Homework I

邵彦骏 19307110036

1. 通过经验风险最小化推导极大似然估计。证明模型是条件概率分布，当损失函数是对数损失函数时，经验风险最小化等价于极大似然估计。

   PROOF

   当损失函数时对数损失函数时，

$$R_{emp}(f) = -\frac{1}{N}\sum_{i=1}^{N}\log\left(p(y|x_i,\Theta)\right)$$

   此时的经验风险最小化等价于，

$$\underset{\Theta}{argmax}\sum_{i=1}^{N}\log\left(p(y|x_i,\Theta)\right)$$

   就是极大似然估计$\underset{\Theta}{argmax}\, l(\Theta)$。

2. The Hoeffding's inequality:

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])\geq t)\leq exp(-\frac{2nt^2}{(b-a)^2})$$

   The Hoeffding's Lemma: Let Z be a bounded random variable with $Z\in[a,b]$. Then

$$\mathbb{E}[exp(\lambda(Z-\mathbb{E}[Z]))]\leq exp(\frac{\lambda^2(b-a)^2}{8})$$"

   PROOF

$$\begin{aligned}
\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])\geq t) &= \mathbb{P}(e^{\lambda\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])}\geq e^{\lambda nt}) \\
&\leq \mathbb{E}(e^{\lambda\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])})e^{-\lambda nt}\quad (Markov\ \ inequality) \\
&= \prod_{i=1}^{n}\mathbb{E}[e^{\lambda(Z_i-\mathbb{E}[Z_i])}]e^{-\lambda nt} \\
&\leq exp(n[\frac{\lambda^2(b-a)^2}{8}-\lambda t])
\end{aligned}$$

As this inequality holds $\forall\lambda>0$, we have,

$$\underset{\lambda>0}{min}[\frac{\lambda^2(b-a)^2}{8}-\lambda t] = -\frac{2t^2}{(b-a)^2}$$

Therefore,

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^{n}(Z_i-\mathbb{E}[Z_i])\geq t)\leq exp(-\frac{2nt^2}{(b-a)^2})$$

3. 有监督学习的应用

   1. 问题背景：在社交网络中有很多复杂的结构，但有些好友关系的建立却能够被简单地预测，这是因为它们之间总是存在相似度。一些简单的机器学习模型就能很好地预测出这些潜在关系，并且为用户推荐这些潜在好友。

2. 自变量：网络图中的边，$x_{ij} = \mathbf{1}\{节点i与节点j有关联\}$

   因变量：网络图中可能的边，$y_{ij} = \mathbf{1}\{预测节点i与节点j有关联\}$

3. 可以通过社交网络中节点相似度的度量，作为二分类模型的输入，使用支持向量机或者朴素贝叶斯模型对两个节点之间有边（输出为1）和没有边（输出为0）进行预测。

4. Please read the background and then prove the following results.
   Background:
   Let $\mathbf{y} = \Psi(\mathbf{x})$, where $\mathbf{y}$ is an m-element vector, and $\mathbf{x}$ is an n-element vector. Denote

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_m}{\partial x_2} \\ \vdots & \vdots & \vdots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix}$$

Prove the results:

(a) Let $\mathbf{y} = \mathbf{A}\mathbf{x}$, where $\mathbf{y}$ is $\mathrm{m} \times 1, \mathbf{x}$ is $\mathrm{n} \times 1, \mathbf{A}$ is $\mathrm{m} \times \mathrm{n}$, and $\mathbf{A}$ does not depend on $\mathbf{x}$ then

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}^\top$$

PROOF

With definition, we have $y_i = \sum_{j=1}^{n} \mathbf{A}_{ij} \cdot x_j$ and $(\frac{\partial \mathbf{y}}{\partial \mathbf{x}})_{ij} = \frac{\partial y_j}{\partial x_i} = \mathbf{A}_{ji}$. Therefore, we can infer that $\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \mathbf{A}^T$

(b) Let the scalar α be defined by $\alpha = \mathbf{y}^\mathrm{T}\mathbf{A}\mathbf{x}$, where $\mathbf{y}$ is $\mathrm{m} \times 1, \mathbf{x}$ is $\mathrm{n} \times 1, \mathbf{A}$ is $\mathrm{m} \times \mathrm{n}$, and $\mathbf{A}$ is independent of $\mathbf{x}$ and $\mathbf{y}$, then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{A}^\top \mathbf{y}$$

PROOF

With definition, we have $\alpha = \mathbf{y}^\mathrm{T}\mathbf{A}\mathbf{x} = \sum_{j=1}^{n}\sum_{k=1}^{m} A_{kj} \cdot y_k \cdot x_j$. Therefore, we can derive,

$$(\frac{\partial \alpha}{\partial \mathbf{x}})_i = \frac{\partial \alpha}{\partial x_i} = \sum_{k=1}^{m} A_{ki} \cdot y_k = (\mathbf{A}^T y)_i$$

And now we prove that $\frac{\partial \alpha}{\partial \mathbf{x}} = \mathbf{A}^\top \mathbf{y}$

