HW 机器学习概论

PB19151769　　马宇骁

目录

1 HW1

作业

- 1. 计算$\frac{\partial \ln \det(\boldsymbol{A})}{\partial x}$
- 2. 书习题1.2
- 3. 已知随机变量 $\boldsymbol{x}=[\boldsymbol{x}_1,\boldsymbol{x}_2]\sim\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, 计算$P(\boldsymbol{x}_1), P(\boldsymbol{x}_1|\boldsymbol{x}_2)$
- 4. 证明范数 $\|\boldsymbol{x}\|_p$是凸函数
- 5. 证明判定凸函数的0阶和1阶条件相互等价

$$\forall x,y\in dom(f),\forall t\in[0,1], f(tx+(1-t)y)\le tf(x)+(1-t)f(y)$$

$$\updownarrow$$

$$\forall\, x,y\in dom(f)\ ,\ f(y)\ge f(x)+\nabla f(x)^{\mathsf{T}}(y-x)$$

1.1

先由：

$$\frac{\partial \det(A)}{\partial A_{ij}}=\frac{\partial \sum_k A_{ik}\,\mathrm{adj}^T(A)_{ik}}{\partial A_{ij}}=\sum_k \frac{\partial A_{ik}\,\mathrm{adj}^T(A)_{ik}}{\partial A_{ij}}=\sum_k \frac{\partial A_{ik}}{\partial A_{ij}}\mathrm{adj}^T(A)_{ik}+\sum_k A_{ik}\frac{\partial\,\mathrm{adj}^T(A)_{ik}}{\partial A_{ij}}$$

其中，$\frac{\partial\,\mathrm{adj}^T(A)_{ik}}{\partial A_{ij}}=0$，有：

$$\frac{\partial \det(A)}{\partial A_{ij}}=\sum_k \frac{\partial A_{ik}}{\partial A_{ij}}\mathrm{adj}^T(A)_{ik}=\mathrm{adj}^T(A)_{ij}$$

因此，

$$\begin{aligned}\frac{\partial \ln\det(A)}{\partial x}&=\frac{1}{\det(A)}\sum_{i,j}\frac{\partial \det(A)}{\partial A_{ij}}\frac{\partial A_{ij}}{\partial x}\\&=\frac{1}{\det(A)}\sum_{i,j}\mathrm{adj}^T(A)_{ij}\frac{\partial A_{ij}}{\partial x}\\&=\frac{1}{\det(A)}\sum_{i,j}\mathrm{adj}(A)_{ji}\frac{\partial A_{ij}}{\partial x}\\&=\frac{1}{\det(A)}\sum_{j=1}^{n}\left(A^*\frac{\partial A}{\partial x}\right)_{jj}\\&=\mathrm{tr}\left(A^{-1}\frac{\partial A}{\partial x}\right)\end{aligned}$$

1.2

题中的表 1.1 如图1.

编号	色泽	根蒂	敲声	好瓜
1	青绿	蜷缩	浊响	是
2	乌黑	蜷缩	沉闷	是
3	青绿	硬挺	清脆	否
4	乌黑	稍蜷	沉闷	否

图 1: 西瓜数据集

由题可知：

仅从表中发现有 3 个因素，每个因素有 2, 3, 3 种取值，假设空间总共有 $3 \times 4 \times 4 + 1 = 49$ 种假设，其中有 1 种是空集，不考虑空时：

全部不泛化 $2 \times 3 \times 3 = 18$ 种假设

一个属性泛化：$2 \times 3 + 3 \times 3 + 2 \times 3 = 21$ 种假设

两个属性泛化：$2 + 3 + 3 = 8$ 种假设

三属性泛化：1 种假设

用这 48 种假设的排列组合来组成析合范式，展开序列为:

$$1, C_{48}^{1}, C_{48}^{2}, \cdots, C_{48}^{48}$$

共 49 个数，左边的 1 代表‘空’，一个都不选，右边的 1 代表全部选。

如果 k=48，$2^{48} - 1$ 种 (排除一种都不选的情况).

如果 0<k<48，$\sum_{i=0}^{k} C_{48}^{i} - 1 = \sum_{i=1}^{k} C_{48}^{i}$ 种.

(以上结果均需去重)

若 k 足够大，并考虑重复，则演变为在不泛化的 18 种中选择，共有 2^{18}–1 种假设。

1.3

令

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \mathrm{Cov}(x_1, x_2) \\ \mathrm{Cov}(x_2, x_1) & \sigma_2^2 \end{pmatrix}$$

其中，$\rho = \frac{\mathrm{Cov}(x_1, x_2)}{\sigma_1 \sigma_2}$

由分布可知：

$$f(x_1, x_2) = \frac{1}{2\pi (1-\rho^2)^{\frac{1}{2}} \sigma_1 \sigma_2} e^{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right) + \left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}$$

所以，

$$\begin{aligned}
P(x_1) &= \int_{-\infty}^{x_1}\int_{-\infty}^{+\infty} f\left(x_1, x_2\right)\,\mathrm{d}\,x_2\,\mathrm{d}\,x_1 \\
&= \int_{-\infty}^{x_1}\int_{-\infty}^{+\infty} \frac{1}{2\pi\left(1-\rho^2\right)^{\frac{1}{2}}\sigma_1\sigma_2} e^{-\frac{1}{2\left(1-\rho^2\right)}\left[\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2-2\rho\left(\frac{x_1-\mu_1}{\sigma_1}\right)\left(\frac{x_2-\mu_2}{\sigma_2}\right)+\left(\frac{x_2-\mu_2}{\sigma_2}\right)^2\right]}\,\mathrm{d}\,x_2\,\mathrm{d}\,x_1 \\
&= \int_{-\infty}^{x_1} f_{x_1}(x_1)\,\mathrm{d}\,x_1 \\
&= \int_{-\infty}^{x_1} \frac{1}{\sigma_1\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2}\,\mathrm{d}\,x_1
\end{aligned}$$

$$\begin{aligned}
P(x_1|x_2) &= \int_{-\infty}^{x_1} \frac{f(x_1,x_2)}{f_{x_2}(x_2)}\,\mathrm{d}\,x_1 \\
&= \int_{-\infty}^{x_1} \frac{1}{\sqrt{2\pi}\sigma_1\sqrt{1-\rho^2}} e^{-\frac{(w-\rho v)^2}{2\left(1-\rho^2\right)}}\,\mathrm{d}\,x_1
\end{aligned}$$

其中，

$$w=\frac{(x_1-\mu_1)^2}{\sigma_1^2},\ v=\frac{(x_2-\mu_2)^2}{\sigma_2^2}.$$

1.4

p-范数（$p \geq 1$）：

$$f(X) = \|X\|_p = (\sum_{i=1}^{n} X_i^p)^{\frac{1}{p}}$$

要证明 f 为凸函数，

及证明 $f(\lambda\mathbf{x}+(1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x})+(1-\lambda)f(\mathbf{y})$ 对任意的 $0 \leq \lambda \leq 1$ 成立

由 Minkowski 不等式可知：

设 S 是一个度量空间，$1 \leq p \leq \infty$ ，$f, g \in L^p(S)$ ，那么 $f+g \in L^p(S)$ ，我们有:

$$\|f+g\|_p \leq \|f\|_p + \|g\|_p$$

即，

$$\left(\sum_{i=1}^{n}\left(x_i+y_i\right)^p\right)^{\frac{1}{p}} \leq \left(\sum_{i=1}^{n} x_i^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n} y_i^p\right)^{\frac{1}{p}}$$

如果 $1<p<\infty$ ，等号成立当且仅当 $k \geq 0$ ，$f=kg$ 或 $g=kf$

故，

$$\begin{aligned}
\|\lambda\mathbf{x}+(1-\lambda)\mathbf{y}\|_{\mathrm{p}} &\leq \|\lambda\mathbf{x}\| + \|(1-\lambda)\mathbf{y}\|_{\mathrm{p}} \\
&= \lambda\|\mathbf{x}\|_{\mathrm{p}} + (1-\lambda)\|\mathbf{y}\|_{\mathrm{p}}
\end{aligned}$$

凸函数得证.

1.5

- 充分性

令 $a = 1 - t$, $a \in [0, 1]$,

$$
\begin{aligned}
& f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) \\
\Longrightarrow\ & af(y) \geq f[(1-a)x + ay] - (1-a)f(x) = f[x + a(y-x)] + af(x) - f(x) \\
\Longrightarrow\ & f(y) \geq f(x) + \lim_{a \to 0} \frac{f[x + a(y-x)] - f(x)}{a(y-x)}(y-x) \\
\Longrightarrow\ & f(y) \geq f(x) + \nabla f(x)^T (y-x)
\end{aligned}
$$

- 必要性

令 $z = tx + (1-t)y$

$$f(y) \geq f(z) + \nabla f(z)^T (y - z) \qquad (1)$$

$$f(x) \geq f(z) + \nabla f(z)^T (x - z) \qquad (2)$$

$(1-t) * (1) + t * (2)$ 可得：

$$
\begin{aligned}
& \mathrm{t}f(x) + (1-t)f(y) \geq f(z) + \nabla f(z)^T (t(1-t)(x-y) + t(1-t)(y-x)) = f(z) \\
\Longrightarrow\ & f(tx + (1-t)y) \leq tf(x) + (1-t)f(y)
\end{aligned}
$$

2 HW2

作业

48

- 习题2.2
- 习题2.4
- 习题2.5
- 习题2.9

2.1

- 10 折交叉验证

分成 10 份每次取 9 份，其中正反各一半即 45 个样本，训练结果随机预测则错误率的期望是 0.5.

- 留一法

留出样本为正则训练 50 反 49 正预测反，反之预测正，结果一定错，错误率期望 1.

