期中1：(10分)给定二分类数据集$D=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_m,y_m)\}$，假设有分类算法SVM、多层感知机、决策树，请通过实验方法比较这三个算法的优劣性.

模型评估，数据集划分，评价指标选择(度量)

1.模型评估：给定二分类数据集$D=\{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_m,y_m)\}$，测试算法在训练集和测试集上的误差，经验误差和泛化误差

过拟合

学习器把训练样本学习的“太好”，将训练样本本身的特点当做所有样本的一般性质，导致泛化性能下降

优化目标加正则项

early stop

欠拟合

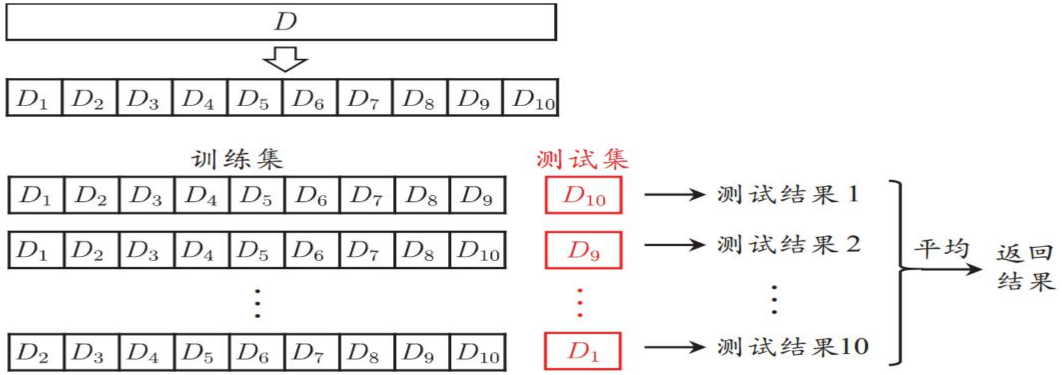
对训练样本的一般性质尚未学好

决策树:拓展分支

神经网络：增加训练轮数

2.数据集划分：

数据集划分的方法包括，留出法（划分时保证正负样本在训练集、测试集中的分布与数据集的一致）；K折交叉验证法（K最常用的取值为10）；留一法；自助法



10 折交叉验证示意图

3.性能评估:

数据集$D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$为二分类，主要测试对比三种算法在数据集上的错误率和分类精度，其他任务则采用不同的评测指标（查准率，查全率，P-R曲线等）。

错误率：分错样本占样本总数的比例

$$E(f, D) = \frac{1}{m} \sum_i \mathbb{I}(f(\boldsymbol{x}_i) \neq y_i)$$

精度：分对样本占样本总数的比率

$$acc(f, D) = \frac{1}{m} \sum_i \mathbb{I}(f(\boldsymbol{x}_i) = y_i)$$

期中2：(20分)对率回归模型用sigmoid函数实现二分类，若替换为softmax函数，可以实现多分类。请给出线性多分类的损失函数（6分）并计算参数梯度（6分），同时基于梯度下降法给出算法学习的伪代码.（8分）

$$h_\theta(x_i)=\begin{bmatrix}p(y_i=1|x_i;\theta)\\ p(y_i=2|x_i;\theta)\\ \vdots\\ p(y_i=k|x_i;\theta)\end{bmatrix}=\frac{1}{\sum_{j=1}^{k}e^{\theta_j^T x_i}}\begin{bmatrix}e^{\theta_1^T x_i}\\ e^{\theta_2^T x_i}\\ \vdots\\ e^{\theta_k^T x_i}\end{bmatrix} \tag{1}$$

损失函数：

$$L(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k}1\{y_i=j\}\log\frac{e^{\theta_j^T x_i}}{\sum_{l=1}^{k}e^{\theta_l^T x_i}}\right]$$

上式中1{·}表示为示性函数，当$y^{(i)}=j$时，函数值为1，否则为0，当类别为正确时，函数值为1.

$$\theta = \begin{bmatrix} \theta_1^T \\ \theta_2^T \\ \vdots \\ \theta_k^T \end{bmatrix}$$

