先來推 posterior distribution.

1. Bayesian Linear Regression

$p(w|x,t,\alpha,\beta) \propto p(t|x,w,\beta)\,p(w|\alpha)$.

$p(w|\alpha) = N(w|m_0,S_0) \propto \exp(-\frac{1}{2}(w-m_0)^T S_0^{-1}(w-m_0)) \propto \exp((w-m_0)^T S_0^{-1}(w-m_0))$

$p(t|x,w,\beta) = \prod_{n=1}^{N} N(t|w^T\phi(x_n),\beta^{-1}) = \prod_{n=1}^{N} \frac{1}{(2\pi\beta)^{\frac{1}{2}}} \exp(-\frac{1}{2\beta^{-1}}(t-w^T\phi(x_n))^2)$

$\propto \exp(\sum_{n=1}^{N}(t_n-w^T\phi(x_n))^T\beta(t_n-w^T\phi(x_n)))$.

$\to p(w|x,t,\alpha,\beta) \propto p(t|x,w,\beta)\,p(w|\alpha) \propto \exp(\sum_{n=1}^{N}(t_n-w^T\phi(x_n))^T\beta(t_n-w^T\phi(x_n)))\exp((w-m_0)^T S_0^{-1}(w-m_0))$

$\log[p(w|w,t,\alpha,\beta)] = \sum_{n=1}^{N}(t_n-w^T\phi(x_n))^T\beta(t_n-w^T\phi(x_n)) + (w-m_0)^T S_0^{-1}(w-m_0)$.

$= \sum_{n=1}^{N}(t_n^T-\phi^T(x_n)w)\beta(t_n-w^T\phi(x_n)) + (w-m_0)^T S_0^{-1}(w-m_0)$.

前項乘開 $= \sum_{n=1}^{N}[t_n^T\beta t_n - t_n^T\beta w^T\phi(x_n) - \phi^T(x_n)w\beta t_n + \phi^T(x_n)w\beta w^T\phi(x_n)] + (w-m_0)^T S_0^{-1}(w-m_0)$

$= \sum_{n=1}^{N}[\underbrace{t_n^T\beta t_n}_{\text{跟w無關}} - t_n^T\beta w^T\phi(x_n) - \phi^T(x_n)w\beta t_n + w^T\phi(x_n)\beta\phi^T(x_n)w] + (w-m_0)^T S_0^{-1}(w-m_0)$.

因為 $\underset{M\times N\,N\times M}{\Phi^T\Phi} = \sum_{n=1}^{N}\phi(x_n)\phi(x_n)^T$, $\underset{M\times N\,N}{\Phi^T t} = \sum_{n=1}^{N} t_n\phi(x_n)^T$.

$\Rightarrow -\sum_{n=1}^{N} t_n^T\beta w^T\phi(x_n) - \sum_{n=1}^{N}\phi^T(x_n)w\beta t_n + \sum_{n=1}^{N} w^T\phi(x_n)\beta\phi^T(x_n)w + w^TS_0^{-1}w - w^TS_0^{-1}m_0$

$- m_0^TS_0^{-1}w + \underbrace{m_0^TS_0^{-1}m_0}_{\text{與w無關}}$

$\Rightarrow -\beta w^T\Phi^Tt - \beta t^T\Phi w + \beta w^T\Phi^T\Phi w + w^TS_0^{-1}w - w^TS_0^{-1}m_0 - m_0^TS_0^{-1}w$

令 posterior 的 mean: m_N

的 standard variant S_N

目標 $(w-m_N)^T S_N^{-1}(w-m_N) = w^TS_N^{-1}w - w^TS_N^{-1}m_N - m_N^TS_N^{-1}w + d$

$\to w^T(\beta\Phi^T\Phi + S_0^{-1})w$

$-w^T(\underbrace{\beta\Phi^Tt + S_0^{-1}m_0}_{\to S_N^{-1}m_N})$

$-(\beta t^T\Phi + m_0^TS_0^{-1})w$

可得知. $\to S_N^{-1} = \beta\Phi^T\Phi + S_0^{-1}$.