2.2

- 真正例率（TPR）

$$TPR = \frac{\text{预测为正例且真实为正例的数量}}{\text{真实为正例的数量}} = \frac{TP}{TP+FN}$$

- 假正例率（FPR）

$$FPR = \frac{\text{预测为正例且真实为反例的数量}}{\text{真实为反例的数量}} = \frac{FP}{TN+FP}$$

- 查准率（P）

$$P = \frac{\text{预测为正例且真实为正例的数量}}{\text{预测为正例的数量}} = \frac{TP}{TP+FP}$$

- 查全率（R）

$$R = \frac{\text{预测为正例且真实为正例的数量}}{\text{真实为正例的数量}} = \frac{TP}{TP+FN}$$

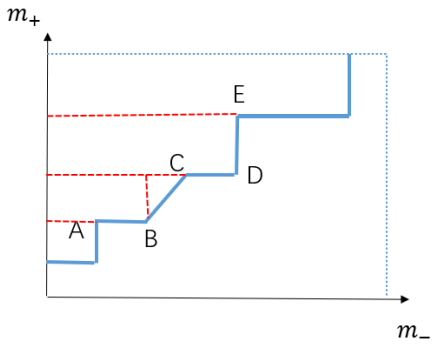
2.3

证明

$$\text{AUC} = 1 - \ell_{\text{rank}}$$

$$\ell_{\text{rank}} = \frac{1}{m^+m^-} \sum_{x^+ \in D^+} \sum_{x^- \in D^-} \left(\mathbb{I}\left(f\left(\boldsymbol{x}^+\right) < f\left(\boldsymbol{x}^-\right)\right) + \frac{1}{2}\mathbb{I}\left(f\left(\boldsymbol{x}^+\right) = f\left(\boldsymbol{x}^-\right)\right) \right)$$

可以将 ROC 曲线分为三部分：平行于 y 轴的；平行于 x 轴的；斜线段。



令 S_i 为 ROC 与 y 轴围成的面积块。则证明题目等式即证明 ℓ_{rank} 是 ROC 曲线与轴围成的全部面积。

- 平行于 y 轴的

此时，$\sum_{x^+\in D^+}\sum_{x^-\in D^-}\left(\frac{1}{2}\mathbb{I}(f(\boldsymbol{x}^+)=f(\boldsymbol{x}^-))\right)=0$ ，
且 $\sum_{x^+\in D^+}\sum_{x^-\in D^-}\frac{1}{m^+m^-}\mathbb{I}(f(\boldsymbol{x}^+)<f(\boldsymbol{x}^-))=\sum_{i,\text{平行 y 轴}}x_in_i=\sum_{i,\text{平行 y 轴}}S_i$ ，
故，不妨令 ℓ_{rank} 前半部分求和是 A，后半部分求和是 B（方便打字），则：

$\sum_{i,\text{平行 y 轴}}S_i=A+B$ 得证；

- 平行于 x 轴的

此时，$A+B=0$ ，$S=0$ 显然，因此，

$\sum_{i,\text{平行 x 轴}}S_i=A+B$ 得证；

- 斜线段

此时，面积可以看成开始的点做 y 轴平行的线向上至结束的点的高度的与 y 轴围成的矩形加上斜线段与刚刚做的辅助线围成的三角形的面积之和。

其中，A 为前者矩形的面积，B 是三角形的面积，故

$\sum_{i,\text{斜线段}}S_i=A+B$ 得证.

综上，$\text{AUC}=1-\ell_{\text{rank}}$ 得证。

2.4

χ^2 检验过程：

对于我们要研究的指标，总体被分为 k 类：$A_i,i=1,2,\ldots,k$。理论上我们知道或者猜想这 k 类数据占总体的比例为：

H_0 : A_i 的占比为 $p_i,i=1,2,\ldots,k$

检验这个原假设，根据收集的每个分类 A_i 的实际频数 n_i ，结合理论频数，有：

$$\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}\sim\chi^2(k-1)$$

再通过查表比较大小，$\chi^2<\chi^2_{\alpha,k-1}$ 则没有充分理由拒绝原假设，认为实际符合理论。

（如果是连续数据则分段然后同上述离散的计算方式）

3 HW3

作业

- 3.2
- 3.7
- 在LDA多分类情形下，试计算类间散度矩阵S_b的秩并证明
- 给出公式3.45的推导过程
- 证明$\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$是投影矩阵，并对线性回归模型从投影角度解释

3.1

- (3.18)

$$y = \frac{1}{1 + e^{-(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+b)}}$$

一阶导数：

$$\frac{dy}{dw} = \frac{xe^{-\left(w^T x+b\right)}}{\left(1 + e^{-(w^T x+b)}\right)^2} = x\left(y - y^2\right)$$

二阶导数：

$$\frac{d}{dw^T}\left(\frac{dy}{dw}\right) = x(1-2y)\left(\frac{dy}{dw}\right)^T = xx^T y(y-1)(1-2y)$$

对于 XPX^T ，有任意向量 Z ，使得：

$$Z^T XPX^T Z = \sum_i P_{ii} V_i^2 \geq 0$$

故，XPX^T 是半正定的。再由 $y \in (0.5, 1)$ 时，$y(y-1)(1-2y) < 0$ ，此时，(3.18) 的 Hessian 矩阵不是半正定的。故，(3.18) 是非凸的。

- (3.27)

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{m}\left(-y_i\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i + \ln\left(1 + e^{\boldsymbol{\beta}^{\mathrm{T}}\hat{\boldsymbol{x}}_i}\right)\right)$$

二阶导数：

$$\frac{d}{d\beta^T}\left(\frac{dl}{d\beta}\right) = xx^T p_1(x;\beta)(1 - p_1(x;\beta))$$

其中，$p \in (0, 1)$ ，故 $p_1(x;\beta)(1-p_1(x;\beta) \geq 0$ ，(3.27) 的 Hessian 矩阵是半正定的。由此，(3.27) 是凸的。

3.2

当 $3 \leq k \leq 7$ 时，需要长度 $2^{k-1}-1$ 的 ECOC 编码。在类别为 4 时，其可行的编码有 7 种，

	f0	f1	f2	f3	f4	f5	f6
c1	1	1	1	1	1	1	1
c2	0	0	0	0	1	1	1
c3	0	0	1	1	0	0	1
c4	0	1	0	1	0	1	0

当码长为 9 时，那么之后加任意两个编码，即为最优编码，因为此时再加任意的编码都是先有编码的反码，此时，类别之间最小的海明距离都为 4，不会再增加。

3.3

在 LDA 多分类情形下, 试计算类间散度矩阵 $\mathrm{S}_b=\sum_{\mathrm{i}=1}^{\mathrm{N}} m_i\left(\mu_{\mathrm{i}}-\mu\right)\left(\mu_{\mathrm{i}}-\mu\right)^{\mathrm{T}}$ 的秩并证明.

$$\mathrm{S}_b=\left[\left(\mu_1-\mu\right),\left(\mu_2-\mu\right)\ldots\left(\mu_N-\mu\right)\right]\left(\begin{array}{ccc} m_1 & & \\ & \ddots & \\ & & m_N \end{array}\right)\left(\begin{array}{c} \left(\mu_1-\mu\right)^T \\ \left(\mu_2-\mu\right)^T \\ \vdots \\ \left(\mu_N-\mu\right)^T \end{array}\right)$$

令 $M=\operatorname{diag}\left(m_1, m_2, \ldots, m_N\right), \mathrm{A}=\left(\mu_1-\mu, \mu_2-\mu, \ldots, \mu_N-\mu\right)^T$ ，则：

$$\begin{aligned}
\operatorname{rank}(\mathrm{S}_b) &= \operatorname{rank}(A^T M A) \\
&= \operatorname{rank}(A^T M^{\frac{1}{2}} M^{\frac{1}{2}} A) \\
&= \operatorname{rank}\left(A^T M^{\frac{1}{2}}\right)\left(A^T M^{\frac{1}{2}}\right)^T \\
&= \operatorname{rank}\left(A^T M^{\frac{1}{2}}\right) \\
&= \operatorname{rank}(A^T) \\
&= \operatorname{rank}\left(\mu_1-\mu, \mu_2-\mu, \ldots, \mu_N-\mu\right)
\end{aligned}$$

再由

$$\sum_{i=1}^{N} \boldsymbol{m_i \mu_i}=\left(\sum_{i=1}^{N} \boldsymbol{m_i}\right) \boldsymbol{\mu}$$

$$\sum_{i=1}^{N} \boldsymbol{m_i}\left(\boldsymbol{\mu_i}-\boldsymbol{\mu}\right)=\mathbf{0}$$

可知：$\operatorname{rank}(\mathrm{S}_b) \leq N-1$.

3.4

由题得：

若 $\boldsymbol{W}$ 为 (3.45) 的解。则对于任意常数与其的积也为解。不失一般性，令 $\operatorname{tr}(\boldsymbol{W}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{W}) = 1$，则 (3.45) 等价为：

$$\begin{aligned}&\min_{\boldsymbol{W}} -\boldsymbol{W}^{\mathrm{T}}\mathbf{S}_b\boldsymbol{W}\\&\text{s.t. } \boldsymbol{W}^{\mathrm{T}}\mathbf{S}_w\boldsymbol{W} = 1.\end{aligned}$$

拉格朗日函数为：

$$L(\boldsymbol{W}, \lambda) = -\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W}\right) + \lambda\left(\operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W}\right) - 1\right)$$

于是有：

$$\begin{aligned}\mathrm{d}L &= -\operatorname{tr}\left((\mathrm{d}\boldsymbol{W})^{\mathrm{T}}\boldsymbol{S}_b\boldsymbol{W} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\,\mathrm{d}\boldsymbol{W}\right) + \lambda\operatorname{tr}\left((\mathrm{d}\boldsymbol{W})^{\mathrm{T}}\boldsymbol{S}_w\boldsymbol{W} + \boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\,\mathrm{d}\boldsymbol{W}\right)\\&= -\operatorname{tr}\left(\left(\boldsymbol{S}_b\boldsymbol{W}\right)^{\mathrm{T}}\mathrm{d}\boldsymbol{W}\right) - \operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_b\,\mathrm{d}\boldsymbol{W}\right) + \lambda\left(\operatorname{tr}\left(\left(\boldsymbol{S}_w\boldsymbol{W}\right)^{\mathrm{T}}\mathrm{d}\boldsymbol{W}\right) + \operatorname{tr}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{S}_w\,\mathrm{d}\boldsymbol{W}\right)\right)\\&= \operatorname{tr}\left(\left[\lambda\boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{S}_w^{\mathrm{T}} + \boldsymbol{S}_w\right) - \boldsymbol{W}^{\mathrm{T}}\left(\boldsymbol{S}_b^{\mathrm{T}} + \boldsymbol{S}_b\right)\right]\mathrm{d}\boldsymbol{W}\right)\\&= \operatorname{tr}\left(\left[\left(\lambda\left(\boldsymbol{S}_w^{\mathrm{T}} + \boldsymbol{S}_w\right) - \left(\boldsymbol{S}_b^{\mathrm{T}} + \boldsymbol{S}_b\right)\right)\boldsymbol{W}\right]^{\mathrm{T}}\mathrm{d}\boldsymbol{W}\right)\end{aligned}$$

故可得到：

$$\frac{\partial L}{\partial W} = \left(\lambda\left(\boldsymbol{S}_{\mathrm{w}}^{\mathrm{T}} + \boldsymbol{S}_w\right) - \left(\boldsymbol{S}_b^{\mathrm{T}} + \boldsymbol{S}_b\right)\right)\boldsymbol{W}$$

偏导数为 0 时，有：

$$\boldsymbol{S}_b\boldsymbol{W} = \lambda\boldsymbol{S}_w\boldsymbol{W}$$

证毕.

