统计机器学习 Homework 01

姓名: 雍崔扬
学号: 21307140051

Problem 1

证明《统计学习方法》习题 1.2:
当模型是条件概率分布且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.

Solution:
设假设空间 $\mathscr{F} = \{p(y|x;\theta) : \theta \in \Theta\}$ 中的条件概率分布由参数 θ 唯一确定.
设训练集为 $D_{\text{train}} = \{(x_i, y_i) : i = 1, \ldots, n\}$ 且其数据独立同分布，联合概率密度函数为 $p_{XY}(x, y)$
损失函数 $\text{loss}(y, p(y|x;\theta)) = -\log(p(y|x;\theta))$
则经验风险函数为:

$$\begin{aligned}
\text{Risk}_{\text{emp}}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\text{loss}(y_i, p(y_i|x_i;\theta)) \\
&= -\frac{1}{n}\sum_{i=1}^{n}\log(p(y_i|x_i;\theta)) \\
&= -\frac{1}{n}\log\left\{\prod_{i=1}^{n}p(y_i|x_i;\theta)\right\} \\
&= -\frac{1}{n}ll(\theta)
\end{aligned}$$

注意到 $ll(\theta) := \log\{\prod_{i=1}^{n}p(y_i|x_i;\theta)\}$ 就是训练样本 D_{train} 在条件分布 $p(y|x;\theta)$ 的对数似然函数.
经验风险最小化策略即求解优化问题:

$$\min_{\theta\in\Theta}\text{Risk}_{\text{emp}}(\theta) = \max_{\theta\in\Theta} ll(\theta)$$

表明此时经验风险最小化得到的最优参数 $\theta_\star$ 的问题就等价于极大似然估计得到极大似然解 θ_{MLE} 的问题.

Problem 2

证明 Hoeffding 引理:
若随机变量 X 满足 $\text{E}[X] = 0$ 且 $\text{P}\{X \in [a, b]\} = 1$，则我们有:

$$\text{E}[e^{sX}] \le \exp\{\frac{1}{8}s^2(b-a)^2\} \ \ (\forall\ s \in \mathbb{R})$$

Solution:
根据 e^{sx} 在 $[a, b]$ 上的凸性可知:

$$e^{sx} \le \frac{b-x}{b-a}e^{sa} + \frac{x-a}{b-a}e^{sb} \quad (x \in [a, b])$$

将有限点的凸组合推广到级数或积分，我们有:

$$\text{E}[e^{sX}] = \frac{b-\text{E}[X]}{b-a}e^{sa} + \frac{\text{E}[X]-a}{b-a}e^{sb} \quad (\text{note that E}[X] = 0)$$

$$
\begin{aligned}
&= \frac{b}{b-a}e^{sa} + \frac{-a}{b-a}e^{sb} \\
&= e^{sa}\left(\frac{b}{b-a} - \frac{a}{b-a}e^{s(b-a)}\right) \\
&= e^{sa}\left[1 + \frac{a}{b-a}(1-e^{s(b-a)})\right] \\
&= \exp\left\{\frac{a}{b-a}s(b-a) + \log\left(1 + \frac{a}{b-a}(1-e^{s(b-a)})\right)\right\} \\
&= \exp\{g(s(b-a))\}
\end{aligned} \tag{2.1}
$$

其中 $g(h) := \frac{a}{b-a}h + \log\left(1 + \frac{a}{b-a}(1-e^h)\right)$

$$g(h) = \frac{a}{b-a}h + \log\left(1 + \frac{a}{b-a}(1-e^h)\right)$$

$$g(0) = 0$$

$$g'(h) = \frac{a}{b-a} + \frac{1}{1+\frac{a}{b-a}(1-e^h)}\left(-\frac{a}{b-a}\right)e^h = \frac{b}{b-a} - \frac{\frac{b}{b-a}}{\frac{b}{b-a} - \frac{a}{b-a}e^h} = \frac{b}{b-a} - \frac{b}{b-ae^h}$$

$$g'(0) = 0$$

$$g''(h) = -\frac{\mathrm{d}}{\mathrm{d}h}\left\{\frac{b}{b-ae^h}\right\} = -\frac{bae^h}{(b-ae^h)^2} \le \frac{\frac{1}{4}(b-ae^h)^2}{(b-ae^h)^2} = \frac{1}{4}$$