参数梯度：

$$\begin{aligned}
\frac{\partial L(\theta)}{\partial \theta_j} &= -\frac{1}{m}\frac{\partial}{\partial \theta_j}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y_i = j\right\}\log\frac{e^{\theta_j^T x_i}}{\sum_{l=1}^{k} e^{\theta_l^T x_i}}\right] \\
&= -\frac{1}{m}\frac{\partial}{\partial \theta_j}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\left\{y_i = j\right\}\left(\theta_j^T x_i - \log\sum_{l=1}^{k} e^{\theta_l^T x_i}\right)\right] \\
&= -\frac{1}{m}\left[\sum_{i=1}^{m} 1\left\{y_i = j\right\}\left(x_i - \sum_{j=1}^{k}\frac{e^{\theta_j^T x_i}\cdot x_i}{\sum_{l=1}^{k} e^{\theta_l^T x_i}}\right)\right] \\
&= -\frac{1}{m}\left[\sum_{i=1}^{m} x_i 1\left\{y_i = j\right\}\left(1 - \sum_{j=1}^{k}\frac{e^{\theta_j^T x_i}}{\sum_{l=1}^{k} e^{\theta_l^T x_i}}\right)\right] \\
&= -\frac{1}{m}\left[\sum_{i=1}^{m} x_i\left(1\left\{y_i = j\right\} - \sum_{j=1}^{k} 1\left\{y_i = j\right\}\frac{e^{\theta_j^T x_i}}{\sum_{l=1}^{k} e^{\theta_l^T x_i}}\right)\right] \\
&= -\frac{1}{m}\left[\sum_{i=1}^{m} x_i\left(1\left\{y_i = j\right\} - \frac{e^{\theta_j^T x_i}}{\sum_{l=1}^{k} e^{\theta_l^T x_i}}\right)\right] \\
&= -\frac{1}{m}\left[\sum_{i=1}^{m} x_i\left(1\left\{y_i = j\right\} - p\left(y_i = j \mid x_i;\theta\right)\right)\right]
\end{aligned}$$

伪代码：

设输入数据$X = \{x_1, x_2, x_3, \cdots\cdots, x_m\}$，共m个数据样本，组成$m \times n$的矩阵，输出的类别为$Y = \{y_1, y_2, y_3, \cdots\cdots, y_m\}$，其中$y_i$是一个$1 \times k$的one-hot矩阵，$P = \{p_1, p_2, p_3, \cdots\cdots, p_m\}$，对应于一个$m \times k$的矩阵，λ表示正则化参数。

$$\frac{\partial L(\theta)}{\partial \theta} = -\frac{1}{m}(y - P)^T X + \lambda\theta$$

1. 设置训练的周期T，学习率alpha等超参数
2. 初始化权重参数θ，对样本中数据进行one-hot编码
3. 循环T个周期：
   计算$m \times k$的分数矩阵Scores=X.dot(θ)
   计算$m \times 1$的矩阵softmax(Scores)
   计算loss函数
   根据上式求解梯度dw
   θ=θ-alpha*dw
4. 输出参数θ

期中3：下表表示的二分类数据集，具有三个属性A,B,C，样本标记为两类"+"，"-"。请运用你学过的知识完成如下问题：

实例	A	B	C	类别
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-
10	F	F	2.0	+

(a) 整个训练样本关于类属性的熵是多少（3分）
(b) 数据集中A，B两个属性的信息增益各是多少（3分）
(c) 对于属性C，计算所有可能划分的信息增益（4分）
(d) 根据Gini指数，A和B两个属性哪个是最优划分（4分）
(e) 采用算法C4.5，构造决策树（6分）

1. $\text{Entropy} =- 2 * \frac{5}{10} * \log_2 \frac{5}{10} = 1$
2. $\text{Gain(A)} = 1 - \left(\frac{4}{10} * \left(-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right) + \frac{6}{10} * \left(-\frac{2}{6}\log_2\frac{2}{6} - \frac{4}{6}\log_2\frac{4}{6}\right)\right)$ =0.125

$$\text{Gain(B)} = 1 - \left(\frac{5}{10} * \left(-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right) + \frac{5}{10} * \left(-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right)\right) = 0.029$$

