Project Report of Machine Learning (Assignment 4)

Author: 赵文浩 23020201153860 (计算机科学系)

■ Task Description

试推导 Bayesian Linear Regression 中 y_{new} 的概率分布。

P.S. 本次任务涉及较多符号计算，因此这里采用纸质的形式展示推导过程。

23020201153860 赵文浩
计算机科学系

Assignment 4 — Bayesian Linear Regression

I. Review of Bayesian

○ 贝叶斯公式基本形式：$\underbrace{P(A|B)}_{\text{后验 posterior}} = \underbrace{P(A)}_{\text{先验 prior}} \cdot \frac{\overbrace{P(B|A)}^{\text{似然函数估计 likelihood}}}{P(B)}$

○ Bayesian Method 两步骤

Step ① Inference — 推理求后验：posterior (w)

Step ② Prediction — 根据 Inference 预测：$x^* \rightarrow y^*$

II. Linear Regression

这里对线性回归进行定义说明，以便推导。

给定数据：Data $\Rightarrow D = \{\langle x_1, y_1\rangle, \langle x_2, y_2\rangle \cdots \langle x_n, y_n\rangle\}$ $i=1,n$, $x_i \in R^d$, $y_i \in R$

$$\underset{n\times d}{X} = (\underbrace{x_1, x_2, \cdots, x_n}_{x_i:\, d\times 1})^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & & x_{nd} \end{pmatrix}$$

$$\underset{n\times 1}{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad w = \begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix} \; w_{d\times 1}$$

模型：Model $\rightarrow \underbrace{y_i}_{\text{标量 } y_i \in R} = \underbrace{\overbrace{w}^{\text{未知，待求解参数}}{}^T x_i}_{\text{确定性}} + \underbrace{\varepsilon}_{\text{偶然性(噪音)}}$，其中 $\varepsilon \sim N(0, \beta^{-1})$

$\beta \uparrow \beta^{-1} \downarrow$ 准确性 $\uparrow$

$\beta \downarrow \beta^{-1} \uparrow$ 准确性 $\downarrow$

☆假设 w 也服从 Gaussian 分布

且 $p(w) = N(0, \alpha^{-1} I)$　　易知 $y_i \sim N(w^T x_i, \beta^{-1})$

III Linear Regression with Bayesian

3.1 Inference — Likelihood

Inference 的目的是求参数 w 的后验分布 $p(w|D)$

由 Section I 可知：$\underbrace{p(w|D)}_{\text{后验}} \propto p(w) \cdot p(D|w) = \underbrace{p(Y|w,X)}_{\text{likelihood}} \underbrace{p(w)}_{\text{prior}}$ Ⅲ

y_i 独立同分布

$$= \prod_{i=1}^{n} \cancel{p(y_i)}\, N(w^T x_i, \beta^{-1}) \cdot N(0, \alpha^{-1} I)$$

其中 likelihood：$\underbrace{p(Y|w,X)}_{\text{Log}} = \underbrace{\log\Big[\prod_{i=1}^{n} p(y_i|w,x_i)}_{\text{log}} = \prod_{i=1}^{n} N(w^T x_i, \beta^{-1})\Big]$

$$\propto \sum_{i=1}^{n} \frac{\beta^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} \cdot \exp\left\{-\frac{\beta}{2}(y_i - w^T x_i)^2\right\}$$

$$\propto \exp\left\{-\frac{\beta}{2}\sum_{i=1}^{n}(y_i - w^T x_i)^2\right\}$$

$$= \exp\left\{-\frac{\beta}{2}(Y - \underbrace{X}_{n\times d}\underbrace{w}_{d\times 1})^T (Y - Xw)\right\}$$

年　月　日　第　页　1

类似地，prior：$p(w) = N(0, \alpha^{-1} I)$

$$\propto \exp\{-\frac{\alpha}{2} w^T w\}$$

3.2 Inference — posterior

后验 posterior：$p(w|D) \propto p(y|w, x) \cdot p(w)$

$$\propto \exp\{-\frac{\beta}{2}(y - xw)^T(y - xw)\} \cdot \exp\{-\frac{\alpha}{2} w^T w\}$$

$$= \exp\{-\frac{\beta}{2}(y - xw)^T(y - xw) - \frac{\alpha}{2} w^T w\}$$

$$= \exp\{-\frac{\beta}{2}(y^T y - 2y^T x w + w^T x^T x w) - \frac{\alpha}{2} w^T w\}$$

$$= \exp\{\underbrace{-\frac{\beta}{2} w^T x^T x w - \frac{\alpha}{2} w^T w}_{\text{二次项}} + \underbrace{\beta y^T x w}_{\text{一次项}} - \frac{\beta}{2} y^T y\} \quad \text{①}$$