根据 Taylor 定理可知: 存在 $\theta \in (0,1)$ 使得:

$$g(h) = g(0) + g'(0)h + \frac{1}{2}g''(\theta h)h^2 \le 0 + 0 + \frac{1}{2}\cdot\frac{1}{4}\cdot h^2 = \frac{1}{8}h^2$$

将上述结果代入 (2.1) 式可知:

$$\mathrm{E}[e^{sX}] = \exp\{g(s(b-a))\} \le \exp\{\frac{1}{8}[s(b-a)]^2\}$$

命题得证.

Problem 3

设真实模型为 f，训练得到的模型为 $\hat{f}$
现有独立于训练集的数据 (x_0, y_0) (其中 X_0 为非随机的给定值)，设 $y_0 = f(x_0) + \varepsilon$ (其中 ε 为零均值的随机噪音)
试证明:

$$\mathrm{E}[(y_0 - \hat{f}(x_0))^2] = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\varepsilon)$$

Solution:

$$
\begin{aligned}
&\mathrm{E}[(y_0 - \hat{f}(x_0))^2] \\
&= \mathrm{E}[(y_0 - f(x_0) + f(x_0) - \mathrm{E}[\hat{f}(x_0)] + \mathrm{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] \\
&= \mathrm{E}[(\varepsilon + f(x_0) - \mathrm{E}[\hat{f}(x_0)] + \mathrm{E}[\hat{f}(x_0)] - \hat{f}(x_0))^2] \quad (\text{note that } \varepsilon \text{ is independent with } \hat{f} \text{ and } x_0) \\
&= \mathrm{E}[\varepsilon^2] + \{f(x_0) - \mathrm{E}[\hat{f}(x_0)]\}^2 + \mathrm{E}[(\hat{f}(x_0) - \mathrm{E}[\hat{f}(x_0)])^2] \\
&\qquad + 2\mathrm{E}[f(x_0) - \hat{f}(x_0)]\mathrm{E}[\varepsilon] + 2\{f(x_0) - \mathrm{E}[\hat{f}(x_0)]\}\{\mathrm{E}[\hat{f}(x_0)] - \mathrm{E}[\hat{f}(x_0)]\} \\
&= \mathrm{E}[\varepsilon^2] + \{f(x_0) - \mathrm{E}[\hat{f}(x_0)]\}^2 + \mathrm{E}[(\hat{f}(x_0) - \mathrm{E}[\hat{f}(x_0)])^2] + 0 + 0 \\
&= \mathrm{Var}(\varepsilon) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}[\hat{f}(x_0)]
\end{aligned}
$$

Problem 4

阅读以下材料:

