1. 证明《统计学习方法》习题 1.2：

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。

解：设条件概率分布的参数为θ，那么：

经验风险为

$$R_{emp}(\theta) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, P(y_i|x_i;\theta))$$

$$= -\frac{1}{N}\sum_{i=1}^{N} \log P(y_i|x_i;\theta) \quad \cdots (1)$$

而 log-likelihood function 为

$$l(\theta) = l(\theta; X, Y) = \sum_{i=1}^{N} \log P(y_i|x_i;\theta) \quad \cdots (2)$$

由(1)和(2)可知，

$$\theta = \arg\min_{\theta} R_{emp}(\theta) = \arg\max_{\theta} l(\theta)$$

即 $\min R_{emp}(\theta) \Leftrightarrow \max l(\theta)$ □

2. 请证明下述 Hoeffding 引理：

Lemma 1. *Let X be a random variable with $E(X) = 0$ and $P(X \in [a, b]) = 1$. Then it holds*

$$E\{\exp(sX)\} \leq \exp\{s^2(b-a)^2/8\}.$$

证：$\frac{de^{sX}}{dX} = se^{sX}$ $\quad \frac{d^2e^{sX}}{dX^2} = s^2e^{sX} > 0$

$\Rightarrow$ e^{sx} 关于 x 是凸函数

由 Jensen 不等式，当 $x \in [a,b]$ 时

$$e^{sx} \le \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}$$

两边取期望，

$$E(e^{sx}) \le E\left(\frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa} \right)$$

利用 $E(X)=0$ 可知

$$E\left(\frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa} \right) = \frac{-a}{b-a} e^{sb} + \frac{b}{b-a} e^{sa}$$

下面说明 $\frac{-a}{b-a} e^{sb} + \frac{b}{b-a} e^{sa} \le \exp\left\{ \frac{s^2(b-a)^2}{8} \right\}$,

由于 $E(X)=0$ 且 $a \le x \le b$

若 $a=0$，必有 $P(X=a)=1$

若 $b=0$，必有 $P(X=b)=1$

这两种情况下，原不等式显然成立

故只考虑 $a<0<b$ 的情况

令 $\lambda = \frac{-a}{b-a} = \frac{0-a}{b-a} \in (0,1)$

则 $1-\lambda = \frac{b}{b-a} = \frac{b-0}{b-a} \in (0,1)$

$$\Rightarrow \lambda e^{sb} + (1-\lambda) e^{sa}$$

$$\lambda = \frac{-a}{b-a}$$

$$a = -\lambda \cdot b-a$$

$$= e^{sa} \left(1-\lambda + \lambda e^{s(b-a)} \right)$$

$$= e^{-s\lambda(b-a)} \left(1-\lambda + \lambda e^{s(b-a)} \right)$$

再令 $u = s(b-a) \in \mathbb{R}$

$$= e^{-\lambda u} \left(1-\lambda + \lambda e^{u} \right)$$

取对数，并记：

$$g(u) = -\lambda u + \ln\left(1-\lambda + \lambda e^{u} \right)$$

$$g'(u) = -\lambda + \frac{\lambda e^{u}}{1-\lambda+\lambda e^{u}} \Rightarrow g'(0) = 0$$

$$g''(u) = \frac{\lambda e^{u}(1-\lambda+\lambda e^{u}) - \lambda e^{u}(\lambda e^{u})}{(1-\lambda+\lambda e^{u})^2}$$

$$= \frac{\lambda e^{u}(1-\lambda+\lambda e^{u} - \lambda e^{u})}{(1-\lambda+\lambda e^{u})^2}$$

$$= \frac{\lambda e^{u}}{1-\lambda+\lambda e^{u}} \cdot \left(\frac{1-\lambda}{1-\lambda+\lambda e^{u}} \right) \leq \frac{1}{4}$$

$\left(z \cdot (1-z) \leq \frac{1}{4}, \ z \in (0,1) \right)$

由于 $g(\mu) = g(0) + g'(0)\cdot\mu + \frac{1}{2}g'(\xi)\mu^2$

$$\leq 0 + 0\cdot\mu + \frac{1}{2}\cdot\frac{1}{4}\cdot\mu^2$$

$$= \frac{1}{8}\mu^2$$

将 $\mu = s(b-a)$ 代入，得到：

$$g(\mu) \leq \frac{1}{8}s^2(b-a)^2$$

即：$\frac{-a}{b-a}e^{sb} + \frac{b}{b-a}e^{sa} \leq \exp\left\{\frac{s^2(b-a)^2}{8}\right\}$ □