3.5

证明 $X(X^TX)^{-1}X^T$ 是投影矩阵, 并对线性回归模型从投影角度解释.

令 $P = X(X^TX)^{-1}X^T$,

$$\begin{aligned}P^2 &= (X(X^TX)^{-1}X^T)(X(X^TX)^{-1}X^T)\\&= XI(X^TX)^{-1}X^T\\&= X(X^TX)^{-1}X^T\\&= P\end{aligned}$$

且，

$$P^T = (X(X^TX)^{-1}X^T)^T = X(X^TX)^{-1}X^T = P$$

对 X，

$$PX = X(X^TX)^{-1}X^TX = XI = X$$

故 P 是投影矩阵。

- 解释

对于 Ax = b 来说，并不是任何时候都有解，实际上大多数这种类型的方程都无解。A 的列空间的含义是方程组有解时 b 的取值空间，当 b 不在 A 的列空间时，方程无解。具体来说，当 A 是行数大于列数的长方矩阵时，方程组中的方程大于未知数个数，肯定无解。

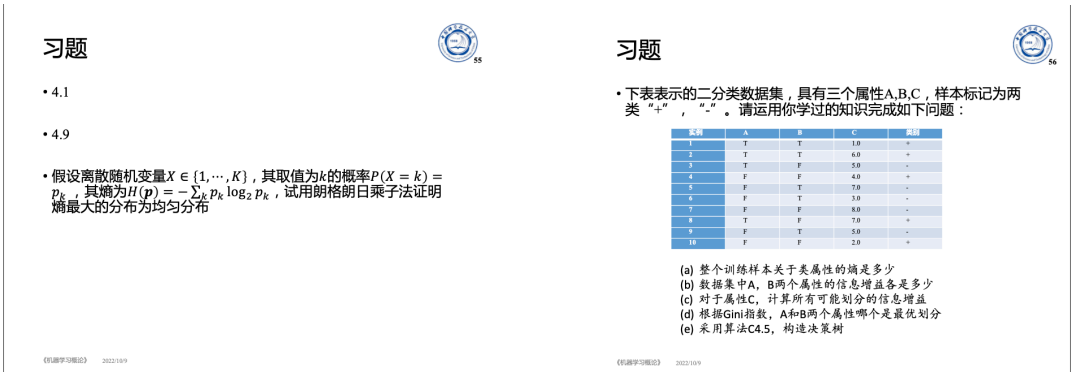
虽然方程无解，这就需要转而寻找最接近可解问题的解。对于无解方程 Ax = b，Ax 总是在列空间里（因为列空间本来就是由 Ax 确定的，和 b 无关），而 b 就不一定了，所以需要微调 b，将 b 变成列空间中最接近它的一个，Ax = b 变成了：$A\hat{x} = p$

p 就是 A 的列空间的投影，$\hat{x}$ 表示此时的 x 已经不是原来 Ax = b 中的 x，$A\hat{x} \neq b$，因为方程无解，所以不可能有 Ax = b。此时问题转换为寻找最好的 A ，让它与原方程的差距最小：

假设 A 的秩是 2，此时列空间的维度也是 2，列空间将是一个平面，平面上的向量有无数个，其中最接近 b 的当然是 b 在平面上的投影，只有 b -p 才能产生最小的误差向量.

所以说，想要求解不等方程 Ax = b，需要将 b 微调成它在 A 的列空间上的投影（列空间上的向量很多，b 在列空间上的投影是唯一的）。如果 b 在列空间中，b 的投影就是它自己；如果 b 正交于列空间，它的投影是 0。

4 HW4



4.1

因为决策树是通过属性来划分，相同属性的样本最终肯定会进入相同的叶节点。一个叶节点只有一个分类，如果样本属性相同而分类不同, 必然产生训练误差。反之，决策树只会在当前样本集合是同一类或者所有属性相同时才会停止划分，最终得到训练误差为 0 的决策树。

4.2

数据集 D 的纯度 Gini：

$$\text{Gini}(D) = 1 - \sum_{k=1}^{|y|} {p_k}^2 = \sum_{k=1}^{|y|} p_k(1 - p_k)$$

属性 a 的纯度 Gini-index：

$$\text{Gini\_}index(D, a) = \sum_{v=1}^{v} \frac{|D^v|}{|D|} \text{Gini}\left(D^v\right)$$

记 $\tilde{D}$ 为 D 中在属性 a 上没有缺失值的样本子集，$\tilde{D}^v$ 为 $\tilde{D}$ 在属性 a 上取值为 a^v 的样本子集。属性 a 的基尼指数可推广为:

$$\text{Gini\_}\, index(D,a) = p \times \text{Gini\_}\, index(\tilde{D},a) = p \times \sum_{v=1}^{V} \tilde{v} Gini\left(D^v\right)$$

4.3

拉格朗日函数为：

$$L(\boldsymbol{p},\lambda) = -\sum_{i=1}^{K} p_i \log_2 p_i + \lambda\left(\sum_{i=1}^{K} p_i - 1\right)$$

则：

$$\frac{\partial L(p,\lambda)}{\partial p_i} = -\log_2 p_i - \frac{1}{\ln 2} + \lambda = 0$$

$$\frac{\partial L(p,\lambda)}{\partial \lambda} = \sum_{i=1}^{K} p_i - 1 = 0$$

且 $p_1 = p_2 = \cdots = p_k = 2^{\lambda - \frac{1}{\ln 2}}$, 由 $\sum_{i=1}^{K} p_i = 1$ 得：

$$p_1 = p_2 = \cdots = p_k = \frac{1}{K}$$

拉朗格由上述方程组解出 p_i 及 λ，如此求得的 p_i，就是函数 $L(\boldsymbol{p},\lambda)$ 在附加条件 $\sum_{i=1}^{K} p_i = 1$ 下的可能的极值点。若这样的点只有一个，由实际问题可直接确定此即所求的点，因此熵的最大分布为均匀分布。

4.4

4.4.1

10 个样本中，正例 5 个，负例 5 个，故：

$$\text{Ent}(D) = -(0.5\log_2 0.5) \times 2 = 1$$

4.4.2

A 中'T' 4 个（里面正例 3 个），'F' 6 个（里面正例 2 个）；B 中'T' 5 个（里面正例 2 个），'F' 5 个（里面正例 3 个），故：

$$\begin{aligned}
\text{Gain}(D,A) &= \text{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \text{Ent}\left(D^v\right) \\
&= 1 + 0.4 \times (0.75\log_2 0.75 + 0.25\log_2 0.25) + 0.6 \times (\frac{2}{6}\log_2\frac{2}{6} + \frac{4}{6}\log_2\frac{4}{6}) \\
&= 0.1245
\end{aligned}$$

$$
\begin{aligned}
\operatorname{Gain}(D, B) &= \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\
&= 1 + 0.5 \times (0.6 \log_2 0.6 + 0.4 \log_2 0.4) + 0.5 \times (0.6 \log_2 0.6 + 0.4 \log_2 0.4) \\
&= 0.0290
\end{aligned}
$$

4.4.3

属性 C 有 8 个取值 1,2,3,4,5,6,7,8. 其中 5（2 负）和 7（一正一负）有两个样本其余一个。7 种二分阈值的信息增益分别为：

1. 1, 2-8:

$$
\begin{aligned}
\operatorname{Gain}(D, C) &= \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\
&= 1 - \frac{1}{10} \times 0 - \frac{9}{10} \times 0.9911 \\
&= 0.1080
\end{aligned}
$$

2. 1-2, 3-8:

$$
\begin{aligned}
\operatorname{Gain}(D, C) &= \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\
&= 1 - \frac{2}{10} \times 0 - \frac{8}{10} \times 0.9544 \\
&= 0.2364
\end{aligned}
$$

3. 1-3, 4-8:

$$
\begin{aligned}
\operatorname{Gain}(D, C) &= \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\
&= 1 - \frac{3}{10} \times 0.9183 - \frac{7}{10} \times 0.9852 \\
&= 0.0349
\end{aligned}
$$

4. 1-4, 5-8:

$$
\begin{aligned}
\operatorname{Gain}(D, C) &= \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\
&= 1 - \frac{4}{10} \times 0.8113 - \frac{6}{10} \times 0.9183 \\
&= 0.1245
\end{aligned}
$$

5. 1-5, 6-8:

$$
\begin{aligned}
\operatorname{Gain}(D, C) &= \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\
&= 1 - \frac{6}{10} \times 1 - \frac{4}{10} \times 1 \\
&= 0
\end{aligned}
$$

6. 1-6, 7-8:

$$\begin{aligned}\operatorname{Gain}(D, C) &= \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\ &= 1 - \frac{7}{10} \times 0.9852 - \frac{3}{10} \times 0.9183 \\ &= 0.0148\end{aligned}$$

7. 1-7, 8:

$$\begin{aligned}\operatorname{Gain}(D, C) &= \operatorname{Ent}(D) - \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Ent}\left(D^v\right) \\ &= 1 - \frac{9}{10} \times 0.9911 - \frac{1}{10} \times 0 \\ &= 0.1080\end{aligned}$$