> **(1) 向量求导**
> 给定 $x \in \mathbb{R}^n$ 和 $y \in \mathbb{R}^m$ ，考虑 $\frac{\partial y}{\partial x}$
>
> - 对于一般的情况，我们记:
>
> $$\frac{\partial y}{\partial x} := \left[\frac{\partial y_i}{\partial x_j}\right] = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_m}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{m\times n}$$
>
> $$\nabla_x y := \left(\frac{\partial y}{\partial x}\right)^{\mathrm{T}} = \left[\frac{\partial y_j}{\partial x_i}\right] = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_m}{\partial x_1} \\ \vdots & & \vdots \\ \frac{\partial y_1}{\partial x_n} & \cdots & \frac{\partial y_m}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n\times m}$$
>
> - 若 x 是一个标量 (即 $n=1$)，则我们有:
>
> $$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y_1}{\partial x} \\ \vdots \\ \frac{\partial y_m}{\partial x} \end{bmatrix} \in \mathbb{R}^{m\times 1}$$
>
> $$\nabla_x y = \left(\frac{\partial y}{\partial x}\right)^{\mathrm{T}} = \begin{bmatrix} \frac{\partial y_1}{\partial x} & \cdots & \frac{\partial y_m}{\partial x} \end{bmatrix} \in \mathbb{R}^{1\times m}$$
>
> - 若 y 是一个标量 (即 $m=1$)，则我们有:
>
> $$\frac{\partial y}{\partial x} = \begin{bmatrix} \frac{\partial y}{\partial x_1} & \cdots & \frac{\partial y}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1\times n}$$
>
> $$\nabla_x y = \left(\frac{\partial y}{\partial x}\right)^{\mathrm{T}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \vdots \\ \frac{\partial y}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{n\times 1}$$
>
> 若标量 y 关于 n 维向量 x 二阶可微，则我们记:
>
> $$\frac{\partial^2 y}{\partial x^2} := \frac{\partial}{\partial x}\nabla_x y = \left[\frac{\partial}{\partial x_j}\frac{\partial y}{\partial x_i}\right] = \left[\frac{\partial^2 y}{\partial x_j \partial x_i}\right] = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_1} \\ \vdots & & \vdots \\ \frac{\partial^2 y}{\partial x_1 \partial x_n} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n\times n}$$
>
> $$\nabla_x^2 y := \left(\frac{\partial^2 y}{\partial x^2}\right)^{\mathrm{T}} = \left[\frac{\partial^2 y}{\partial x_i \partial x_j}\right] = \begin{bmatrix} \frac{\partial^2 y}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2 y}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 y}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 y}{\partial x_n \partial x_n} \end{bmatrix} \in \mathbb{R}^{n\times n}$$
>
> ---
>
> **(链式法则)**
> 给定 $x \in \mathbb{R}^n, y \in \mathbb{R}^r, z \in \mathbb{R}^m$，则 $\frac{\partial z}{\partial x} \in \mathbb{R}^{m\times n}$, $\frac{\partial y}{\partial x} \in \mathbb{R}^{r\times n}$, $\frac{\partial z}{\partial y} \in \mathbb{R}^{m\times r}$ 满足:

$$\frac{\partial z}{\partial x} = \frac{\partial z}{\partial y}\frac{\partial y}{\partial x}$$

$$\nabla_x z = \left(\frac{\partial z}{\partial x}\right)^{\mathrm{T}} = \left(\frac{\partial z}{\partial y}\frac{\partial y}{\partial x}\right)^{\mathrm{T}} = \left(\frac{\partial y}{\partial x}\right)^{\mathrm{T}}\left(\frac{\partial z}{\partial y}\right)^{\mathrm{T}} = \nabla_x y \nabla_y z$$

(2) 矩阵求导

我们假设 X 没有特殊结构 (例如对称性和正定性等等) 以保证 X 的元素都是相互独立的.
这个基本假设可以表示为:

$$\frac{\partial X_{kl}}{\partial X_{ij}} = \delta_{ik}\delta_{lj} = \begin{cases} 1 & \text{if } i = k \text{ and } l = j \\ 0 & \text{otherwise} \end{cases}$$

其中 $\delta_{ij} := \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases}$ 代表 Kronecker δ-函数.

考虑矩阵 $X \in \mathbb{R}^{m\times n}$ 对标量 y 的求导:

$$\frac{\partial X}{\partial y} = \left[\frac{\partial X_{ij}}{\partial y}\right] = \begin{bmatrix} \frac{\partial X_{11}}{\partial y} & \cdots & \frac{\partial X_{1n}}{\partial y} \\ \vdots & & \vdots \\ \frac{\partial X_{m1}}{\partial y} & \cdots & \frac{\partial X_{mn}}{\partial y} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

考虑矩阵 $X \in \mathbb{R}^{m\times n}$ 对其自身元素 X_{ij} 的求导:

$$\frac{\partial X}{\partial X_{ij}} = E_{ij}$$

其中 $E_{ij} \in \mathbb{R}^{m\times n}$ 在 (i, j) 位置上为 1，在其余位置为零.