3. 请列举一个实际中有监督学习的应用，请说明（1）问题背景、（2）因变量和自变量分别是什么，以及（3）通过机器学习建模如何解决该实际问题。

解：(1) 在现代医学中，需要对病人的病灶部位拍片检查，但检查结果往往需富有经验的医生通过仔细观察影像才能看出。在缺乏医生的地区，可通过有监督的机器学习所得模型辅助诊断。

(2) 自变量：拍片所得影像的像素点矩阵 M

因变量：病灶的坐标 (x,y) 以及病灶的类型 ω_i

(如炎症、积液、良性肿瘤、恶性肿瘤……)

(3) 将已经由专业医生标出病灶坐标与病灶类型的影片

作为数据集，建立学习模型为：

$$f(M_i) = (x_i, y_i, w_i)$$

其中 M_i 是输入的像素矩阵，x_i, y_i 是判断病灶的坐标，w_i 是病灶类型

再设定损失函数为：

$$L_i = \lambda L_1(x_i, y_i) + \mu L_2(w_i) + \gamma L_3$$

其中：$L_1(x_i, y_i) = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2}$

x_i, y_i 是机器判断的坐标，x_0, y_0 是真实坐标。

$$L_2(w_i) = \begin{cases} 1 & w_i \text{ 分类正确} \\ 0 & w_i \text{ 分类错误} \end{cases}$$

其中 w_i 是机器给出的判断

L_3 为度量模型复杂度的某种范数

λ, μ, γ 是用于调节模型的参数

经验风险为：

$$R_{emp} = \frac{1}{N}\sum_{i=1}^{N} L_i(x_i, y_i, w_i)$$

使 R_{emp} 在训练集上与测试集上都尽可能小

则可以用训练好的模型 $f(\cdot)$ 去判断未标定的数据 M_i 以帮助医生快速诊断。 $\square$

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

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

证: $y_i = \sum_{k=1}^{n} A_{ik} x_k$

$$\frac{\partial y_i}{\partial x_k} = A_{ik} \Rightarrow \left(\frac{\partial y}{\partial x^T}\right)_{ik} = A_{ik}$$

$$\Rightarrow \frac{\partial y}{\partial x^T} = A$$

(b) Let the scalar α be defined by $\alpha = \mathbf{y}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{y}$ is m×1, $\mathbf{x}$ is n×1, $\mathbf{A}$ is m×n, and $\mathbf{A}$ is independent of $\mathbf{x}$ and $\mathbf{y}$, then $\frac{\partial \alpha}{\partial \mathbf{x}^\top} = \mathbf{x}^\top \mathbf{A}^\top \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}^\top}\right) + \mathbf{y}^\top \mathbf{A}.$

$(y^T A x)'$
$= y^T A + x^T A^T \frac{\partial y}{\partial x^T}$

证: $\frac{\partial \alpha}{\partial x^T} = \left[\frac{\partial \alpha}{\partial x_1}, \cdots, \frac{\partial \alpha}{\partial x_n}\right]$

$$\alpha = \sum_{i=1}^{m} \sum_{j=1}^{n} y_i A_{ij} x_j$$

$$\frac{\partial \alpha}{\partial x_k} = \sum_{i=1}^{m} y_i A_{ik} + \sum_{i=1}^{m} \sum_{j=1}^{n} \frac{\partial y_i}{\partial x_k} A_{ij} x_j$$

记 $\left(\frac{\partial y}{\partial x^T}\right)_k = \left[\frac{\partial y_1}{\partial x_k}, \frac{\partial y_2}{\partial x_k}, \cdots, \frac{\partial y_m}{\partial x_k}\right]^T$ (第k列)

则 $\frac{\partial \alpha}{\partial x_k} = y^T \cdot A_k + \left(\frac{\partial y}{\partial x^T}\right)_k^T \cdot A \cdot x$

$$= y^T A_k + x^T A^T \left(\frac{\partial y}{\partial x^T}\right)_k$$

$$\Rightarrow \frac{\partial \alpha}{\partial x^T} = y^T A + x^T A^T \left(\frac{\partial y}{\partial x^T}\right)$$