3. 排列C属性值

1.0	2.0	3.0	4.0	5.0	5.0	6.0	7.0	7.0	8.0
+	+	-	+	-	-	+	-	+	-

所有可能划分及信息增益

	0.5		1.5		2.5		3.5		4.5		5.5		6.5		7.5		8.5	
	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>	<=	>
+	0	5	1	4	2	3	2	3	3	2	3	2	4	1	5	0	5	0
-	0	5	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0
Gain	0		0.108		0.236		0.035		0.125		0		0.035		0.108		0	

4. $\mathrm{Gini}(A)=\frac{4}{10}*\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)+\frac{6}{10}*\left(1-\left(\frac{2}{6}\right)^2-\left(\frac{4}{6}\right)^2\right)=0.417$

$$\mathrm{Gini}(B)=\frac{5}{10}*\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right)+\frac{5}{10}*\left(1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2\right)=0.48$$

由于 $\mathrm{Gini}(A)<\mathrm{Gini}(B)$，A比B更可取

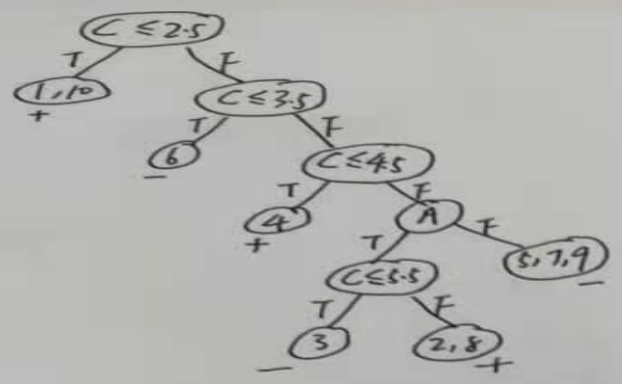
5. 各属性的增益率

$$\text{Gain rate}(A) = \frac{0.125}{-\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10}} = 0.128$$

$$\text{Gain rate}(B) = \frac{0.029}{-\frac{5}{10}\log_2\frac{5}{10} - \frac{5}{10}\log_2\frac{5}{10}} = 0.029$$

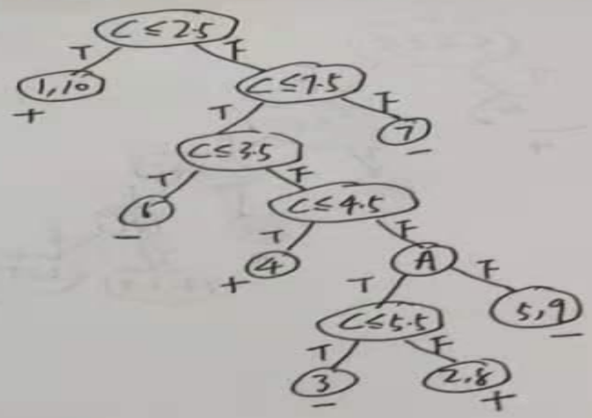
0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.3
0	0.230	0.328	0.040	0.128	0	0.040	0.230	0

根据 $C \leq 2.5$将数据集划分为{1,10}和{2,3,4,5,6,7,8,9}
同理，划分节点{2,3,4,5,6,7,8,9}直到节点中仅包含一种类别，得到如下图决策树：



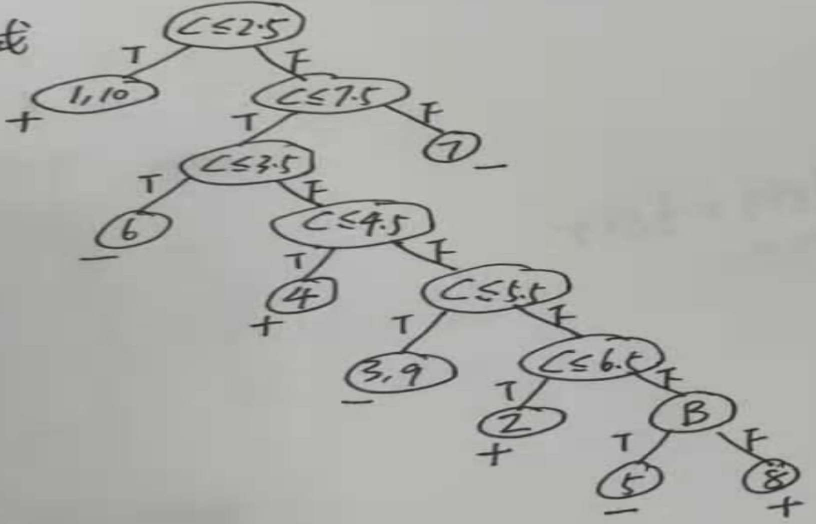
C≤2.5
1,10
C≤3.5
6
C≤4.5
4
A
C≤5.5
3
2,8
5,7,9