乘积求导法则:

$$\frac{\partial (XY)}{\partial \alpha} = \frac{\partial X}{\partial \alpha} Y + X \frac{\partial Y}{\partial \alpha}$$

其中 X, Y 是矩阵，α 是标量.

考虑标量 $y = f(X)$ 关于矩阵 $X \in \mathbb{R}^{m\times n}$ 的求导，我们记:

$$\frac{\partial y}{\partial X} := \left[\frac{\partial y}{\partial X_{ji}}\right] = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{m1}} \\ \vdots & & \vdots \\ \frac{\partial y}{\partial x_{1n}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{n\times m}$$

$$\nabla_X y := \left(\frac{\partial y}{\partial X}\right)^{\mathrm{T}} = \left[\frac{\partial y}{\partial X_{ij}}\right] = \begin{bmatrix} \frac{\partial y}{\partial x_{11}} & \cdots & \frac{\partial y}{\partial x_{1n}} \\ \vdots & & \vdots \\ \frac{\partial y}{\partial x_{m1}} & \cdots & \frac{\partial y}{\partial x_{mn}} \end{bmatrix} \in \mathbb{R}^{m\times n}$$

证明以下命题:
假设下面出现的 A, x, y, z 没有特殊结构，即其自身元素是相互独立的.

- ① 设 $y = Ax$，其中 $y \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n$ 且 A, x 相互独立，则 $\frac{\partial y}{\partial x} = A$
- ② 设标量 α 满足 $\alpha = y^{\mathrm{T}}Ax$，其中 $y \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n$ 且 A 独立于 x, y
则有 $\frac{\partial \alpha}{\partial x} = x^{\mathrm{T}}A^{\mathrm{T}}\frac{\partial y}{\partial x} + y^{\mathrm{T}}A$

- ③ 设标量 α 满足 $\alpha = x^{\mathrm{T}}Ax$，其中 $A \in \mathbb{R}^{n\times n}, x \in \mathbb{F}^n$ 且 A, x 相互独立，则有 $\frac{\partial\alpha}{\partial x} = x^{\mathrm{T}}(A + A^{\mathrm{T}})$
- ④ 设标量 α 满足 $\alpha = y^{\mathrm{T}}Ax$，其中 $y \in \mathbb{R}^m, A \in \mathbb{R}^{m\times n}, x \in \mathbb{R}^n$
若 x, y 都是 $z \in \mathbb{R}^q$ 的函数，且 A 独立于 z，则有 $\frac{\partial\alpha}{\partial z} = x^{\mathrm{T}}A^{\mathrm{T}}\frac{\partial y}{\partial z} + y^{\mathrm{T}}A\frac{\partial x}{\partial z}$
- ⑤ 设 $A \in \mathbb{R}^{n\times n}$ 非奇异且元素是标量 α 的函数，则有 $\frac{\partial A^{-1}}{\partial\alpha} = -A^{-1}\frac{\partial A}{\partial\alpha}A^{-1}$

Solution:

- **命题 ① 的证明:**

$$\left[\frac{\partial y}{\partial x}\right]_{ij} = \frac{\partial y_i}{\partial x_j} = \frac{\partial}{\partial x_j}\sum_{k=1}^{n} a_{ik}x_k = a_{ij} \ \ (\forall\, i, j)$$
$$\Leftrightarrow$$
$$\frac{\partial y}{\partial x} = A$$

- **命题 ② 的证明:**

$$\left[\frac{\partial\alpha}{\partial x}\right]_i = \frac{\partial\alpha}{\partial x_i} = \frac{\partial}{\partial x_i}\sum_{j=1}^{m}\left\{y_j\sum_{k=1}^{n}a_{jk}x_k\right\} = \sum_{j=1}^{m}y_j a_{ji} + \sum_{j=1}^{m}(Ax)_j\frac{\partial y_j}{\partial x_i} = \left[y^{\mathrm{T}}A + x^{\mathrm{T}}A^{\mathrm{T}}\frac{\partial y}{\partial x}\right]_i$$
$$\Leftrightarrow$$
$$\frac{\partial\alpha}{\partial x} = y^{\mathrm{T}}A + x^{\mathrm{T}}A^{\mathrm{T}}\frac{\partial y}{\partial x}$$

