

T A T A

2 2 2 4

18. 假设你有三个盒子，每个盒子里都有一定数量的苹果和桔子。每次随机选择一个盒子，然后从盒子里选一个水果，并记录你的发现（a代表苹果，o代表橘子）。

不幸的是，你忘了写下你所选的盒子，只是简单地记下了苹果和桔子。假设每个盒子中水果数量如下：

- 盒子一：2 个苹果，2 个桔子
- 盒子二 : 3 个苹果，1 个桔子
- 盒子三: 1 个苹果，3 个桔子。

（1） 请给出 HMM 模型；

（2） 请给出水果序列 $\boldsymbol{x} = (a, a, o, o, o)$对应的最佳盒子序列。

答：（1）初始概率$\boldsymbol{\pi} = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$,

盒子间的转移概率矩阵$\mathbf{A} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$,

发放概率矩阵（给定盒子是选择每种水果的概率）$\mathbf{B} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix}$。

（3） 采用 Viterbi 算法求解给定观测序列对应的隐含状态序列

1) 初始化：$\delta_1(i) = \pi_i b_{i,x_1}$， $x_1 = a$，所以$\boldsymbol{\delta_1} = \left(\frac{1}{3}\times\frac{1}{2}, \frac{1}{3}\times\frac{3}{4}, \frac{1}{3}\times\frac{1}{4}\right) = \left(\frac{1}{6}, \frac{1}{4}, \frac{1}{12}\right)$

$$\varphi_1(1) = \varphi_1(2) = \varphi_1(3) = 0,$$

2)

当$t = 1$时：

$\delta_2(i) = \max\limits_{\{y_1\}} \left[\delta_1(y_1) a_{y_1,i}\right] b_{i,x_2}$， $x_2 = a$，所以：

$$\delta_2(1) = \max_{\{y_1\}} \left[\delta_1(y_1) a_{y_1,1}\right] b_{1,1} = \max_{\{y_1\}} \left[\frac{1}{6}\times\frac{1}{3}, \frac{1}{4}\times\frac{1}{3}, \frac{1}{12}\times\frac{1}{3}\right] \times\frac{1}{2} = \frac{1}{24}$$

$$\delta_2(2) = \max_{\{y_1\}} \left[\delta_1(y_1) a_{y_1,2}\right] b_{2,1} = \max_{\{y_1\}} \left[\frac{1}{6}\times\frac{1}{3}, \frac{1}{4}\times\frac{1}{3}, \frac{1}{12}\times\frac{1}{3}\right] \times\frac{3}{4} = \frac{1}{16}$$

$$\delta_2(3) = \max_{\{y_1\}} \left[\delta_1(y_1) a_{y_1,3}\right] b_{3,1} = \max_{\{y_1\}} \left[\frac{1}{6}\times\frac{1}{3}, \frac{1}{4}\times\frac{1}{3}, \frac{1}{12}\times\frac{1}{3}\right] \times\frac{1}{4} = \frac{1}{48}$$

根据$\varphi_2(i) = \underset{\{y_1\}}{\operatorname{argmax}} \left[\delta_1(y_1) a_{y_1,i}\right]$，得到：

$$\varphi_2(1) = \underset{\{y_1\}}{\operatorname{argmax}} \left[\delta_1(y_1) a_{y_1,1}\right] = 2$$

$$\varphi_2(2) = \underset{\{y_1\}}{\operatorname{argmax}} \left[\delta_1(y_1) a_{y_1,2}\right] = 2$$

$$\varphi_2(3) = \underset{\{y_1\}}{\operatorname{argmax}} \left[\delta_1(y_1) a_{y_1,3}\right] = 2$$

当$t=2$时：

$\delta_3(i)=\max\limits_{\{y_2\}}\left[\delta_2(y_2)a_{y_2,i}\right]b_{i,x_3}$，$x_3=o$，所以：

$$\delta_3(1)=\max_{\{y_2\}}\left[\delta_2(y_2)a_{y_2,1}\right]b_{1,2}=\max_{\{y_2\}}\left[\frac{1}{24}\times\frac{1}{3},\frac{1}{16}\times\frac{1}{3},\frac{1}{48}\times\frac{1}{3}\right]\times\frac{1}{2}=\frac{1}{96}$$

$$\delta_3(2)=\max_{\{y_2\}}\left[\delta_2(y_2)a_{y_2,2}\right]b_{2,2}=\max_{\{y_2\}}\left[\frac{1}{24}\times\frac{1}{3},\frac{1}{16}\times\frac{1}{3},\frac{1}{48}\times\frac{1}{3}\right]\times\frac{1}{4}=\frac{1}{192}$$

$$\delta_3(3)=\max_{\{y_2\}}\left[\delta_2(y_2)a_{y_2,3}\right]b_{3,2}=\max_{\{y_2\}}\left[\frac{1}{24}\times\frac{1}{3},\frac{1}{16}\times\frac{1}{3},\frac{1}{48}\times\frac{1}{3}\right]\times\frac{3}{4}=\frac{1}{64}$$

根据$\varphi_3(i)=\operatorname{argmax}\limits_{\{y_2\}}\left[\delta_2(y_2)a_{y_2,i}\right]$，得到：

$$\varphi_3(1)=\operatorname*{argmax}_{\{y_2\}}\left[\delta_2(y_2)a_{y_2,1}\right]=2$$

$$\varphi_3(2)=\operatorname*{argmax}_{\{y_2\}}\left[\delta_2(y_2)a_{y_2,2}\right]=2$$

$$\varphi_3(3)=\operatorname*{argmax}_{\{y_2\}}\left[\delta_2(y_2)a_{y_2,3}\right]=2$$