$m_N = S_N(\beta\Phi^Tt + S_0^{-1}m_0)$.

$$\boxed{\begin{array}{l} S_0 \text{ when } m_0 = 0 \text{ and } S_0 = \alpha^{-1}I \\ \to m_N = \beta S_N\Phi^Tt \\ \quad S_N^{-1} = \alpha I + \beta\Phi^T\Phi \end{array}}$$

predictive distribution

$$p(t|x,\mathbf{x},\mathbf{t}) = \int_{-\infty}^{\infty} p(t|x,w,\beta)\,p(w|\mathbf{x},\mathbf{t})\,dw$$

$$\propto \int N(t|w^T\phi(x),\beta^{-1})\,N(w|m_N,S_N)$$

$$\propto \int \exp\left(-\frac{\beta}{2}(t-w^T\phi(x))^2\right)\exp\left(-\frac{1}{2}(w-m_N)^T S_N(w-m_N)\right)$$

$$\propto \int \exp\left(-\frac{\beta}{2}(t^2-2w^T\phi(x)t+w^T\phi(x)\phi^T(x)w)-\frac{1}{2}(w^TS_Nw-2w^TS_Nm_N+\cancel{m_N^TS_Nm_N})\right)$$

與t和w無關

取log：$-\frac{1}{2}(\beta t^2-2\beta w^T\phi(x)t+\beta w^T\phi(x)\phi^T(x)w+w^TS_Nw-2w^TS_Nm_N)$

$$= -\frac{1}{2}\left[\,w^T(\beta\phi(x)\phi^T(x)+S_N)w-2w^T(t\beta\phi(x)+S_Nm_N)+\beta t^2\right]$$

目標：$(w-m)^TL(w-m) = w^TLw-2w^TLm+m^TLm$ 對照上方式子.

取 $L=\beta\phi(x)\phi^T(x)+S_N$ ， $Lm=\beta t\phi(x)+S_Nm_N$ ， $m=L^{-1}(\beta t\phi(x)+S_Nm_N)$

底線部份可寫成

$$\left(w^TLw-2w^TLm+m^TLm\right)-m^TLm+\beta t^2$$

$$= (w-m)^TL(w-m)-m^TLm+\beta t^2.$$

$$p(t|x,\mathbf{x},\mathbf{t}) \propto \int \exp\left(-\frac{1}{2}(w-m)^TL(w-m)\right)\exp\left(\frac{1}{2}m^TLm-\frac{1}{2}\beta t^2\right)dw$$

由於前項depend on w，積起來等於1，因此正比於後項

$$\propto \exp\left(\frac{1}{2}m^TLm-\frac{1}{2}\beta t^2\right)$$

$$m^TLm = (\beta t\phi(x)+S_Nm_N)^TL^{-1}\cancel{L}\cancel{L^{-1}}(\beta t\phi(x)+S_Nm_N)$$

沒有變數depend on y.

$$= \beta t\phi^T(x)L^{-1}\beta t\phi(x)+2\beta t\phi^T(x)L^{-1}S_Nm_N+\cancel{m_N^TS_NL^{-1}S_Nm_N}$$

$$= \left(\beta^2\phi^T(x)L^{-1}\phi(x)\right)t^2+\left(2\beta\phi^T(x)L^{-1}S_Nm_N\right)t+\text{const}.$$

$$-\frac{1}{2}(\beta t^2-m^TLm) = -\frac{1}{2}\left[(\beta-\beta^2\phi^T(x)L^{-1}\phi(x))t^2-2(\beta\phi^T(x)L^{-1}S_Nm_N)t+C\right].$$

目標 $-\frac{\lambda}{2}(t-m)^2 = -\frac{1}{2}(\lambda t^2-2\lambda tm+\lambda m^2)$ 與上方式對照

取 $\lambda=\beta(1-\beta\phi^T(x)L^{-1}\phi(x))$ ， $m=\frac{1}{\lambda}\beta\phi^T(x)L^{-1}S_Nm_N$

接著整理 λ 與 m

$\boxed{\text{首先}\ \lambda}$

$$L^{-1}=(\beta\phi(x)\phi^T(x)+S_N)^{-1} = S_N^{-1}-\frac{S_N^{-1}\beta\phi(x)\phi^T(x)S_N^{-1}}{1+\beta\phi^T(x)S_N^{-1}\phi(x)}$$