4.4.4

由基尼值的定义：

$$\begin{aligned}\operatorname{Gini}(D) &= \sum_{k=1}^{|\mathcal{Y}|} \sum_{k' \neq k} p_k p_{k'} \\ &= 1 - \sum_{k=1}^{|\mathcal{Y}|} p_k^2\end{aligned}$$

对于 A，B 的 Gini 指数：

$$\begin{aligned}\text{Gini index}\ (D, A) &= \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Gini}\left(D^v\right) \\ &= 0.4 \times (1 - 0.75^2 - 0.25^2) + 0.6 \times (1 - (\frac{2}{6})^2 - (\frac{4}{6})^2) \\ &= 0.4167\end{aligned}$$

$$\begin{aligned}\text{Gini index}\ (D, B) &= \sum_{v=1}^{V} \frac{|D^v|}{|D|} \operatorname{Gini}\left(D^v\right) \\ &= 0.5 \times (1 - 0.4^2 - 0.6^2) \times 2 \\ &= 0.48\end{aligned}$$

所以 A 的 Gini 指数小，划分最优。

4.4.5

由前几小题可知：C 的信息增益选择第 2 种划分方式增益最大，由此选第二种作为 C 的处理方式。此时，

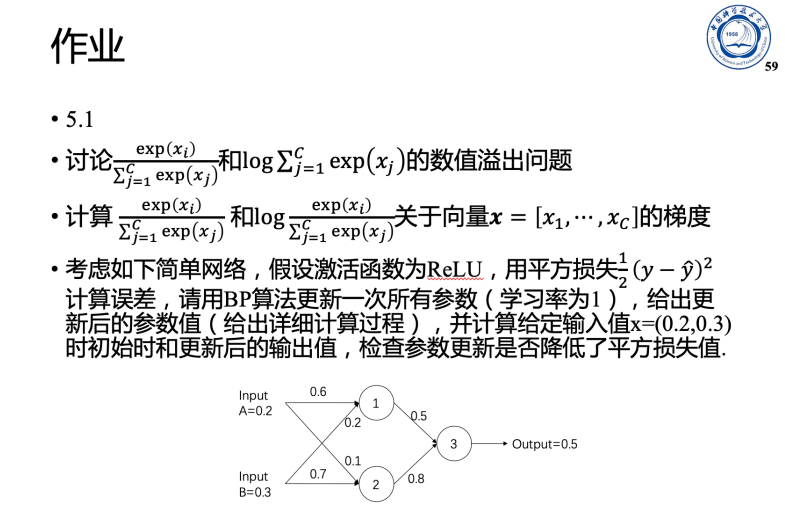
$$\text{GainRatio}(D,A)=\frac{\text{Gain}(D,A)}{\text{IV}(A)}=\frac{0.1245}{0.9710}=0.1232$$
$$\text{GainRatio}(D,B)=\frac{\text{Gain}(D,B)}{\text{IV}(B)}=\frac{0.0290}{1}=0.0290$$
$$\text{GainRatio}(D,C)=\frac{\text{Gain}(D,C)}{\text{IV}(C)}=\frac{0.2364}{0.7219}=0.3275$$

所以第一层选择 C。此时，对于 C 的 1-2 子树全为正例，没必要继续计算，该子树直接判断为正例；对于 3-7 子树有：

$$\text{GainRatio}(D',A)=\frac{\text{Gain}(D',A)}{\text{IV}'(A)}=\frac{0.9544-0.9512}{0.9544}=0.003388$$
$$\text{GainRatio}(D',B)=\frac{\text{Gain}(D',B)}{\text{IV}'(B)}=\frac{0.9544-0.9056}{1}=0.04876$$

所以该子树选择 B 作为第二层。对于第三层，由表格数据可知：B 的‘T’类：A 为‘T’则判断为正类，A 为‘F’则判断为负类；B 的‘F’类：A 为‘T’由于正负都各出现一次，因此都可以，A 为‘F’则判断为正类。

5 HW5



5.1

理想中的激活函数是阶跃函数，但是阶跃函数非连续，在 0 处不可导；线性激活函数没办法完全的拟合阶跃函数。线性函数在定义域内变换情况相同。使用线性函数作为激活函数时，无论是在隐藏层还是在输出层，不管传递多少层，其单元值都还是输入值和 x 的线性组合，若输出层也使用线性函数作为激活函数，那么就等价于线性回归，达不到激活与筛选的目的。

5.2

- $\frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}$

如果 x_i 很大，那么指数操作时可能会大于数据类型容许的最大数字，造成上溢，这将使分母或分子变为 ∞，这就使得结果是 $0, \infty$ 还是 nan 难以得到准确结论。但可以在计算之前对所有的 x_i 减去 $\max(x_i)$ 避免上溢。

- $\log \sum_{j=1}^{n} e^{x_j}$

一样的原因，e^{x_i} 计算可能溢出，也可以做同样处理，取 c=max(x_i),

$$\begin{aligned}
\log\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}^{\mathrm{x_i}}\right) &= \log\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}^{\mathrm{x_i}-\mathrm{c}}\mathrm{e}^{\mathrm{c}}\right) \\
&= \log\left(\mathrm{e}^{\mathrm{c}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}^{\mathrm{x_i}-\mathrm{c}}\right) \\
&= \log\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}^{\mathrm{x_i}-\mathrm{c}}\right) + \log\left(\mathrm{e}^{\mathrm{c}}\right) \\
&= \log\left(\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{e}^{\mathrm{x_i}-\mathrm{c}}\right) + \mathrm{c}
\end{aligned}$$

5.3

记 $f(x) = \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$

$$\frac{\partial f(x)}{\partial x_k} = \begin{cases} -\frac{\exp(x_i+x_k)}{\sum_{j=1}^{C}\exp(x_j)^2}, & k \neq i \\ \frac{\exp(x_k)\sum_{j=1,j\neq k}^{C}\exp(x_j)}{\sum_{j=1}^{C}\exp(x_j)^2}, & k = i \end{cases}$$

记 $g(x) = \log \frac{e^{x_i}}{\sum_{j=1}^{C} e^{x_j}}$

$$\frac{\partial g(x)}{\partial x_k} = \begin{cases} -\frac{\exp(x_k)}{\sum_{j=1}^{C}\exp(x_j)}, & k \neq i \\ 1 - \frac{\exp(x_k)}{\sum_{j=1}^{C}\exp(x_j)}, & k = i \end{cases}$$

5.4

由题得：

$$h_1 = 0.6 \times 0.2 + 0.2 \times 0.3 = 0.18$$

$$h_2 = 0.1 \times 0.2 + 0.7 \times 0.3 = 0.23$$

经过激活函数 ReLU: max(0,x) 变换后值为：

$$out_{h_1} = max(0, h_1) = 0.18$$

$$out_{h_2} = max(0, h_2) = 0.23$$

所以输出值：

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

此时误差为：

$$E = \frac{1}{2}(0.5 - 0.274)^2 = 0.025538$$

BP(标准)：记 $out_{h_1} : o_1$, $out_{h_2} : o_2$ 由于：

$$E = \frac{1}{2}(y - w_5 o_1 - w_6 o_2)^2$$

故可得到：

$$\frac{\partial E}{\partial w_5} = -o_1(y - w_5 o_1 - w_6 o_2)$$

一次更新:

$$\begin{aligned} w_{5new} &= w_{5old} - \eta \frac{\partial E}{\partial w_{5old}} \\ &= 0.5 - 1 \times (-0.18(0.5 - 0.274)) \\ &= 0.54068 \end{aligned}$$

同理：

$$\begin{aligned} w_{6new} &= w_{6old} - \eta \frac{\partial E}{\partial w_{6old}} \\ &= 0.8 - 1 \times (-0.23(0.5 - 0.274)) \\ &= 0.85198 \end{aligned}$$

此时有：$h_1 = o_1, h_2 = o_2$，故由于:

$$h_1 = w_1 A + w_3 B$$

有：

$$\begin{aligned} E &= \frac{1}{2}(y - w_5 o_1 - w_6 o_2)^2 \\ &= \frac{1}{2}(y - w_5 h_1 - w_6 h_2)^2 \\ &= \frac{1}{2}(y - w_5(w_1 A + w_3 B) - w_6 h_2)^2 \\ \frac{\partial E}{\partial w_1} &= -w_5 A(y - w_5(w_1 A + w_3 B) - w_6 h_2) \end{aligned}$$

所以：

$$\begin{aligned} w_{1new} &= w_{1old} - \eta \frac{\partial E}{\partial w_{1old}} \\ &= 0.6 - 1 \times (-0.5 \times 0.2(0.5 - 0.274)) \\ &= 0.6226 \end{aligned}$$

同理：

$$\begin{aligned} w_{3new} &= w_{3old} - \eta \frac{\partial E}{\partial w_{3old}} \\ &= 0.2 - 1 \times (-0.5 \times 0.3(0.5 - 0.274)) \\ &= 0.2339 \end{aligned}$$

$$
\begin{aligned}
w_{2new} &= w_{2old} - \eta \frac{\partial E}{\partial w_{2old}} \\
&= 0.1 - 1 \times (-0.8 \times 0.2(0.5 - 0.274)) \\
&= 0.13616 \\
w_{4new} &= w_{4old} - \eta \frac{\partial E}{\partial w_{4old}} \\
&= 0.7 - 1 \times (-0.8 \times 0.3(0.5 - 0.274)) \\
&= 0.75424
\end{aligned}
$$