或



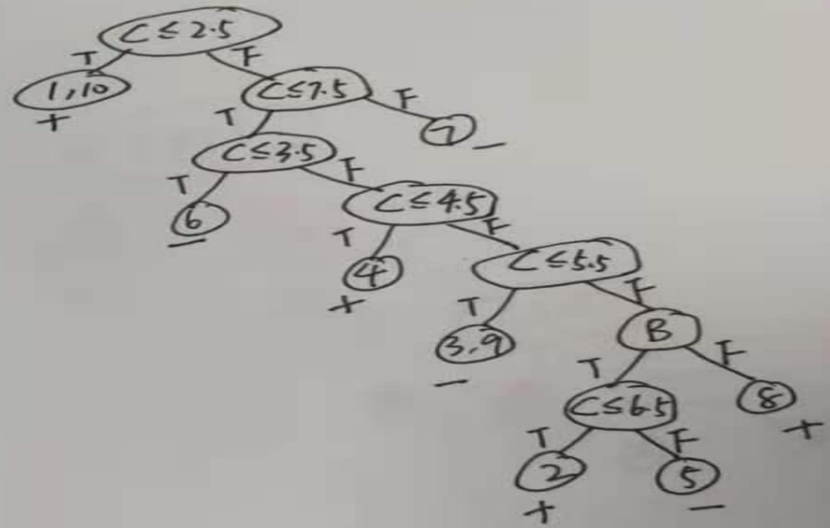
C≤2.5
1,10
C≤7.5
7
C≤3.5
6
C≤4.5
4
A
C≤5.5
5,9
3
2,8

或



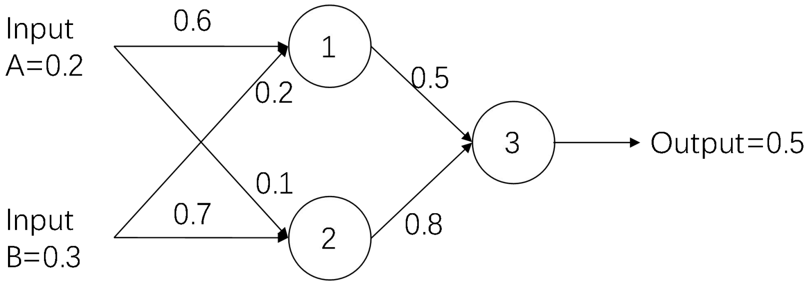
C≤2.5
1,10
C≤7.5
7
C≤3.5
6
C≤4.5
4
C≤5.5
3,9
C≤6.5
2
B
5
8

或

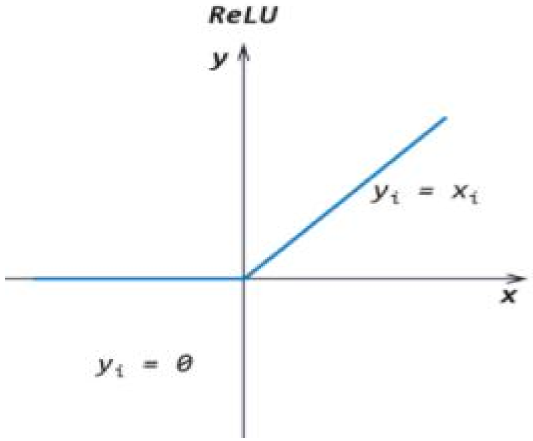


C≤2.5
1,10
C≤7.5
7
C≤3.5
6
C≤4.5
4
C≤5.5
3,9
B
C≤6.5
8
2
5

期中4:考虑如下简单网络，假设激活函数为ReLU，用平方损失$\frac{1}{2}(y-\hat{y})^2$计算误差，请用BP算法更新一次所有参数（学习率为1），给出更新后的参数值（12分，给出详细计算过程），并计算给定输入值x=(0.2,0.3)时初始时和更新后的输出值（5分），检查参数更新是否降低了平方损失值.（3分）