(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 $\mathbf{x}$ is n×1, $\mathbf{A}$ is n×n, and $\mathbf{A}$ does not depend on $\mathbf{x}$, then

$$\frac{\partial \alpha}{\partial \mathbf{x}^\top} = \mathbf{x}^\top(\mathbf{A} + \mathbf{A}^\top).$$

证：由(b)可知：

$$\frac{\partial \alpha}{\partial x^T} = x^T A + x^T A^T \left(\frac{\partial x}{\partial x^T}\right)$$

$$= x^T A + x^T A^T \cdot I_n$$

$$= x^T (A + A^T)$$

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

$$\frac{\partial \alpha}{\partial \mathbf{z}^\top} = \mathbf{x}^\top \mathbf{A}^\top \left(\frac{\partial \mathbf{y}}{\partial \mathbf{z}^\top}\right) + \mathbf{y}^\top \mathbf{A} \left(\frac{\partial \mathbf{x}}{\partial \mathbf{z}^\top}\right).$$

证：$$\frac{\partial \alpha}{\partial z^T} = \left[\frac{\partial \alpha}{\partial z_1}, \cdots, \frac{\partial \alpha}{\partial z_q}\right]$$

$$\alpha = \sum_{i=1}^{m} \sum_{j=1}^{n} y_i A_{ij} x_j$$

$$\frac{\partial \alpha}{\partial z_k} = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{\partial y_i}{\partial z_k} A_{ij} x_j + y_i A_{ij} \frac{\partial x_j}{\partial z_k}\right)$$

记 $\left(\frac{\partial y}{\partial z^T}\right)_k = \left(\frac{\partial y_1}{\partial z_k}, \cdots, \frac{\partial y_m}{\partial z_k}\right)^T$ (第k列)

$\left(\frac{\partial x}{\partial z^T}\right)_k = \left(\frac{\partial x_1}{\partial z_k}, \cdots, \frac{\partial x_n}{\partial z_k}\right)^T$ (第k列)

则有 $$\frac{\partial \alpha}{\partial z_k} = \left(\frac{\partial y}{\partial z^T}\right)_k^T A x + y^T A \left(\frac{\partial x}{\partial z^T}\right)_k$$

$$\Rightarrow \frac{\partial \alpha}{\partial z^T} = x^T A \frac{\partial y}{\partial z^T} + y^T A \frac{\partial x}{\partial z^T}$$

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

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

证：$I = AA^{-1}$

$$\frac{\partial I}{\partial \alpha} = 0$$

$$\frac{\partial AA^{-1}}{\partial \alpha} = \frac{\partial A}{\partial \alpha} \cdot A^{-1} + A \cdot \frac{\partial A^{-1}}{\partial \alpha}$$

$$\Rightarrow \frac{\partial A}{\partial \alpha} \cdot A^{-1} + A \cdot \frac{\partial A^{-1}}{\partial \alpha} = 0$$

$$\Rightarrow \frac{\partial A^{-1}}{\partial \alpha} = -A^{-1}\frac{\partial A}{\partial \alpha}A^{-1} \qquad \square$$

5. Please write $\hat{\mathbf{a}}$ as the solution of the minimization problem:

$$\min_{\mathbf{a}} \|\mathbf{X}\mathbf{a} - \mathbf{y}\|_2 ,$$

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

解：$a = \arg\min\limits_{a} \|Xa - y\|_2 = \arg\min\limits_{a} \frac{1}{2}\|Xa - y\|_2^2$

设 $f(a) = \frac{1}{2}\|Xa - y\|_2^2 = \frac{1}{2}(Xa-y)^T(Xa-y)$

$$= \frac{1}{2}(a^T X^T X a - 2y^T X a + y^T y)$$

$$\frac{\partial f(a)}{\partial a^T} = a^T X^T X - y^T X$$

令 $\frac{\partial f(a)}{\partial a^T}=0 \Rightarrow a^T X^T X - y^T X = 0$

由于 $X^T X$ 可逆，则 $a=(X^T X)^{-1} X^T y$ $\square$