(c)For the special case in which the scalar α is given by the quadratic form $\alpha = \mathbf{x}^\mathrm{T}\mathbf{A}\mathbf{x}$ where x is $\mathrm{n} \times 1, \mathbf{A}$ is $\mathrm{n} \times \mathrm{n}$, and $\mathbf{A}$ does not depend on x, then

$$\frac{\partial \alpha}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^\mathrm{T})\mathbf{x}$$

PROOF

With definition, we have $\alpha = \sum_{j=1}^{n}\sum_{k=1}^{m} A_{kj} \cdot x_k \cdot x_j$. Therefore, we can derive,

$$(\frac{\partial\alpha}{\partial\mathbf{x}})_i = \frac{\partial\alpha}{\partial x_i} = \sum_{k=1}^{m}(A_{ki}+A_{ik})\cdot x_k = (\mathbf{A}+\mathbf{A}^{\mathrm{T}}x)_i$$

And now we prove that $\dfrac{\partial\alpha}{\partial\mathbf{x}} = \left(\mathbf{A}+\mathbf{A}^{\mathrm{T}}\right)\mathbf{x}$

(d) Let the scalar α be defined by $\alpha = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{y}$ is $\mathrm{m}\times 1$, $\mathbf{x}$ is $\mathrm{n}\times 1$, $\mathbf{A}$ is $\mathrm{m}\times\mathrm{n}$, and both $\mathbf{y}$ and $\mathbf{x}$ are functions of the vector $\mathbf{z}$, while $\mathbf{A}$ does not depend on $\mathbf{z}$. Then

$$\frac{\partial\alpha}{\partial\mathbf{z}} = \frac{\partial\mathbf{y}}{\partial\mathbf{z}}\mathbf{A}\mathbf{x} + \frac{\partial\mathbf{x}}{\partial\mathbf{z}}\mathbf{A}^{\top}\mathbf{y}$$

PROOF

With definition, we have $\alpha = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x} = \sum\limits_{j=1}^{n}\sum\limits_{k=1}^{m} A_{kj}\cdot y_k\cdot x_j$. Therefore, we will have,

$$\begin{aligned}\frac{\partial\alpha}{\partial\mathbf{z}} = \frac{\partial(\sum\limits_{j=1}^{n}\sum\limits_{k=1}^{m} A_{kj}\cdot y_k\cdot x_j)}{\partial\mathbf{z}} &= \sum_{k=1}^{m}\frac{\partial y_k}{\partial\mathbf{z}}\mathbf{A}\mathbf{x}_k + \sum_{j=1}^{n}\frac{\partial x_j}{\partial\mathbf{z}}\mathbf{A}^{\mathrm{T}}\mathbf{y}_j \\ &= \frac{\partial\mathbf{y}}{\partial\mathbf{z}}\mathbf{A}\mathbf{x} + \frac{\partial\mathbf{x}}{\partial\mathbf{z}}\mathbf{A}^{\top}\mathbf{y}\end{aligned}$$

(e) Let $\mathbf{A}$ be a nonsingular, $\mathrm{m}\times\mathrm{m}$ matrix whose elements are functions of the scalar parameter α. Then

$$\frac{\partial\mathbf{A}^{-1}}{\partial\alpha} = -\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\alpha}\mathbf{A}^{-1}$$

PROOF

First of all, we can have

$$\frac{\partial\mathbf{A}\mathbf{B}}{\partial\alpha} = \frac{\partial\mathbf{A}}{\partial\alpha}\mathbf{B} + \mathbf{A}\frac{\partial\mathbf{B}}{\partial\alpha}$$

Hence,

$$\frac{\partial\mathbf{A}\mathbf{A}^{-1}}{\partial\alpha} = \frac{\partial\mathbf{A}}{\partial\alpha}\mathbf{A}^{-1} + \mathbf{A}\frac{\partial\mathbf{A}^{-1}}{\partial\alpha} = \frac{\partial\mathbf{I}}{\partial\alpha} = 0$$

And reorganize the equation,

$$\frac{\partial\mathbf{A}^{-1}}{\partial\alpha} = -\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\alpha}\mathbf{A}^{-1}$$

(4) Please write $\hat{a}$ as the solution of the minimization problem:

$$\min_a \|\mathbf{X}a-\mathbf{y}\|$$

where $\mathbf{X}$ is a $\mathrm{n}\times\mathrm{p}$ matrix and $\mathbf{y}$ is a $\mathrm{n}\times 1$ vector. $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is nonsingular.

SOLUTION

$$\min_a \|\mathbf{X}a-\mathbf{y}\| \Leftrightarrow \min_a \|\mathbf{X}a-\mathbf{y}\|^2$$

Take derivative on the right-hand term to minimize it.

$$\frac{\partial\|\mathbf{X}a-\mathbf{y}\|^2}{\partial a} = 2\mathbf{X}^T(\mathbf{X}a-\mathbf{y}) = 0$$

Since $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is nonsingular, we can have the optimal solution,

$$a = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}\mathbf{y}$$