Relu：



$$\frac{\partial y}{\partial x} = \begin{cases} 0, x < 0 \\ 1, \ x > 0 \end{cases}$$

1.Input：

$0.2 \times 0.6 + 0.3 \times 0.2$=0.18

Output:0.18

Error:e1=g*w1=0.226*0.5=0.113

2.Input：

$0.3 \times 0.7 + 0.2 \times 0.1$=0.23

Output:0.23

Error:e2=g*w2=0.1808

3.Input：

$0.5 \times 0.18 + 0.8 \times 0.23$=0.274

Output:0.274

Error: g=0.226

Loss1=$\frac{1}{2}(0.5-0.274)^2 = 0.0255$

参数更新：

W1=w1+g*0.18=0.541

W2=w2+g*0.23=0.852

W3=w3+e1*0.2=0.623

W4=w4+e2*0.2=0.136

W5=w5+e1*0.3=0.234

W6=w6+e2*0.3=0.754

更新后：

Node1=0.2*0.623+0.3*0.234=0.195

Node2=0.2*0.136+0.3*0.754=0.253

Node3=0.195*0.541+0.253*0.852=0.321

Loss2=0.5*$(0.5-0.321)^2$=0.0160

损失降低了

期中5:

（a）SVM可直接求解优化问题$\min_{\boldsymbol{w},b}\frac{1}{2}\|\boldsymbol{w}\|^2+C\sum_{i=1}^{m}\max\left(0,1-y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)\right)$，请计算该目标函数关于参数的梯度，并基于梯度下降法给出算法伪代码.（12分）

(b) 支持向量回归的对偶问题如下，

$$\max_{\boldsymbol{\alpha},\widehat{\boldsymbol{\alpha}}} g(\boldsymbol{\alpha},\widehat{\boldsymbol{\alpha}})=-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\widehat{\alpha}_i)(\alpha_j-\widehat{\alpha}_j)\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_{i=1}^{m}\left(y_i(\widehat{\alpha}_i-\alpha_i)-\epsilon(\widehat{\alpha}_i+\alpha_i)\right)$$

�.�. $C\geqslant\boldsymbol{\alpha},\ \widehat{\boldsymbol{\alpha}}\ \geqslant 0$ and $\sum_{i=1}^{m}(\alpha_i-\widehat{\alpha}_i)=0$

请将该问题转化为类似于如下标准型的形式（$\boldsymbol{u},\boldsymbol{v},\boldsymbol{K}$均已知），

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha})=\boldsymbol{\alpha}^\top\boldsymbol{v}-\frac{1}{2}\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha}$$

�.�. $C\geqslant\boldsymbol{\alpha}\geqslant 0$ and $\boldsymbol{\alpha}^\top\boldsymbol{u}=0$

例如在软间隔SVM中$\boldsymbol{v}=\boldsymbol{1},\ \boldsymbol{u}=\boldsymbol{y},\ \boldsymbol{K}[i,j]=y_iy_j\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)$.（8分）

- $\frac{\partial J(w,b)}{\partial w}=w+C\sum_{i=1}^{m}I\{y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)\le 1\}(-y_i\phi(\boldsymbol{x}_i))=\mathrm{w}-\mathrm{C}\sum_{i:y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)\le 1}y_i\phi(\boldsymbol{x}_i)$
- $\frac{\partial J(w,b)}{\partial b}=C\sum_{i=1}^{m}I\{y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)\le 1\}(-y_i)=-\mathrm{C}\sum_{i:y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i)+b)\le 1}y_i$