此时，更新后的输出值为：

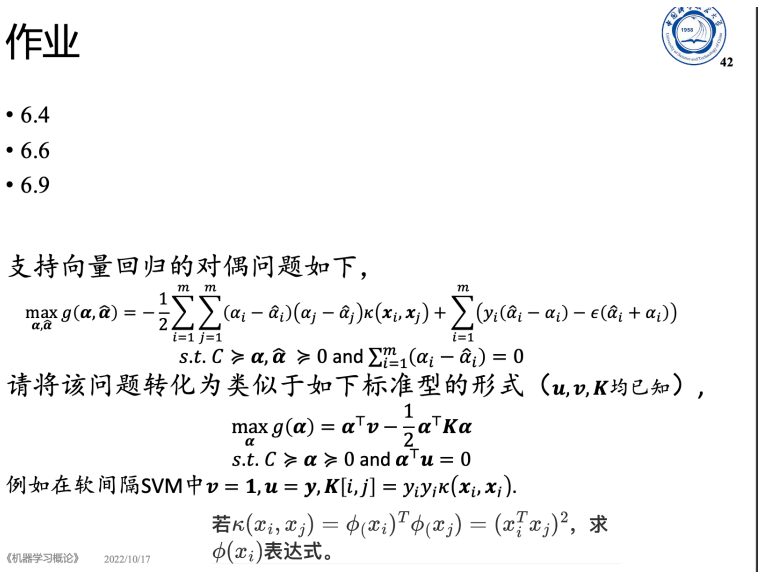
$$
\begin{aligned}
Output_{new} &= max((0.6226 \times 0.2 + 0.2339 \times 0.3), 0) \times 0.54068 \\
&\quad + max((0.13616 \times 0.2 + 0.75424 \times 0.3), 0) \times 0.85198 \\
&= 0.32125
\end{aligned}
$$

平方损失：

$$
E_{new} = \frac{1}{2}(0.5 - 0.32125)^2 = 0.01598 < 0.025538
$$

因此参数更新降低了平方损失值。

6 HW6



作业

42

- 6.4
- 6.6
- 6.9

支持向量回归的对偶问题如下，

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

$$s.t.\ C \geqslant \boldsymbol{\alpha},\hat{\boldsymbol{\alpha}} \geqslant 0 \text{ and } \textstyle\sum_{i=1}^{m}(\alpha_i - \hat{\alpha}_i) = 0$$

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），

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

$$s.t.\ C \geqslant \boldsymbol{\alpha} \geqslant 0 \text{ and } \boldsymbol{\alpha}^{\top}\boldsymbol{u} = 0$$

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

若$\kappa(x_i,x_j) = \phi_{(}x_i)^T\phi_{(}x_j) = (x_i^T x_j)^2$，求$\phi(x_i)$表达式。

《机器学习概论》 2022/10/17

6.1

线性判别分析能够解决多分类问题, 而 SVM 只能解决二分类问题。故当两类样本线性可分时，且处理二分类问题时等价。

6.2

SVM 的目的是求出与支持向量有最大化距离的直线, 以每个样本为圆心，该距离为半径做圆，可以近似认为圆内的点与该样本属于相同分类。如果出现了噪声，那么这个噪声所带来的错误分类也将最大化，所以 SVM 对噪声是很敏感的。

6.3

使用对率损失函数 ℓ_{log} 来代替式 (6.29) 中的 0/1 损失函数，则几乎就得到了对率回归模型 (3.27)，即：
由 SVM 模型：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \quad \text{s.t. } y_i\left(\boldsymbol{w}^\top \boldsymbol{x}_i + b\right) \geq 1 \quad i = 1, \cdots, m$$

软间隔的支持向量机：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m} \ell_{0/1}\left(y_i\left(\boldsymbol{w}^\top \phi\left(\boldsymbol{x}_i\right) + b\right) - 1\right)$$

令 $y_i\left(\boldsymbol{w}^\top \phi\left(\boldsymbol{x}_i\right) + b\right) = Z$，对率损失：

$$\ell_{\log}(z) = \log\left(1 + e^{-z}\right) = \log\left(\frac{1+e^z}{e^z}\right) = \log\left(1 + e^z\right) - z$$

即核对率回归：

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\left(-z + \log\left(1 + e^z\right)\right)$$

$$\ell(\beta) = \sum_{i=1}^{m}\left(-y_i\beta^\top \widehat{x}_i + \log\left(1 + e^{\beta^\top \widehat{x}_i}\right)\right)$$

6.4

令

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha \\ \hat{\alpha} \end{pmatrix} \in \mathbb{R}^{2m}, v = \begin{pmatrix} -\boldsymbol{y} - \epsilon \\ \boldsymbol{y} - \epsilon \end{pmatrix} \in \mathbb{R}^{2m}$$

$$\boldsymbol{K} = \begin{pmatrix} K & -K \\ -K & K \end{pmatrix} \in \mathbb{R}^{2m\times 2m} \text{ 其中} K_{ij} = \boldsymbol{\kappa}\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right)$$

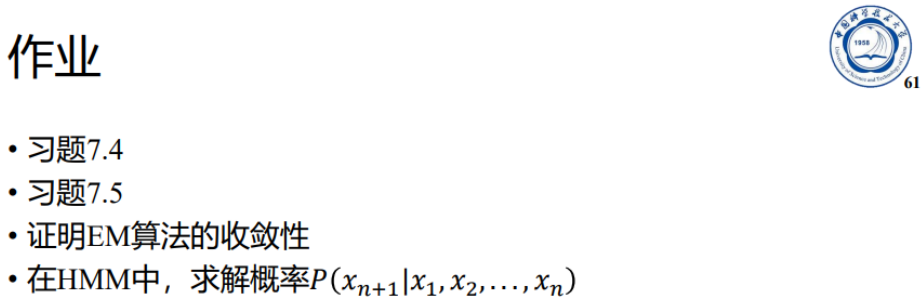
则可以得到结论.

$$
\begin{aligned}
\kappa\left(x_i, x_j\right) &= \left(x_i^T x_j\right)^2 \\
&= \left(x_i^T x_j\right)\left(x_i^T x_j\right) \\
&= \left(\sum_{p=1}^{d} x_{pi} x_{pj}\right)\left(\sum_{q=1}^{d} x_{qi} x_{qj}\right) \\
&= \sum_{p=1}^{d}\sum_{q=1}^{d} x_{pi} x_{pj} x_{qi} x_{qj} \\
&= \begin{pmatrix} x_1x_1 \\ x_1x_2 \\ \vdots \\ x_1x_d \\ \vdots \\ x_2x_d \\ \vdots \\ x_dx_1 \\ \vdots \\ x_dx_d \end{pmatrix} \times \begin{pmatrix} x_1x_1 \\ x_1x_2 \\ \vdots \\ x_1x_d \\ \vdots \\ x_2x_d \\ \vdots \\ x_dx_1 \\ \vdots \\ x_dx_d \end{pmatrix}
\end{aligned}
$$

则可以得到：

$$
\phi : \boldsymbol{x} \in \mathbb{R}^d \rightarrow \begin{pmatrix} x_1x_1 \\ x_1x_2 \\ \vdots \\ x_1x_d \\ \vdots \\ x_2x_d \\ \vdots \\ x_dx_1 \\ \vdots \\ x_dx_d \end{pmatrix} \in \mathbb{R}^{d^2}
$$

7 HW7



7.1

实践中使用式 (7.15) 决定分类类别时, 若数据的维数非常高,则概率连乘 $\prod_{i=1}^{d} P(x_i|c)$ 的结果通常会非常接近于 0 从而导致下溢. 试述防止下溢的可能方案.

(7.15):

$$h_{nb}(\boldsymbol{x}) = \arg\max_{c\in\mathcal{Y}} P(c) \prod_{i=1}^{d} P\left(x_i \mid c\right)$$

若连乘的式子太多, 导致乘积接近 0。由于属性个数是已知的, 可以对每个乘式做适当次的开方处理，可以保证结果不会为 0。另外也可以对各项取对数，当累加太多时，可能导致和接近负无穷。可以对每个加数除以属性的个数, 来防止溢出。

7.2

试证明：二分类任务中两类数据满足高斯分布且方差相同时，线性判别分析产生贝叶斯最优分类器.

假设 1 类样本均值为 u_1, 2 类样本均值为 u_2.

由题意，数据满足高斯分布且方差相同，因此，当样本足够大时，认为等价于：线性判别分析公式 $J = \frac{\left|w^T(u_1-u_2)\right|^2}{w^T(\Sigma_1+\Sigma_2)w}$ 求最大值；对 $\frac{1}{J} = \frac{w^T(\Sigma_1+\Sigma_2)w}{|w^T(u_1-u_2)|^2} = \sum_i \frac{(1-y_i)\left|w^T(x_i-u_1)\right|^2+y_i\left|w^T(x_i-u_2)\right|^2}{|w^T(u_1-u_2)|^2}$ 求最小值。

最优贝叶斯分类器使得每个训练样本的后验概率 $P(c|x)$ 最大，对应线性判别分析中，离对应分类的中心距离平方除以两个分类中心的距离平方越小。即求

$$\sum_i \frac{(1-y_i)\left|w^T\left(x_i-u_1\right)\right|^2+y_i\left|w^T\left(x_i-u_2\right)\right|^2}{\left|w^T\left(u_1-u_2\right)\right|^2}$$

的最小值。

此时，显然有两个式子相同。故，线性判别分析产生最优贝叶斯分类器。

7.3

要证明 EM 算法收敛，即证明对数似然函数的值在迭代中一直增大。即

$$\sum_{i=1}^{m} \log P\left(x^{(i)}; \theta^{j+1}\right) \geq \sum_{i=1}^{m} \log P\left(x^{(i)}; \theta^{j}\right)$$

由于

$$L\left(\theta, \theta^{j}\right) = \sum_{i=1}^{m} \sum_{z^{(i)}} P\left(z^{(i)} \mid x^{(i)}; \theta^{j}\right) \log P\left(x^{(i)}, z^{(i)}; \theta\right)$$

令

$$H\left(\theta, \theta^{j}\right) = \sum_{i=1}^{m} \sum_{z^{(i)}} P\left(z^{(i)} \mid x^{(i)}; \theta^{j}\right) \log P\left(z^{(i)} \mid x^{(i)}; \theta\right)$$

相减得到

$$\sum_{i=1}^{m} \log P\left(x^{(i)}; \theta\right) = L\left(\theta, \theta^{j}\right) - H\left(\theta, \theta^{j}\right)$$