简便很多的方法: (原题的背景知识只给出了导数的定义，这让我不敢用导数的相关公式)

$$\begin{aligned}\frac{\partial\alpha}{\partial x} &= \frac{\partial}{\partial x}y^{\mathrm{T}}Ax \\ &= y^{\mathrm{T}}\left(\frac{\partial}{\partial x}Ax\right) + (Ax)^{\mathrm{T}}\left(\frac{\partial y}{\partial x}\right) \\ &= y^{\mathrm{T}}A + x^{\mathrm{T}}A^{\mathrm{T}}\frac{\partial y}{\partial x}\end{aligned}$$

- **命题 ③ 的证明:**
直接应用命题 ② 的结论，我们有

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

另法:

$$\begin{aligned}\frac{\partial\alpha}{\partial x} &= \frac{\partial}{\partial x}x^{\mathrm{T}}Ax \\ &= x^{\mathrm{T}}\left(\frac{\partial}{\partial x}Ax\right) + (Ax)^{\mathrm{T}}\left(\frac{\partial x}{\partial x}\right) \\ &= x^{\mathrm{T}}A + x^{\mathrm{T}}A^{\mathrm{T}}I_n \\ &= x^{\mathrm{T}}(A + A^{\mathrm{T}})\end{aligned}$$

- **命题 ④ 的证明:**
应用命题 ② 的结论和求导的链式法则可知:

$$
\begin{aligned}
\frac{\partial \alpha}{\partial z} &= \frac{\partial \alpha}{\partial x}\frac{\partial x}{\partial z} \\
&= \left(y^{\mathrm{T}}A + x^{\mathrm{T}}A^{\mathrm{T}}\frac{\partial y}{\partial x}\right)\frac{\partial x}{\partial z} \\
&= y^{\mathrm{T}}A\frac{\partial x}{\partial z} + x^{\mathrm{T}}A^{\mathrm{T}}\frac{\partial y}{\partial x}
\end{aligned}
$$

- **命题 ⑤ 的证明:**

$$
\begin{gathered}
AA^{-1} = I_n \\
\Leftrightarrow \\
\frac{\partial}{\partial \alpha}(AA^{-1}) = \frac{\partial A}{\partial \alpha}A^{-1} + A\frac{\partial A^{-1}}{\partial \alpha} = \frac{\partial I_n}{\partial \alpha} = 0_{n\times n} \\
\Leftrightarrow \\
\frac{\partial A^{-1}}{\partial \alpha} = -A^{-1}\frac{\partial A}{\partial \alpha}A^{-1}
\end{gathered}
$$

Problem 5

设 $X \in \mathbb{R}^{m\times n}, a \in \mathbb{R}^n, y \in \mathbb{R}^m$ 且 $X^{\mathrm{T}}X$ 非奇异
试推导如下最小二乘问题的解 $\hat{x}$ 的解析形式:

$$
\min_x \|y - Xa\|_2
$$

Solution:
上述问题等价于 $\min_x \|y - Xa\|_2^2$

$$
\begin{aligned}
\nabla_a \|y - Xa\|_2^2 &= \nabla_a (y - Xa)^{\mathrm{T}}(y - Xa) = 2X^{\mathrm{T}}Xa - 2X^{\mathrm{T}}y \\
\nabla_a^2 \|y - Xa\|_2^2 &= 2X^{\mathrm{T}}X \succ 0 \ (\text{note that } X^{\mathrm{T}}X \text{ is non-singular})
\end{aligned}
$$

注意到目标函数是严格凸的，且在定义域上一阶连续可微
因此最小二乘解 $\hat{a}$ 即目标函数的驻点 (即使得梯度为零的点)
令 $\nabla_a \|y - Xa\|_2^2 = 2X^{\mathrm{T}}Xa - 2X^{\mathrm{T}}y = 0_n$ 可得 $\hat{a} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y$

The End