算法伪代码：大致步骤 初始化数据，参数，梯度更新，学习率，终止条件

期中5:

$$\max_{\boldsymbol{\alpha},\widehat{\boldsymbol{\alpha}}} g(\boldsymbol{\alpha},\widehat{\boldsymbol{\alpha}}) =-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\widehat{\alpha}_i)(\alpha_j-\widehat{\alpha}_j)\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)+\sum_{i=1}^{m}\left(y_i(\widehat{\alpha}_i-\alpha_i)-\epsilon(\widehat{\alpha}_i+\alpha_i)\right)$$

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{v}-\frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m}(\alpha_i-\widehat{\alpha}_i)(\alpha_j-\widehat{\alpha}_j)\kappa(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jk_{ij}-\widehat{\alpha}_i\alpha_jk_{ij}-\alpha_i\widehat{\alpha}_jk_{ij}+\widehat{\alpha}_i\widehat{\alpha}_jk_{ij}$$

令$\alpha^* = \binom{\alpha}{\widehat{\alpha}}$，则有$\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jk_{ij}-\widehat{\alpha}_i\alpha_jk_{ij}-\alpha_i\widehat{\alpha}_jk_{ij}+\widehat{\alpha}_i\widehat{\alpha}_jk_{ij}=\alpha^{*\mathrm{T}}K\alpha^*$,其中，K=$\begin{bmatrix} k & -k \\ -k & k \end{bmatrix}$

令 $v=\binom{-y-\varepsilon}{y-\varepsilon}$,则有$\sum_{i=1}^{m}\left(y_i(\widehat{\alpha}_i-\alpha_i)-\epsilon(\widehat{\alpha}_i+\alpha_i)\right)=\alpha^{*\mathrm{T}}v$

因此，原式形变为$\max_{\alpha^*} g(\alpha^*) = \alpha^{*\mathrm{T}}\boldsymbol{v}-\frac{1}{2}\alpha^{*\mathrm{T}}\boldsymbol{K}\alpha^*$ s.t. $C\geq\alpha^*\geq 0,\alpha^{*\mathrm{T}}v=0$

期中6:

6.(10分)假设数据集$D=\{x_1,x_2,\dots,x_m\}$, 任意x_i是从均值为μ、方差为$\frac{1}{\lambda}$的正态分布$\mathcal{N}(\mu,\lambda^{-1})$中独立采样而得到。试通过极大似然估计法求解μ和$\beta$.（4分）

假设μ和λ也是随机变量，在未知数据集D时分别满足正态分布和伽玛分布，即μ~$\mathcal{N}(\mu_0,\sigma_0^2)$，而λ~$\mathrm{Gam}(a,b)=\frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)$，其中$\Gamma(a)$为伽玛函数，请用贝叶斯定理求解μ和λ的后验分布$p(\mu|D)$和$p(\lambda|D)$.（6分）

解：1、$f(x,\mu,\lambda)=\left(\frac{\sqrt{\lambda}}{\sqrt{2\pi}}\right)^m exp\left\{-\frac{\lambda}{2}\sum_{i=1}^m(x_i-\mu)^2\right\}$,

则$l(x,\mu,\lambda)=\log f(x,\mu,\lambda)=-\frac{m}{2}\log 2\pi+\frac{m}{2}\log\lambda-\frac{\lambda}{2}\sum_{i=1}^m(x_i-\mu)^2$

$$\frac{\partial l}{\partial\mu}=\lambda\sum_{i=1}^m(x_i-\mu)=0\qquad\Rightarrow\quad\mu=\frac{1}{m}\sum_{i=1}^m x_i$$

$$\frac{\partial l}{\partial\lambda}=\frac{m}{2\lambda}-\frac{1}{2}\sum_{i=1}^m(x_i-\mu)^2\Rightarrow\lambda=\frac{m}{\sum_{i=1}^m(x_i-\bar{x})^2}$$

期中6:

6.(10分)假设数据集$D=\{x_1,x_2,\ldots,x_m\}$, 任意x_i是从均值为μ、方差为$\frac{1}{\lambda}$的正态分布$\mathcal{N}(\mu,\lambda^{-1})$中独立采样而得到。试通过极大似然估计法求解μ和$\beta$.（4分）