$\beta w^T y^T x$

推理得到 $p(w|D)$ 是关于 w 的 Gaussian 分布，不妨假设：

$$\boxed{p(w|D) = N(w|u_n, \Sigma_n)}$$

u_n, Σ_n ↳代表n个元素

展开 $p(w|D) = N(u_n, \Sigma_n)$

$$\propto \exp\{-\frac{1}{2}(w - u_n)^T \Sigma_n^{-1}(w - u_n)\}$$

$$= \exp\{-\frac{1}{2}(w^T \Sigma_n^{-1} w - 2w^T \Sigma_n^{-1} u_n + u_n^T \Sigma_n^{-1} u_n)\} \quad \text{②}$$

对比①、②两式，可得：$\begin{cases} \Sigma_n^{-1} = \beta X^T X + \alpha I \\ \Sigma_n^{-1} u_n = \beta X^T y \end{cases}$ 即：$\begin{cases} u_n = \beta \Sigma_n X^T y \\ \Sigma_n = (\beta X^T X + \alpha I)^{-1} \end{cases}$

☆ ⇒ 考虑 $p(w|D)$ 高斯分布 极值点，即：$\hat{w} = u_n$

$$= \beta \Sigma_n X^T y$$

$$= \beta(\beta X^T X + \alpha I)^{-1} X^T y$$

当 $\alpha = \lambda$，且 $\beta = 1$ 时，$w^* = (X^T X + \lambda I)^{-1} X^T y$，与 Ridge 回归（L2 正则）的结果相同

3.3 Prediction

预测 Prediction 的目的是从新数据样本 x^*，结合给定的数据集 D，预测 y^*，即预测分布：

$$p(y|x, D) = \int p(y, w|x, D) dw$$

$x \to x^{new}$ ；w 为向量

$*$ w 与 x^{new} 无关；y^{new} 与 D 无关

$$= \int p(w|x, D) \cdot p(y|x, w, D) dw$$

$$= \int \underbrace{p(w|D)}_{N(u_n, \Sigma_n)} \cdot \underbrace{p(y|x, w)}_{N(w^T x, \beta^{-1})} dw$$

x ↳列向量 $(d \times 1)$

↳见下页.

上接 3.3 Prediction

*注意：这里的 x 为列向量 $(d\times 1)$

$$p(y|x,D)=\int p(w|D)\cdot p(y|x,w)\,dw$$

$p(w|D) \to N(u_n,\Sigma_n)$ 3.2已求解；$p(y|x,w)\to N(w^Tx,\beta^{-1})$，$x$ ↳ $d\times1$ 列向量

$$=\int N(u_n,\Sigma_n)\,N(w^Tx,\beta^{-1})\,dw$$

$$\propto \exp\{-\tfrac{1}{2}(w-u_n)^T\Sigma_n^{-1}(w-u_n)\}\cdot\exp\{-\tfrac{\beta}{2}(y-w^Tx)^2\}$$

w^Tx：$1\times d,\ d\times1$ → 标量，可写成平方形式

$$=\exp\{-\tfrac{1}{2}(\underbrace{w^T\Sigma_n^{-1}w-2w^T\Sigma_n^{-1}u_n+\boxed{u_n^Tu_n}_{const}}_{part1}\underbrace{+\beta y^2-2\beta yw^Tx+\beta w^Txx^Tw}_{part2\ 标量})\}$$

整体记为 $\#$。

先只考虑 $\#$ 部分：

$$\#\propto w^T\Sigma_n^{-1}w-2w^T\Sigma_n^{-1}u_n+\beta y^2-2\beta yw^Tx+\beta w^Txx^Tw$$

$$=w^T(\beta xx^T+\Sigma_n^{-1})w-2w^T(\beta yx+\Sigma_n^{-1}u_n)+\beta y^2 \quad ③$$

↓配方

ASSUME ⟹ 假设 $\#=(w-m)^TL^{-1}(w-m)+\beta y^2-\boxed{?}$

$$=\boxed{w^TL^{-1}w-2w^TL^{-1}m+m^TL^{-1}m}+\beta y^2-m^TL^{-1}m \quad ④$$

$w\sim N(m,L)$

对比③、④，可得：
$$\begin{cases}L^{-1}=\beta xx^T+\Sigma_n^{-1}\\ L^{-1}m=\beta yx+\Sigma_n^{-1}u_n\end{cases}$$，即：
$$\begin{cases}m=L(\beta yx+\Sigma_n^{-1}u_n)\\ L=(\beta xx^T+\Sigma_n^{-1})^{-1}\end{cases}$$

因此代入 $p(y|x,D)$ 可得：

$$p(y|x,D)\propto\exp\{-\tfrac{1}{2}[\underbrace{(w-m)^TL^{-1}(w-m)}_{N(m,L)}+\beta y^2-m^TL^{-1}m]\}$$

$$\propto\exp\{-\tfrac{1}{2}(\beta y^2-m^TL^{-1}m)\}$$

$$=\exp\{-\tfrac{1}{2}[\beta y^2-(\beta yx+\Sigma_n^{-1}u_n)^T\underbrace{L^TL^{-1}}_{I}L(\beta yx+\Sigma_n^{-1}u_n)]\}$$

