1

1. 推导线性回归模型的贝叶斯估计。给定训练数据集 $\{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$，其中 $\mathbf{x}_i \in R^m, y_i \in R$。考虑线性回归模型：$y = \mathbf{w}^T\mathbf{x} + \varepsilon$，其中，假设 ε 服从零均值、方差为 σ_0^2 的高斯分布。若已知 $\mathbf{w}$ 服从零均值、协方差矩阵为 Λ 的高斯分布，其中 Λ 是对角矩阵，对角元素为 $\{\lambda_{jj}\}_{j=1}^{m}$。试基于最大后验概率准则估计模型参数，给出详细推导步骤。

2. 在题目 2 中，若 $\mathbf{w}$ 采用共轭先验，即假设 $\mathbf{w}$ 服从零均值、协方差矩阵为 Λ 是的高斯分布，请基于完全贝叶斯准则推导回归模型，给出预测性分布密度函数和每个参数的推导步骤。（提示：回归函数的分布密度；后验分布密度；预测性分布密度函数；“配方法”）

1

3. 对于包含 K 个类别的分类问题，分析比较一下线性判别函数模型、概率生成模型和概率判别模型。提示，可以从以下几个方面比较：模型的输入和输出、模型形式是否是线性、目标函数、以及学习算法（求解参数的方法）等。

模型名称	线性判别函数模型	概率生成模型	概率判别模型
典型例子	感知机	朴素贝叶斯	Logistic 回归
模型的输入	x	X	X
模型的输出	f(x)=sing(w·x+b)	y = arg maxP(Y= c_k)∏P(X|Y)	$\hat{w} = argmax_w J(w) = argmax_w \sum_{i=1}^{N}(y_i \log p_i + (1-y_i)\log p_0)$
模型形式是否是线性	是	否	否
目标函数	[illegible] $L(w) = \sum_{x_i \in D_{wrong}} -y_i w^T x_i$	后验概率	似然函数
学习算法（求解参数的方法）	梯度下降	后验概率最大化	极大似然估计

已知： Data : $\{(x_i, y_i)\}_{i=1}^{N}$

$x_i \in \mathbb{R}^p$ · $y_i \in \mathbb{R}$,

$$X = (x_1\ x_2\ \cdots\ x_N)^T = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & & \\ x_{N1} & \cdots & x_{Np} \end{pmatrix}_{N \times p}$$

$Y = (y_1\ y_2 \cdots y_N)^T$ $N\times1$ 维.

Model: $\begin{cases} f(x) = w^T x \\ y = f(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \end{cases}$

Bayesian Method: ① Inference : posterior(w).

② Prediction: $x^* \to y^*$

① Inference. $P(w|Data) = P(w|X,Y) = \dfrac{P(w, Y|X)}{P(Y|X)} = \dfrac{\overbrace{P(Y|w,X)}^{\text{likelihood}}\ \overbrace{P(w)}^{\text{prior}}}{\int P(Y|w,X) \cdot P(w)\,dw}$

$= P(w|X) P(Y|X)$

$P(w, Y|X)$

$= P(Y|w,X) \cdot \dfrac{P(w|X)}{P(w)}$

$$P(Y|w,X) = \prod_{i=1}^{N} P(y_i|w; x_i) = \prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2).$$

$P(w) = N(0, \Sigma_p)$

知识点 ↘

(共轭：Gaussian 分布是自共轭的)

$$\underset{\text{Gaussian.}}{P(w|Data)} \propto \underset{\text{Gaussian}}{P(Y|w,X)}\ \underset{\text{Gaussian}}{P(w)} \propto \prod_{i=1}^{N} N(y_i|w^T x_i, \sigma^2) \cdot N(0, \Sigma_p) \xrightarrow{\text{总可以得到:}} N(\mu_w, \Sigma_w)$$

后验也一定是高斯的.

$$P(Y|w,X) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{-\frac{1}{2\sigma^2}(y_i - w^T x_i)^2\right\}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^{N} (y_i - w^T x_i)^2\right\}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}} \sigma^N} \exp\left\{-\frac{1}{2}(Y - Xw)^T \sigma^{-2} I\, (Y - Xw)\right\}$$

$$= N(Xw, \sigma^{-2} I)$$

这一步的具体步骤：

$$\sum_{i=1}^{N} (y_i - w^T x_i)^2 = (y_1 - w^T x_1\ \ y_2 - w^T x_2\ \cdots\ y_N - w^T x_N) \begin{pmatrix} y_1 - w^T x_1 \\ y_2 - w^T x_2 \\ \vdots \\ y_N - w^T x_N \end{pmatrix}$$

↓

$(y_1\ y_2 \cdots y_N) - w^T(x_1\ x_2 \cdots x_N)$

$= (Y^T - w^T X^T)(Y - Xw)$

$= (Y - Xw)^T (Y - Xw)$

$$P(w|Data) \propto N(Xw, \sigma^{-2} I)\, N(0, \Sigma_p)$$

$$\propto \exp\left\{-\frac{1}{2}(Y - Xw)^T \sigma^{-2} I\,(Y - Xw)\right\} \cdot \exp\left\{-\frac{1}{2} w^T \Sigma_p^{-1} w\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(Y^T - w^T X^T)(Y - Xw) - \frac{1}{2} w^T \Sigma_p^{-1} w\right\}$$

$$= \exp\left\{-\frac{1}{2\sigma^2}(Y^T Y - 2Y^T X w - \cancel{w^T X^T Y} + w^T X^T X w) - \frac{1}{2} w^T \Sigma_p^{-1} w\right\}$$

这两项互为转置且为实数.

上式中的二次项： $-\dfrac{1}{2\sigma^2} w^T X^T X w - \dfrac{1}{2} w^T \Sigma_p^{-1} w = \dfrac{1}{2} w^T \underbrace{(\sigma^{-2} X^T X + \Sigma_p^{-1})}_{\Sigma_w^{-1} \to \text{记为} A} w$

上式中的一次项： $-\dfrac{1}{2\sigma^2}(-2) Y^T X w = \underbrace{\sigma^{-2} Y^T X}\, w$

$\mu_w^T \Sigma_w^{-1} = \mu_w^T A$.

则 $\mu_w = (\sigma^{-2} Y^T X)^T A^{-1} = \sigma^{-2} A^{-1} X^T Y$

综上，$P(w|Data) = N(\mu_w, \Sigma_w)$，其中，$\mu_w = \sigma^{-2} A^{-1} X^T Y$

$\Sigma_w = A^{-1}$ $\quad (A = \sigma^{-2} X^T X + \Sigma_p^{-1})$

综上，$P(w|Data) = N(\mu_w, \Sigma_w)$，其中，$\mu_w = \sigma^{-2} A^{-1} X^T Y$

$\Sigma_w = A^{-1}$ $\quad (A = \sigma^{-2} X^T X + \Sigma_p^{-1})$

② Prediction

Model : $\begin{cases} f(x) = W^T x = x^T w \\ y = f(x) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2) \end{cases}$

给出样本 x^*，则 $f(x^*) = x^{*T} w$

$\because w \sim N(\mu_w, \Sigma_w) \quad \therefore f(x^*) \sim N(x^{*T}\mu_w, x^{*T}\Sigma_w x^*)$

$\therefore y^* = f(x^*) + \varepsilon. \sim N(x^{*T}\mu_w, x^{*T}\Sigma_w x^* + \sigma^2)$