假设μ和λ也是随机变量，在未知数据集D时分别满足正态分布和伽玛分布，即μ~$\mathcal{N}(\mu_0,\sigma_0^2)$，而λ~$\text{Gam}(a,b)=\frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)$，其中$\Gamma(a)$为伽玛函数，请用贝叶斯定理求解μ和λ的后验分布$p(\mu|D)$和$p(\lambda|D)$.（6分）

解：2、$P(D,\lambda,\mu)=P(D|\mu,\lambda)p(\lambda)p(\mu)=p(\lambda)p(\mu)\prod_{i=1}^m p(x_i|\mu,\lambda)$

$$=\frac{1}{\sqrt{2\pi\sigma_0^2}}\exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)\frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)\left(\frac{\lambda}{2\pi}\right)^{\frac{m}{2}}\exp\left(-\frac{\lambda}{2}\sum_i^m(x_i-\mu)^2\right)$$

$$\propto\exp\left(-\frac{\lambda}{2}\sum_i^m(x_i-\mu)^2-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2-b\lambda\right)\lambda^{\frac{m}{2}+a-1}$$

$$=\exp\left(-\frac{1}{2}\left(\left(\lambda m+\frac{1}{\sigma_0^2}\right)\mu^2-2\left(\lambda\sum_i^m x_i+\frac{\mu_0}{\sigma_0^2}\right)\mu\right)-\frac{\lambda}{2}\sum_i^m x_i^2-\frac{1}{2\sigma_0^2}\mu_0^2-b\,\lambda\right)\lambda^{\frac{m}{2}+a-1}$$

$$=\exp\left(-\frac{\lambda_m}{2}(\mu-\mu_m)^2+\frac{1}{2}\frac{\left(\lambda\sum_i^m x_i+\frac{\mu_0}{\sigma_0^2}\right)^2}{\lambda m+\frac{1}{\sigma_0^2}}-\frac{\lambda}{2}\sum_i^m x_i^2-\frac{1}{2\sigma_0^2}\mu_0^2-b\lambda\right)\lambda^{\frac{m}{2}+a-1}$$

记$\lambda_m=\lambda m+\frac{1}{\sigma_0^2}$

$$\mu_m=\frac{\lambda\sum_i^m x_i+\frac{\mu_0}{\sigma_0^2}}{\lambda m+\frac{1}{\sigma_0^2}}$$

期中6:

6.(10分)假设数据集$D=\{x_1,x_2,\dots,x_m\}$, 任意x_i是从均值为μ、方差为$\frac{1}{\lambda}$的正态分布$\mathcal{N}(\mu,\lambda^{-1})$中独立采样而得到。试通过极大似然估计法求解μ和$\beta$.（4分）

假设μ和λ也是随机变量，在未知数据集D时分别满足正态分布和伽玛分布，即μ~$\mathcal{N}(\mu_0,\sigma_0^2)$，而λ~$\mathrm{Gam}(a,b)=\frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)$，其中$\Gamma(a)$为伽玛函数，请用贝叶斯定理求解μ和λ的后验分布$p(\mu|D)$和$p(\lambda|D)$.（6分）

解：2、$P(D,\lambda,\mu)=P(D|\mu,\lambda)p(\lambda)p(\mu)=p(\lambda)p(\mu)\prod_{i=1}^m p(x_i|\mu,\lambda)$

记$\lambda_m=\lambda m+\frac{1}{\sigma_0^2}$

$$\mu_m=\frac{\lambda\sum_i^m x_i+\frac{\mu_0}{\sigma_0^2}}{\lambda m+\frac{1}{\sigma_0^2}}$$

$$\propto\exp\left(-\frac{\lambda_m}{2}(\mu-\mu_m)^2+\frac{1}{2}\frac{\left(\lambda\sum_{i=1}^m x_i+\frac{\mu_0}{\sigma_0^2}\right)^2}{\lambda m+\frac{1}{\sigma_0^2}}-\frac{\lambda}{2}\sum_i^m x_i^2-\frac{1}{2\sigma_0^2}\mu_0^2-b\lambda\right)\lambda^{\frac{m}{2}+a-1}$$