↓配方

ASSUME ⟹ 假设 $y\sim N(u,\lambda^{-1})$

$$=\exp\{-\tfrac{1}{2}[\beta y^2-(\beta yx+\Sigma_n^{-1}u_n)^TL(\beta yx+\Sigma_n^{-1}u_n)]\}$$

$$\propto\exp\{-\tfrac{1}{2}[\underbrace{(\beta-\beta^2x^TLx)}_{\triangle}y^2-2\underbrace{\beta yx^TL\Sigma_n^{-1}u_n}_{☆}]\}$$

有：$p(y|x,D)\propto\exp\{-\tfrac{1}{2}\lambda(y-u)^2\}$
$$=\exp\{-\tfrac{1}{2}(\underbrace{\lambda}_{\triangle}y^2-\underbrace{2\lambda uy}_{☆}+u^2\lambda)\}$$

因此
$$\begin{cases}\lambda=\beta(1-\beta x^TLx)\\ u=\lambda^{-1}\beta x^TL\Sigma_n^{-1}u_n\end{cases}$$
β、L、Σ_n^{-1} 均已知。

↳ 见下页

年 月 日 第 页 3

↑上接

*note

$$\left\{\begin{array}{l}\lambda=\beta(1-\beta X^TLX) \\ u=\lambda^{-1}\beta X^TL\Sigma_n^{-1}u_n\end{array}\right.$$

其中 $L=(\underbrace{\beta X}_{u}\underbrace{X^T}_{v}+\underbrace{\Sigma_n^{-1}}_{A})^{-1}$ ⇙ $(A+uv^T)^{-1}=A^{-1}-\frac{A^{-1}uv^TA^{-1}}{1+v^TA^{-1}u}$

$$=\Sigma_n-\frac{\Sigma_n\beta XX^T\Sigma_n}{1+X^T\Sigma_n\beta X}=\Sigma_n-\frac{\beta\Sigma_nXX^T\Sigma_n}{1+\beta X^T\Sigma_nX}$$

因此将L代入 X^TLX 中可得:

$$X^TLX=X^T\Sigma_nX-\frac{\beta X^T\Sigma_nXX^T\Sigma_nX}{1+\beta X^T\Sigma_nX}$$

$$=\frac{X^T\Sigma_nX+\boxed{\beta X^T\Sigma_nX\cdot X^T\Sigma_nX-\beta X^T\Sigma_nXX^T\Sigma_nX}}{1+\beta X^T\Sigma_nX}=\frac{X^T\Sigma_nX}{1+\beta X^T\Sigma_nX}$$

$$\therefore\lambda=\beta(1-\beta X^TLX)$$

$$=\beta\left(1-\beta\frac{X^T\Sigma_nX}{1+\beta X^T\Sigma_nX}\right)$$

$$=\beta*\frac{1+\boxed{\beta X^T\Sigma_nX-\beta X^T\Sigma_nX}}{1+\beta X^T\Sigma_nX}$$

$$=\frac{\beta}{1+\beta X^T\Sigma_nX}$$

$$u=\lambda^{-1}\beta X^TL\Sigma_n^{-1}u_n$$

$$=(\beta^{-1}+X^T\Sigma_nX)\beta X^TL\Sigma_n^{-1}u_n$$

$$=(\beta^{-1}+X^T\Sigma_nX)\beta X^T\left(\Sigma_n-\frac{\beta\Sigma_nXX^T\Sigma_n}{1+\beta X^T\Sigma_nX}\right)\Sigma_n^{-1}u_n$$

$$=\boxed{(1+\beta X^T\Sigma_nX)}X^T\cdot\frac{u_n+\boxed{u_n\beta X^T\Sigma_nX-\beta\Sigma_nXX^T\Sigma_n\Sigma_n^{-1}u_n}}{\boxed{1+\beta X^T\Sigma_nX}}$$

$$=X^Tu_n$$

Var(y) ⇒ $\lambda^{-1}=\beta^{-1}+X^T\Sigma_nX$

$w^*=(X^TX+\lambda I)^{-1}X^Ty$ ($\alpha=\lambda$,同时) $=\beta(\beta X^TX+\alpha I)^{-1}X^Ty$

综上所述 $p(y|X,D)=N(X^Tu_n,\ \beta^{-1}+X^T\Sigma_nX)$,其中 $\left\{\begin{array}{l}u_n=\beta\Sigma_nX^Ty \\ \Sigma_n=(\beta X^TX+\alpha I)^{-1}\end{array}\right.$

预测分布 ⇒ Gaussian 分布

↓ 见section 3.2.

4