分别取 θ 为 θ^{j+1} 和 θ^{j}, 再相减得到

$$\sum_{i=1}^{m} \log P\left(x^{(i)}; \theta^{j+1}\right) - \sum_{i=1}^{m} \log P\left(x^{(i)}; \theta^{j}\right) = \left[L\left(\theta^{j+1}, \theta^{j}\right) - L\left(\theta^{j}, \theta^{j}\right)\right] - \left[H\left(\theta^{j+1}, \theta^{j}\right) - H\left(\theta^{j}, \theta^{j}\right)\right]$$

此时，要证明 EM 算法的收敛性，只需要证明上式的右边是非负的即可。

由于 θ^{j+1} 使得 $L(\theta, \theta^{j})$ 极大，故

$$L\left(\theta^{j+1}, \theta^{j}\right) - L\left(\theta^{j}, \theta^{j}\right) \geq 0$$

再由于

$$\begin{aligned}
H\left(\theta^{j+1}, \theta^{j}\right) - H\left(\theta^{j}, \theta^{j}\right) &= \sum_{i=1}^{m} \sum_{z^{(i)}} P\left(z^{(i)} \mid x^{(i)}; \theta^{j}\right) \log \frac{P\left(z^{(i)} \mid x^{(i)}; \theta^{j+1}\right)}{P\left(z^{(i)} \mid x^{(i)}; \theta^{j}\right)} \\
&\leq \sum_{i=1}^{m} \log \left(\sum_{z^{(i)}} P\left(z^{(i)} \mid x^{(i)}; \theta^{j}\right) \frac{P\left(z^{(i)} \mid x^{(i)}; \theta^{j+1}\right)}{P\left(z^{(i)} \mid x^{(i)}; \theta^{j}\right)} \right) \\
&= \sum_{i=1}^{m} \log \left(\sum_{z^{(i)}} P\left(z^{(i)} \mid x^{(i)}; \theta^{j+1}\right) \right) \\
&= 0
\end{aligned}$$

故，EM 算法的收敛性得证.

7.4

$$\begin{aligned}
P\left(x_{n+1} \mid x_1, x_2, \ldots, x_n\right) &= \frac{P\left(x_1, x_2, \ldots, x_{n+1}\right)}{P\left(x_1, x_2, \ldots, x_n\right)} \\
&= \frac{P\left(y_1\right) P\left(x_1 \mid y_1\right) \prod_{k=2}^{n+1} P\left(y_k \mid y_{k-1}\right) P\left(x_k \mid y_k\right)}{P\left(y_1\right) P\left(x_1 \mid y_1\right) \prod_{k=2}^{n} P\left(y_k \mid y_{k-1}\right) P\left(x_k \mid y_k\right)} \\
&= P\left(y_{n+1} \mid y_n\right) P\left(x_{n+1} \mid y_{n+1}\right)
\end{aligned}$$

8 HW8

HW8&9 如下：

- 8.2
- 8.8
- 给定任意的两个相同长度向量$\boldsymbol{x}, \boldsymbol{y}$，其余弦距离为$1-\frac{\boldsymbol{x}^{\top} \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}$，证明余弦距离不满足传递性，而余弦夹角$\arccos \left(\frac{\boldsymbol{x}^{\top} \boldsymbol{y}}{|\boldsymbol{x}||\boldsymbol{y}|}\right)$满足
- 证明k-means算法的收敛性
- 在k-means算法中替换欧式距离为其他任意的度量，请问“聚类簇”中心如何计算？

8.1

式 (8.5)

$$\ell_{\exp}(H \mid \mathcal{D})=\mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}}\left[e^{-f(\boldsymbol{x}) H(\boldsymbol{x})}\right]$$

由式 (8.4)

$$H(\boldsymbol{x})=\sum_{t=1}^{T} \alpha_t h_t(\boldsymbol{x})$$

再由式 (8.11)

$$\alpha_t=\frac{1}{2} \ln \left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

可以看出分类器的权重只与分类器的错误率负相关。

对于式 (8.5) 的情况时，由于 $f(x) \in\{+1,-1\}$，$H(x)$ 为实数，当两者同号时，有 $f(x) H(x)>0$，此时，$e^{-f(x) H(x)}=e^{-|H(x)|}<1$，$|H(x)|$ 越大损失函数越小；两者异号时，有 $f(x) H(x)<0$，此时，$e^{-f(x) H(x)}=e^{|H(x)|}>1$，$|H(x)|$ 越大损失函数越大。且由于指数函数的性质，在两个区间的时候损失函数都是单调的。

由此，当使用任意损失函数 $\ell(-f(x) H(x))$，且对于 $H(x)$ 在区间 $[-\infty, \delta](\delta>0)$ 单调递减的情况，仍然满足损失函数是分类任务原本的 0 / 1 损失函数的一致的替代损失函数。

8.2

MultiBoosting 由于集合了 Bagging，Wagging，AdaBoost，可以有效的降低误差和方差，特别是误差，但是训练成本和预测成本都会显著增加。

Iterative Bagging 相比 Bagging 会降低误差，但是方差上升。由于 Bagging 本身就是一种降低方差的算法，所以 Iterative Bagging 相当于 Bagging 与单分类器的折中。

9 HW9

9.1

- 余弦距离不满足传递性：

  不妨令空间中三个点的坐标如下：

$$x = (1,0), y = (0,1), z = (\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2})$$

则，余弦距离有：

$$d(x,y) = 1, d(x,z) = 1 - \frac{\sqrt{2}}{2}, d(z,y) = 1 - \frac{\sqrt{2}}{2}$$

此时，$d(x,z) + d(y,z) < d(x,y)$. 故余弦距离不满足传递性。

- 余弦夹角满足传递性:

  证明：

$$\arccos\left(\frac{(\boldsymbol{x}^\top \boldsymbol{y})}{|\boldsymbol{x}||\boldsymbol{y}|}\right) \leq \arccos\left(\frac{(\boldsymbol{x}^\top \boldsymbol{z})}{|\boldsymbol{x}||\boldsymbol{z}|}\right) + \arccos\left(\frac{(\boldsymbol{z}^\top \boldsymbol{y})}{|\boldsymbol{z}||\boldsymbol{y}|}\right)$$

  两侧同时取 cos，由于在区间 $[0,\pi]$ 上单调递减，于是在该区间有：

$$\begin{aligned}
\frac{(\boldsymbol{x}^\top \boldsymbol{y})}{|\boldsymbol{x}\|\boldsymbol{y}|} &\geq \cos\left(\arccos\left(\frac{(\boldsymbol{x}^\top \boldsymbol{z})}{|\boldsymbol{x}\|\boldsymbol{z}|}\right) + \arccos\left(\frac{(\boldsymbol{z}^\top \boldsymbol{y})}{|\boldsymbol{z}||\boldsymbol{y}|}\right)\right) \\
&= \frac{(\boldsymbol{x}^\top \boldsymbol{z})}{|\boldsymbol{x}\|\boldsymbol{z}|} * \frac{(\boldsymbol{z}^\top \boldsymbol{y})}{|\boldsymbol{z}||\boldsymbol{y}|} - \sqrt{\left(1 - \left(\frac{(\boldsymbol{x}^\top \boldsymbol{z})}{|\boldsymbol{x}\|\boldsymbol{z}|}\right)^2\right) * \left(1 - \left(\frac{(\boldsymbol{z}^\top \boldsymbol{y})}{|\boldsymbol{z}\|\boldsymbol{y}|}\right)^2\right)} \\
&= \frac{\boldsymbol{x}^\top (\boldsymbol{z}\boldsymbol{z}^\top) \boldsymbol{y}}{|\boldsymbol{x}\|\boldsymbol{z}||\boldsymbol{z}\|\boldsymbol{y}|} - \sqrt{\left(1 - \left(\frac{(\boldsymbol{x}^\top \boldsymbol{z})}{|\boldsymbol{x}\|\boldsymbol{z}|}\right)^2\right) * \left(1 - \left(\frac{(\boldsymbol{z}^\top \boldsymbol{y})}{|\boldsymbol{z}\|\boldsymbol{y}|}\right)^2\right)} \\
&= \frac{(\boldsymbol{x}^\top \boldsymbol{y})}{|\boldsymbol{x}\|\boldsymbol{y}|} - \sqrt{\left(1 - \left(\frac{\boldsymbol{x}^\top \boldsymbol{z})}{|\boldsymbol{x}\|\boldsymbol{z}|}\right)^2\right) * \left(1 - \left(\frac{(\boldsymbol{z}^\top \boldsymbol{y})}{|\boldsymbol{z}||\boldsymbol{y}|}\right)^2\right)}
\end{aligned}$$

由于：$\sqrt{\left(1 - \left(\frac{\boldsymbol{x}^\top \boldsymbol{z})}{|\boldsymbol{x}\|\boldsymbol{z}|}\right)^2\right) * \left(1 - \left(\frac{(\boldsymbol{z}^\top \boldsymbol{y})}{|\boldsymbol{z}||\boldsymbol{y}|}\right)^2\right)} \geq 0$，故不等式得证。余弦夹角满足传递性。

9.2

证明 k-means 算法的收敛性

即证明：

$$\min J\left(u_1, u_2, \ldots u_k\right) = 1/2 \sum_{i=1}^{m} \sum_{j=1}^{k} \left(x_i - u_j\right)^2$$

1）单调；2）有界。

1. 单调

   (a) 更新 m 个点的归属类别：将 m 个点归属到已有的 k 个中心，其中的判定就是离哪个中心点近，就归属到那一类，此时损失函数为 $J_0(u_1, u_2, \ldots u_k)$. 在中心确定的情况下，如果不归属到最近的中心，归属到其他中心，此刻损失函数假设为 $J_1(u_1, u_2, \ldots u_k)$，一定有 $J_0 \leq J_1$. （假如其中 m-1 个点都归属到最近的中心了，某一个点没有归属到最近中心，那么如果把这个点放到最近的那个中心，损失函数就会降低）

   (b) 假设第一个类别有 N_1 个点,中心为 u_j,这个类别的损失函数部分为 $1/2\sum_{i=1}^{N_1}(x_i - u_1)^2$，且有：$1/2\sum_{i=1}^{N_1}(x_i - u_1)^2 \geq 1/2\sum_{i=1}^{N_1}(x_i - x_{\text{ave}})^2$ 其中x_{ave}，x_{ave} 为其均值中心。固定归属类别，通过把均值中心作为类别中心，降低了 J