$$\propto\sqrt{\frac{\lambda_m}{2\pi}}\exp\left(-\frac{\lambda_m}{2}(\mu-\mu_m)^2+\frac{1}{2}\frac{\left(\lambda\sum_{i=1}^m x_i+\frac{\mu_0}{\sigma_0^2}\right)^2}{\lambda m+\frac{1}{\sigma_0^2}}-\frac{\lambda}{2}\sum_i^m x_i^2-\frac{1}{2\sigma_0^2}\mu_0^2-b\lambda\right)\lambda^{\frac{m}{2}+a-1}\frac{1}{\sqrt{\lambda_m}}$$

$$=\mathcal{N}(\mu|\mu_m,\lambda_m^{-1})\exp\left(\frac{1}{2}\frac{\left(\lambda\sum_{i=1}^m x_i+\frac{\mu_0}{\sigma_0^2}\right)^2}{\lambda m+\frac{1}{\sigma_0^2}}-\frac{\lambda}{2}\sum_i^m x_i^2-b\lambda\right)\lambda^{\frac{m}{2}+a-1}\frac{1}{\sqrt{\lambda_m}}$$

期中6：

6.(10分)假设数据集$D=\{x_1,x_2,\dots,x_m\}$，任意x_i是从均值为μ、方差为$\frac{1}{\lambda}$的正态分布$\mathcal{N}(\mu,\lambda^{-1})$中独立采样而得到。试通过极大似然估计法求解μ和$\beta$.（4分）

假设μ和λ也是随机变量，在未知数据集D时分别满足正态分布和伽玛分布，即$\mu\sim\mathcal{N}(\mu_0,\sigma_0^2)$，而$\lambda\sim\mathrm{Gam}(a,b)=\frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)$，其中$\Gamma(a)$为伽玛函数，请用贝叶斯定理求解μ和λ的后验分布$p(\mu|D)$和$p(\lambda|D)$.（6分）

解：2、$P(D,\lambda,\mu)=P(D|\mu,\lambda)p(\lambda)p(\mu)=p(\lambda)p(\mu)\prod_{i=1}^m p(x_i|\mu,\lambda)$

需要限定$\sigma_0^2=\frac{\hat{\sigma}_0^2}{\lambda}$，$\lambda$才能满足Gamma分布

记$\lambda_m=\lambda\left(m+\frac{1}{\hat{\sigma}_0^2}\right)$

$$\mu_m=\frac{\sum_{i=1}^m x_i+\frac{\mu_0}{\hat{\sigma}_0^2}}{m+\frac{1}{\hat{\sigma}_0^2}}$$

$$\propto\sqrt{\frac{\lambda}{2\pi\hat{\sigma}_0^2}}p(\mu|\mu_m,\lambda_m^{-1})\exp\left(\frac{1}{2}\frac{\lambda\left(\sum_{i=1}^m x_i+\frac{\mu_0}{\hat{\sigma}_0^2}\right)^2}{m+\frac{1}{\hat{\sigma}_0^2}}-\frac{\lambda}{2}\sum_i^m x_i^2-b\lambda\right)\lambda^{\frac{m}{2}+a-1}\frac{1}{\sqrt{\lambda\left(m+\frac{1}{\hat{\sigma}_0^2}\right)}}$$

$$\propto\mathcal{N}(\mu|\mu_m,\lambda_m^{-1})\exp\left(\frac{1}{2}\frac{\lambda\left(\sum_{i=1}^m x_i+\frac{\mu_0}{\hat{\sigma}_0^2}\right)^2}{m+\frac{1}{\hat{\sigma}_0^2}}-\frac{\lambda}{2}\sum_i^m x_i^2-b\lambda\right)\lambda^{\frac{m}{2}+a-1}$$

$$\propto\mathcal{N}(\mu|\mu_m,\lambda_m^{-1})\exp\left(-\left(b+\frac{\lambda}{2}\sum_i^m x_i^2-\frac{1}{2}\frac{\lambda\left(\sum_{i=1}^m x_i+\frac{\mu_0}{\hat{\sigma}_0^2}\right)^2}{m+\frac{1}{\hat{\sigma}_0^2}}\right)\lambda\right)\lambda^{\frac{m}{2}+a-1}$$

$$\mathcal{N}(\mu|\mu_m,\lambda_m^{-1})Gam(a_m,b_m)$$