当$t=3$时：

$\delta_4(i)=\max\limits_{\{y_3\}}\left[\delta_3(y_3)a_{y_3,i}\right]b_{i,x_4}$，$x_4=o$，所以：

$$\delta_4(1)=\max_{\{y_3\}}\left[\delta_3(y_3)a_{y_3,1}\right]b_{1,2}=\max_{\{y_3\}}\left[\frac{1}{96}\times\frac{1}{3},\frac{1}{192}\times\frac{1}{3},\frac{1}{64}\times\frac{1}{3}\right]\times\frac{1}{2}=\frac{1}{384}$$

$$\delta_4(2)=\max_{\{y_3\}}\left[\delta_3(y_3)a_{y_3,2}\right]b_{2,2}=\max_{\{y_3\}}\left[\frac{1}{96}\times\frac{1}{3},\frac{1}{192}\times\frac{1}{3},\frac{1}{64}\times\frac{1}{3}\right]\times\frac{1}{4}=\frac{1}{768}$$

$$\delta_4(3)=\max_{\{y_3\}}\left[\delta_3(y_3)a_{y_3,3}\right]b_{3,2}=\max_{\{y_3\}}\left[\left[\frac{1}{96}\times\frac{1}{3},\frac{1}{192}\times\frac{1}{3},\frac{1}{64}\times\frac{1}{3}\right]\right]\times\frac{3}{4}=\frac{1}{256}$$

根据$\varphi_4(i)=\operatorname{argmax}\limits_{\{y_3\}}\left[\delta_3(y_3)a_{y_3,i}\right]$，得到：

$$\varphi_4(1)=\operatorname*{argmax}_{\{y_3\}}\left[\delta_3(y_3)a_{y_3,1}\right]=3$$

$$\varphi_4(2)=\operatorname*{argmax}_{\{y_3\}}\left[\delta_3(y_3)a_{y_3,2}\right]=3$$

$$\varphi_4(3)=\operatorname*{argmax}_{\{y_3\}}\left[\delta_3(y_3)a_{y_3,3}\right]=3$$

当$t=4$时：

$\delta_5(i)=\max\limits_{\{y_4\}}\left[\delta_4(y_4)a_{y_4,i}\right]b_{i,x_5}$，$x_5=a$，所以：

$$\delta_5(1)=\max_{\{y_4\}}\left[\delta_4(y_4)a_{y_4,1}\right]b_{1,1}=\max_{\{y_4\}}\left[\frac{1}{384}\times\frac{1}{3},\frac{1}{768}\times\frac{1}{3},\frac{1}{256}\times\frac{1}{3}\right]\times\frac{1}{2}=\frac{1}{1536}$$

$$\delta_4(2)=\max_{\{y_4\}}\left[\delta_4(y_4)a_{y_4,2}\right]b_{2,1}=\max_{\{y_4\}}\left[\frac{1}{384}\times\frac{1}{3},\frac{1}{768}\times\frac{1}{3},\frac{1}{256}\times\frac{1}{3}\right]\times\frac{3}{4}=\frac{1}{1024}$$

$$\delta_4(3)=\max_{\{y_4\}}\left[\delta_4(y_4)a_{y_4,3}\right]b_{3,1}=\max_{\{y_4\}}\left[\frac{1}{384}\times\frac{1}{3},\frac{1}{768}\times\frac{1}{3},\frac{1}{256}\times\frac{1}{3}\right]\times\frac{1}{4}=\frac{1}{3072}$$

根据$\varphi_5(i)=\underset{\{y_4\}}{\operatorname{argmax}}\left[\delta_4(y_4)a_{y_4,i}\right]$，得到：

$$\varphi_5(1)=\underset{\{y_4\}}{\operatorname{argmax}}\left[\delta_4(y_4)a_{y_4,1}\right]=3$$

$$\varphi_5(2)=\underset{\{y_4\}}{\operatorname{argmax}}\left[\delta_4(y_4)a_{y_4,2}\right]=3$$

$$\varphi_5(3)=\underset{\{y_4\}}{\operatorname{argmax}}\left[\delta_4(y_4)a_{y_4,3}\right]=3$$

3) 终止：$T=5$

$i_5^*=\underset{\{i\}}{argmax}\delta_5(i)=2$；

最优路径回溯：$i_4^*=\varphi_5(i_5^*)=\varphi_5(2)=3$；

$i_3^*=\varphi_4(i_4^*)=\varphi_4(3)=3$；

$i_2^*=\varphi_3(i_3^*)=\varphi_3(3)=2$；

$i_1^*=\varphi_2(i_2^*)=\varphi_2(3)=2$；

所以水果序列$\boldsymbol{x}=(a,a,o,o,o)$对应的最佳盒子序列 2，2，3，3，2。

19. 请给出当损失函数取交叉熵损失时的梯度提升算法的流程（初始化和迭代过程）。

（提示：交叉熵损失为：$L\big(f(\boldsymbol{x}),y\big)=\ln\Big(1+\exp\big(-yf(\boldsymbol{x})\big)\Big)$，$y\in\{-1,1\}$）

20. 请给出 AdaBoost 算法的流程。（略）