2. 有界

   $J(u_1, u_2, \ldots u_k) \geq 0$

故，单调，有界，收敛。

9.3

在 k-means 算法中替换欧氏距离为其他任意的度量，请问聚类簇“中心”如何计算？

直接进行平均计算得到的均值是符合欧几里得距离定义下的均值的定义的，但是并不一定符合其它距离计算公式。不同距离的定义方式需要设计一套单独的“均值”的定义方式进行聚类簇“中心”的计算。若损失函数是凸的，则直接对其求梯度并令其等于 0，可得到聚类中心。同时，很可能完全定义不出来，如曼哈顿距离。

10 HW10

习题

- 记$\text{err}^*(\boldsymbol{x}) = 1 - \max_{c \in \mathcal{Y}} P(c|\boldsymbol{x})$， $\text{err}(\boldsymbol{x}) = 1 - \sum_c P(c|\boldsymbol{x})P(c|\boldsymbol{z})$，其中$\boldsymbol{z}$为$\boldsymbol{x}$的最近邻，试证明在样本无穷多时

$$\text{err}^*(\boldsymbol{x}) \leq \text{err}(\boldsymbol{x}) \leq \text{err}^*(\boldsymbol{x})\left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}| - 1} \times \text{err}^*(\boldsymbol{x})\right)$$

  提示：柯西-施瓦兹不等式$(\sum_i a_i)^2 \leq n\left(\sum_i a_i^2\right)$

- 10.4

- 求解20页的优化问题

附加题：令$\mathbf{M} = PP^T$，那么下列问题还是凸优化问题吗？试证明之。

$$\min_P \sum_{(x_i, x_j) \in \mathcal{M}} \|x_i - x_j\|_{\mathbf{M}}^2 \quad \text{s. t.} \sum_{(x_i, x_j) \in \mathcal{C}} \|x_i - x_j\|_{\mathbf{M}}^2 \geq 1$$

10.1

由题可知：

$$\mathrm{err}^\star(\boldsymbol{x}) = 1 - \max_{c\in Y} P(c \mid \boldsymbol{x})$$

$$\mathrm{err}(\boldsymbol{x}) = 1 - \sum_c P(c \mid \boldsymbol{x}) P(c \mid \boldsymbol{z})$$

要证 $\mathrm{err}^\star(\boldsymbol{x}) \le \mathrm{err}(\boldsymbol{x})$

即证 $\max_{c\in Y} P(c \mid \boldsymbol{x}) \ge \sum_c P(c \mid \boldsymbol{x}) P(c \mid \boldsymbol{z})$

$$\begin{aligned}\sum_c P(c \mid \boldsymbol{x}) P(c \mid \boldsymbol{z}) &\approx \sum_c P(c \mid \boldsymbol{x})^2 \\ &\le \max_{c\in Y} P(c \mid \boldsymbol{x}) \sum_c P(c \mid \boldsymbol{x}) \\ &= \max_{c\in Y} P(c \mid \boldsymbol{x})\end{aligned}$$

得证.

要证 $\mathrm{err}(\boldsymbol{x}) \le \mathrm{err}^\star(\boldsymbol{x}) \left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \times \mathrm{err}^\star(\boldsymbol{x})\right)$

$$\begin{aligned}\mathrm{err}(x) &= 1 - \sum_x P^2(c \mid x) \\ &= (2 - \mathrm{err}^\star(x))\,\mathrm{err}^\star(x) - \sum_{c|c^\star} P^2(c \mid x)\end{aligned}$$

由 Cauchy 不等式：

$$\sum_{c|c^\star} P^2(c \mid x) \ge \frac{1}{|\mathcal{Y}|}(1 - P(c^\star \mid x))^2 = \frac{1}{|\mathcal{Y}|}\,\mathrm{err}^\star(\boldsymbol{x})^2$$

即：$\mathrm{err}(\boldsymbol{x}) \le \mathrm{err}^\star(\boldsymbol{x}) \left(2 - \frac{|\mathcal{Y}|}{|\mathcal{Y}|-1} \times \mathrm{err}^\star(\boldsymbol{x})\right)$

10.2

在实践中,协方差矩阵XX^T的特征值分解常由中心化后的样本矩阵X的奇异值分解代替,试述其原因.

因为两者等价，下证明：

X 的奇异值分解为：

$$X = U\Sigma V^T, \quad XX^T = U\Sigma V^T \left(U\Sigma V^T\right)^T = U\Sigma V^T V \Sigma^T U^T$$

此时，$V^T V = 1, U^T U = 1$，故，

$$XX^T = U\Sigma\Sigma^T U^T = U\Lambda U^T$$

X 的特征值分解时，$X^T X = P\Lambda P^T$, P=U 则两者等价。

还有，用奇异值分解代替特征分解，节省了计算和存储的成本，且计算精度较高。

10.3

主成分分析—求解

$$\max_{\boldsymbol{W}} \operatorname{tr}\left(\boldsymbol{W}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{W}\right) \quad \text{s.t. } \boldsymbol{W}^T\boldsymbol{W} = \boldsymbol{I}_{d'}$$

使用拉格朗日乘子法可得

$$XX^TW = \Lambda W$$

由题得：拉格朗日函数为：

$$L(w,\theta) = \operatorname{tr}\left(W^TXX^TW\right) - \operatorname{tr}\left(\theta^T\left(w^TW - I\right)\right)$$

其中，θ 乘子矩阵。考虑其中前 d' 个最大的特征值，令 $\theta = diag(\lambda_1, \Lambda_2, \ldots, \lambda_{d'})$ 则：

$$\begin{aligned}\frac{\partial L(W,\theta)}{\partial W} &= \frac{\partial\left(\operatorname{tr}\left(W^TXX^TW\right)\right.}{\partial W} - \frac{\partial}{\partial W}\left(\operatorname{tr}\left(\theta^T\left(W^TW - I\right)\right)\right) \\ &= 2XX^TW - W\theta - W\theta^T \\ &= 2XX^TW - 2W\theta\end{aligned}$$

为 0 时，有

$$XX^TW = 2W\theta$$

代回，有：

$$\begin{aligned}&\max_{W} \operatorname{tr}\left(W^TXx^TW\right) \\ =&\max_{W} \sum_{i=1}^{d'} W_i^TXx^TW_i \\ =&\max_{i=1} \sum_{i}^{'} \lambda_i W_i^TW_i \\ =&\max^{x} \sum_{i=1}^{d'} \lambda_i\end{aligned}$$

11 HW11

作业

45

- 11.5
- 11.7
- PPT 20页：证明回归和对率回归的损失函数的梯度是否满足L-Lipschitz条件，并求出L

11.1

结合图11.2,试举例说明 L1 正则化在何种情形下不能产生稀疏解.

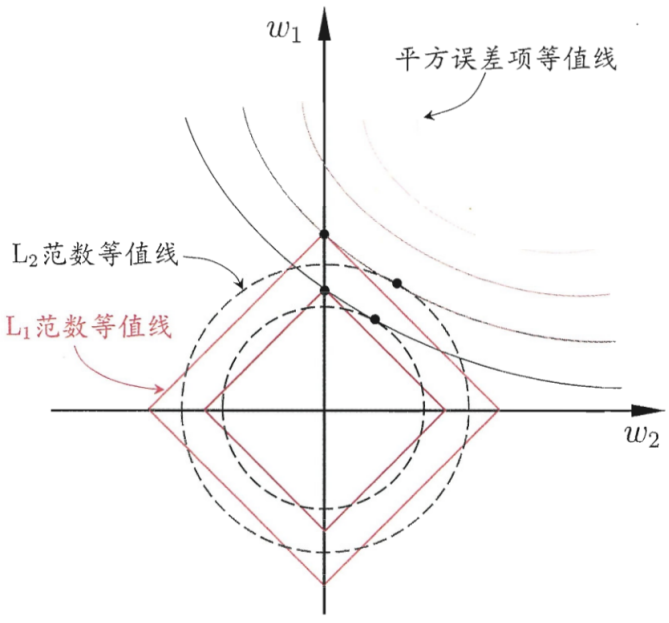


图 11.2 L_1 正则化比 L_2 正则化更易于得到稀疏解

L1 正则化之所以可以产生稀疏解，主要是因为平方误差项等值线与 L1 等值线的第一个交点位于坐标轴上，如书上图 11.2 所示，当平方误差项等值线的曲率比较大时，就会导致其与 L1 等值线的第一个交点不再位于坐标轴上，此时就无法产生稀疏解。

11.2

试述直接求解 L0 范数正则化会遇到的困难.