By Sherman Morriso formula

令 $\gamma=\phi^T(x)S_N^{-1}\phi(x)$

$$\phi^T(x)L^{-1}\phi(x) = \gamma-\frac{\beta\gamma^2}{1+\beta\gamma} = \frac{\gamma}{1+\beta\gamma}$$

$$\lambda=\beta(1-\beta\phi^T(x)L^{-1}\phi(x)) = \beta\left(1-\frac{\beta\gamma}{1+\beta\gamma}\right) = \beta\frac{1}{1+\beta\gamma} = \frac{\beta}{1+\beta\gamma}$$

$$\frac{1}{\lambda} = \frac{1+\beta\gamma}{\beta} = \beta^{-1}+\gamma = \beta^{-1}+\phi^T(x)S_N^{-1}\phi(x)$$

$\boxed{m.}$

$$m = \frac{1}{\lambda}\beta\phi^T(x)L^{-1}S_N m_N = m_N^T\left(\frac{1}{\lambda}\beta S_N L^{-1}\phi(x)\right)$$

證 $\phi(x) = \frac{1}{\lambda}\beta S_N L^{-1}\phi(x)$

$\Rightarrow$ $L S_N^{-1}\phi(x) = \frac{1}{\lambda}\beta\phi(x)$

左式 $= (\beta\phi(x)\phi^T(x) + S_N)S_N^{-1}\phi(x) = \beta\phi(x)\underbrace{\phi^T(x)S_N^{-1}\phi(x)}_{\gamma} + \phi(x)$

$= (\beta\gamma + 1)\phi(x)$

右式 $= \frac{1}{\lambda}\beta\phi(x) = \frac{1+\beta\gamma}{\beta}\beta\phi(x) = (1+\beta\gamma)\phi(x)$.

故立 → 所以 $m = m_N^T\phi(x) \overset{(代入 m_N^T)}{=} ((\beta S_N\Phi^T t)^T\phi(x) = \phi^T(x)\beta S_N\Phi^T t$

$= \beta\phi^T(x)S_N\sum_{n=1}^{N}\phi(x_n)t_n$

整理後.

$$\lambda = \beta^{-1} + \phi^T(x)S_N^{-1}\phi(x)$$

$$m = \beta\phi^T(x)S_N\sum_{n=1}^{N}\phi(x_n)t_n.$$

其中 $S_N^{-1} = \alpha I + \beta\Phi^T\Phi = \alpha I + \beta\sum_{n=1}^{N}\phi(x_n)\phi^T(x_n)$

2. Jensens Inequality.

Basis step $M=2$ $f(\lambda_1 x_1 + \lambda_2 x_2) \le \lambda_1 f(x_1) + \lambda_2 f(x_2)$.

$\because \sum_i \lambda_i = 1$

$\therefore f(\lambda_1 x_1 + (1-\lambda_1) x_2) \le \lambda_1 f(x_1) + (1-\lambda_1) f(x_2)$

$\rightarrow$ equal to equation (1).

Induction Step.

for any integer $K \ge 2$, $f(\sum_{i=1}^{K} \lambda_i x_i) \le \sum_{i=1}^{K} \lambda_i f(x_i)$

Prove $K+1$ is also true.

$$f(\sum_{i=1}^{K+1} \lambda_i x_i) = f(\underline{\lambda_{K+1}}\, \underline{x_{K+1}} + \underline{(1-\lambda_{K+1})} \underline{\sum_{i=1}^{K} \frac{\lambda_i}{1-\lambda_{K+1}} x_i}).$$

$$\le \lambda_{K+1} f(x_{K+1}) + (1-\lambda_{K+1}) f(\sum_{i=1}^{K} \frac{\lambda_i}{1-\lambda_{K+1}} x_i).$$

$$\le \lambda_{K+1} f(x_{K+1}) + (1-\lambda_{K+1}) \sum_{i=1}^{K} \frac{\lambda_i}{1-\lambda_{K+1}} f(x_i)$$

$$= \lambda_{K+1} f(x_{K+1}) + \sum_{i=1}^{K} \lambda_i f(x_i)$$

$$= \sum_{i=1}^{K+1} \lambda_i f(x_i).$$

Hence, equation (2) is true.