L0 范数是不连续的，而且是非凸函数，无法通过优化直接求解，必须采用遍历的方式，因此导致这个问题是个 NP 难问题。

11.3

11.3.1

回归：

$$f(w) = \|Xw - y\|^2$$

故，

$$\nabla f(w) = 2X^T(Xw - y)$$

任取 w_1, w_2,

$$\begin{aligned}\|\nabla f(w_1) - \nabla f(w_2)\|_2 &= 2\left\|X^T(Xw_1 - y) - X^T(Xw_2 - y)\right\|_2 \\ &= 2\left\|X^TX(w_1 - w_2)\right\|_2 \\ &\le 2\left\|X^TX\right\|_2\|w_1 - w_2\|_2\end{aligned}$$

则，满足 L-Lipschitz 条件，且 $\mathcal{L} = 2\left\|X^TX\right\|_2 = 2\|X\|_2^2$

11.3.2

对率回归：

$$f(w) = \sum_{i=1}^{m}(-yw^Tx_i + \ln(1 + e^{w^Tx_i}))$$

故，

$$\nabla f(w) = -\sum_{i=1}^{m}(y_i - \frac{e^{w^Tx_i}}{1 + e^{w^Tx_i}})x_i$$

任取 w_1, w_2,

$$\begin{aligned}\|\nabla f(w_1) - \nabla f(w_2)\|_2 &= \left\|(-\sum_{i=1}^{m}(y_i - \frac{e^{w_1^Tx_i}}{1 + e^{w_1^Tx_i}})x_i) - (-\sum_{i=1}^{m}(y_i - \frac{e^{w_2^Tx_i}}{1 + e^{w_2^Tx_i}})x_i)\right\|_2 \\ &= \left\|\sum_{i=1}^{m}(\frac{e^{w_1^Tx_i}}{1 + e^{w_1^Tx_i}} - \frac{e^{w_2^Tx_i}}{1 + e^{w_2^Tx_i}})x_i\right\|_2\end{aligned}$$

令 $g(\alpha) = \frac{e^\alpha}{1+e^\alpha}$

$$\nabla g(\alpha) = \frac{e^\alpha}{(1 + e^\alpha)^2} \ge 0, \quad \nabla^2 g(\alpha) = -\frac{e^\alpha(e^\alpha - 1)}{(1 + e^\alpha)^3}$$

二阶导数为 0 时，$\alpha = 0$，一阶导数在此处取得极大值 $\frac{1}{4}$.

$$\begin{aligned}\|g(\alpha) - g(\beta)\|_2 &= \left\| \frac{e^{\alpha}}{1+e^{\alpha}} - \frac{e^{\beta}}{1+e^{\beta}} \right\|_2 \\ &= \|\nabla(\xi)(\alpha - \beta)\|_2 \quad \xi \in (\alpha, \beta) \\ &\leq \frac{1}{4}\|\alpha - \beta\|_2\end{aligned}$$

故，

$$\begin{aligned}\|\nabla f(w_1) - \nabla f(w_2)\|_2 &\leq \frac{1}{4} \left\| \sum \left(w_2^T x_i - w_2^T x_i \right) x_i \right\|_2 \\ &= \frac{1}{4} \left\| x^T x \left(w_1 - w_2 \right) \right\|_2 \\ &\leq \frac{1}{4} \left\| x^T x \right\|_2 \|w_1 - w_2\|_2\end{aligned}$$

则，满足 L-Lipschitz 条件，且 $\mathcal{L} = \frac{1}{4} \left\| X^T X \right\|_2 = \frac{1}{4} \|X\|_2^2$

12 HW12

无

13 HW13

无

14 HW14

1. 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$\lambda^{-1}$的正态分布$\mathcal{N}(\mu, \lambda^{-1})$中独立采样而得到。假设$\mu$和$\lambda$的先验分布为$p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\kappa_0\lambda)^{-1})\mathrm{Gam}(\lambda|a_0, b_0)$ 其中 $\mathrm{Gam}(\lambda|a_0, b_0) = \frac{1}{\Gamma(a_0)} b_0^{a_0} \lambda^{a_0 - 1} \exp(-b_0\lambda)$

(1) 请写出联合概率分布$p(D, \mu, \lambda)$

(2) 请写出证据下界（即变分推断的优化目标），并证明其为观测数据边际似然$\sum_{i=1}^{m} \log p(x_i)$的下界

(3) (3) 请用变分推断法近似推断后验概率$p(\mu, \lambda|D)$

2. PPT 46，给出CRF的预测问题的解法

14.1

14.1.1

$$
\begin{aligned}
p\left(D, \mu, \lambda\right) &= p(\mu, \lambda) \cdot p(D \mid \mu, \lambda) \\
&= p(\mu, \lambda) \cdot \prod_{i=1}^{m} p\left(x_i \mid \mu, \lambda\right) \\
&= \frac{1}{\sqrt{2\pi(\kappa_0\lambda)^{-1}}} \exp\left\{-\frac{1}{2(\kappa_0\lambda)^{-1}}\left(\mu-\mu_0\right)^2\right\} \\
&\quad \cdot \frac{1}{\Gamma\left(a_0\right)} b_0^{a_0} \cdot \lambda^{a_0-1} \exp\left\{-b_0\lambda\right\} \cdot \left(\frac{\lambda}{2\pi}\right)^{\frac{m}{2}} \exp\left\{-\frac{\lambda}{2}\sum_{i=1}^{m}\left(x_i-\mu\right)^2\right\}
\end{aligned}
$$

14.1.2

由题可知，变分推断的目的是寻找一个参数的概率密度函数

$$
q(z) \quad \text{s.t. } q^*(z) = \arg\min_{q(z)} KL\left(q(z) \| p(z|x)\right)
$$

其中，KL 散度为

$$
KL(q(z) \| p(z \mid x)) = \mathbf{E}[\log q(z)] - \mathbf{E}[\log p(z, x)] + \log p(x)
$$

再由于 KL 散度的非负性，故可得到：

$$
\begin{aligned}
ELBO &= \mathbf{E}[\log p(z, x)] - \mathbf{E}[\log q(z)] \\
&= \mathbf{E}_q[\log p(\lambda)] + \mathbf{E}_q[\log p(\mu|\lambda)] + \mathbf{E}_q[\log p(x|\mu, \lambda)] - \mathbf{E}_q[\log q(\lambda)] - \mathbf{E}_q[\log q(\mu)]
\end{aligned}
$$

14.1.3

由题可得，

$$
\frac{\partial L}{\partial q_u(u)} = E_\lambda[\log p(\mu \mid \lambda)] + E_\lambda[\log p(D \mid \mu, \lambda)] - \log q(u) = 0
$$

$$
\begin{aligned}
\log q(u) &= \frac{E[\lambda]\kappa_0}{-2}\left(\mu-\mu_0\right)^2 - \frac{E[\lambda]}{2}\sum_{i=1}^{m}\left(x_i-\mu\right)^2 \\
&= -\frac{E[x]}{2}\left[\left(\kappa_0+m\right)\mu^2 + \sum_{i=1}^{m} x_i^2 - 2\mu\left(\kappa_0\mu_0 + m\bar{x}\right)\right] \\
&= -\frac{E[\lambda]}{2}\left[\left(\kappa_0+m\right)\left(\mu - \frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m}\right)^2 + \sum_{i=1}^{m} x_i^2 - \frac{\left(\kappa_0\mu_0+m\bar{x}\right)^2}{\kappa_0+m}\right]
\end{aligned}
$$

令 $\mu_m = \frac{\kappa_0\mu_0+m\bar{x}}{\kappa_0+m}$, $\lambda_m = (\kappa_0+m)E[\lambda]$, 故可以看出：$q_u^*(u) \sim \mathcal{N}(\mu \mid \mu_m, \lambda_m)$.

再由

$$
\frac{\partial L}{\partial q_\lambda(\lambda)} = E_\mu[\log p(D \mid \lambda, \mu)] + E_\mu[\log(\mu \mid \lambda)] + E_\mu[\log p(\lambda)] - \log q(\lambda) = 0
$$

故

$$\log q^*(\lambda) = -\frac{\lambda}{2}E_\mu\left[\kappa_0\left(\mu-\mu_0\right)^2+\sum_{i=1}^{m}\left(x_i-\mu\right)^2\right]+\left(a_0-1\right)\log\lambda-b_0\lambda+\frac{m+1}{2}\log\lambda$$
$$=\left(a_0+\frac{m-1}{2}\right)\log\lambda-\left(b_0+\frac{1}{2}E_\mu\left[\kappa_0\left(\mu-\mu_0\right)^2+\sum_{i=1}^{m}\left(x_i-\mu\right)^2\right]\right)\lambda$$

令 $a_m = a_0+\frac{m+1}{2}$, $b_m = b_0+\frac{1}{2}E_\mu\left[\kappa_0\left(\mu-\mu_0\right)^2+\sum_{i=1}^{m}\left(x_i-\mu\right)^2\right]$,
可以看出：$q^*(\lambda)\sim Gam(\lambda\mid a_m,b_m)$.

故，可以得到：$q^*(\mu,\lambda)\sim\mathcal{N}\left(\mu\mid N_m,\lambda_m^{-1}\right)Gam\left(\lambda\mid a_m,b_m\right)$

在无先验的情况下有：

$$\begin{cases}\mu_0=a_0=b_0=\kappa_0=0\\ E[\lambda]=\frac{a_m}{b_m}\\ E[\mu]=\mu_m=\bar{x}\\ E[\mu^2]=\bar{x}^2+\frac{1}{mE[\lambda]}\end{cases}$$

可以解得：

$$E[\lambda]=\frac{1}{E[x^2]-\bar{x}^2}=\frac{1}{\mathrm{Var}(x)}$$
$$\lambda_m=\frac{m}{\mathrm{Var}(x)}$$
$$b_m=b_0+\frac{1}{2}E_\mu\left[\kappa_0\mu-^2+\sum_{i=1}^{m}\left(x_i-\mu\right)^2\right]$$

记带入后的 λ_m,b_m 为 λ',b'，可以得到：

$$p(\mu,\lambda\mid D)\sim\mathcal{N}\left(\mu\mid\mu_m,\lambda'\right)Gam\left(\lambda\mid am,b'\right)$$

14.2

采用维特比算法，首先求出位置 1 的各个标记 j = 1,2,...,m 的非规范化概率：

$$\delta_1(j)=\omega*F_1\left(y_0=\text{ start},y_1=j,x\right),\quad j=1,2,\ldots,m$$

由递推公式，求出位置 i 的各个标记 l = 1,2,⋯,m 的非规范化概率的最大值，同时记录非规范化概率最大值的路径：

$$\delta_i(l)=\max_{1\le j\le m}\left\{\delta_{i-1}(l)+\omega*F_i\left(y_{i-1}=j,y_i=l,x\right)\right\},\quad l=1,2,\ldots,m$$

$$\Psi_i(l)=\arg\max_{1\le j\le m}\left\{\delta_{i-1}(l)+\omega*F_i\left(y_{i-1}=j,y_i=l,x\right)\right\},\quad l=1,2,\ldots,m$$

直到 i = n 时终止，此时求得非规范化概率的最大值为：

$$\max_y(\omega*F(y,x))=\max_{1\le j\le m}\delta_n(j)$$

以及最优路径的终点

$$y_n^*=\max_{1\le j\le m}\delta_n(j)$$

由此最优路径终点返回

$$y_i^* = \Psi_{i+1}\left(y_{i+1}^*\right), \quad \mathrm{i} = \mathrm{n}-1, \mathrm{n}-2, \ldots, 1$$

最终求得最优路径.