Práctica minería de datos

Xavier Mira Fernandez

February 1, 2016

Contents

1 Ejercicios de descubrimiento **3**
1.1 Ejercicio 1.5 3
1.2 Ejercicio 1.6 3
1.3 Ejercicio 1.11 3
1.4 Ejercicio 1.30 4
1.5 Ejercicio 1.32 5
1.6 Ejercicio 1.33 5
1.7 Ejercicio 1.39 6

2 Ejercicios del tema 2 del libro **7**
2.1 Ejercicio 2.8 7
2.2 Ejercicio 2.13 7
2.3 Ejercicio 2.15 8
2.4 Ejercicio 2.33 9
2.5 Ejercicio 2.44 9
2.6 Ejercicio 2.45 11

3 Ejercicios del tema 3 del libro **12**
3.1 Ejercicio 3.2 12
3.2 Ejercicio 3.7 12
3.3 Ejercicio 3.11 13
3.4 Ejercicio 3.16 14

4 Ejercicios tema 4 del libro **16**
4.1 Ejercicio 4.1 16
4.2 Ejercicio 4.10 17

1 Ejercicios de descubrimiento

1.1 Ejercicio 1.5

Hay que demostrar que $var\{f(x)\}$ satisface (1.39) $= E\{f(x)^2\} - E\{f(x)\}^2$

Esto implica demostrar que $E\{(f(x) - E\{f(x))^2\} = E\{f(x)^2\} - E\{f(x)\}^2$

$$E\{(f(x) - E\{f(x)\})^2\} = E\{f(x)^2\} - E\{f(x)\}^2$$

Si expandimos la parte izquierda

$$E\{f(x)^2 - 2f(x)E\{f(x)\} + E\{f(x)\}^2\}$$

$$E\{f(x)^2\} - 2E\{f(x)\} + E\{f(x)\}^2$$

$$= E\{f(x)^2\} - E\{f(x)\}^2$$

Que es lo que queríamos demostrar

1.2 Ejercicio 1.6

Demostrar que si x e y son independientes su covarianza es 0

Sabiendo que la covarianza $cov\{x, y\} = E\{x, y\} - E\{x\}E\{y\}$ y sabiendo que cuando las variables son independientes $p(x, y) = p(x)p(y)$

$$\rho_{x,y} = \sum_x \sum_y xyP(x, y) - \sum_x xP(x) \sum_y yP(y)$$

Aplicando lo que sabemos sobre las variables independientes

$$\rho_{x,y} = \sum_x \sum_y xP(x)yP(y) - \sum_x xP(x) \sum_y yP(y)$$

Sin más que reordenar tenemos

$$\rho_{x,y} = \sum_x xP(x) \sum_y yP(y) - \sum_x xP(x) \sum_y yP(y) = 0$$

1.3 Ejercicio 1.11

Para μ

Tomamos la log-verosimilitud

$$\ln P(x \mid \mu, \sigma^2) = \frac{-1}{2\sigma^2} \sum_{i=1}^{N} (x_i - \mu)^2 - \frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln 2\pi$$

Derivando respecto a μ, igualando a 0 y resolviendo para μ

$$\frac{-1}{2\sigma^2}(-2) \left(\sum_{i=1}^{N} (x_i) - \mu N \right) = 0$$

$$\frac{1}{\sigma^2}\sum_{i=1}^{N}(x_i) = \frac{1}{\sigma^2}\mu N$$

$$\frac{1}{N}\sum_{i=1}^{N}(x_i) = \mu_{ML}$$

Para σ^2 hacemos lo mismo. Derivamos respecto a σ^2, igualamos a 0 para el punto crítico y resolvemos

$$\frac{\partial}{\partial\sigma^2} f = \frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{N}(x_i-\mu)^2 - \frac{N}{2\sigma^2} = 0$$

$$\frac{1}{2\left(\sigma^2\right)^2}\sum_{i=1}^{N}(x_i-\mu)^2 = \frac{N}{2\sigma^2}$$

Si multiplicamos ambos lados por $2\left(\sigma^2\right)^2$ y dividimos ambos lados por N queda

$$\frac{1}{N}\sum^{N}(x_i-\mu)^2 = \sigma^2_{ML}$$

1.4 Ejercicio 1.30

Evaluar la divergencia de Kullback-Leibler entre 2 gausianas $P(x) = N(x \mid \mu, \sigma^2)$ y $Q(x) = N(x \mid m, s^2)$

$$Kld(p||q) = -\int p(x) - \ln\frac{q(x)}{p(x)}dx$$

Lo separamos para obtener

$$-\int P(x)\ln P(x)dx + \int P(x)\ln q(x)$$

La primera integral la reconocemos como la entropía negativa, que evaluamos directamente $-\frac{1}{2}\left(1+\ln 2\pi\sigma^2\right)$
La segunda integral es

$$\int N(x \mid \mu, \sigma^2)\ln N(x \mid m, s^2)$$

Vamos a usar $P(x)$ en vez de $N(x \mid \mu, \sigma^2)$ para abreviar

$$\int P(x)\frac{1}{2}\left(\ln 2\pi s^2 - \frac{(x-m)^2}{s^2}\right)$$

$$= \frac{1}{2}\left(\int P(x)\ln 2\pi s^2 - \frac{1}{s^2}\int P(x)\left(x-m\right)^2\right)$$

$$= \frac{1}{2}\left(\ln 2\pi s^2 - \frac{1}{s^2}\left(\underbrace{\int P(x)x^2}_{E\{x^2\}=\mu^2+\sigma^2} \underbrace{-2m\int P(x)x}_{2mE\{x\}=2m\mu} + \underbrace{m^2\int P(x)}_{m^2}\right)\right)$$

$$= \frac{1}{2}\left(\ln 2\pi s^2 - \frac{\mu^2+\sigma^2-2m\mu+m^2}{s^2}\right)$$

Si lo juntamos todo

$$= \frac{1}{2}\left(\ln 2\pi s^2 - \frac{\mu^2 + \sigma^2 - 2m\mu + m^2}{s^2} - 1 - \ln 2\pi\sigma^2\right)$$

Por las propiedades del logaritmo y reordenando

$$= \frac{1}{2}\left(-1 + \ln\frac{s^2}{\sigma^2} - \frac{(m-\mu)^2 + \sigma^2}{s^2}\right)$$

1.5 Ejercicio 1.32

No estoy muy seguro pero creo que es así. Para empezar hay que ver que se hace una transformación lineal según (1.27), por lo que

Considerando $P_x(x)$ que se corresponde con $P_y(y)$ las observaciones que caigan en el rango $(x, x + \delta x)$ para δ pequeño, seran transformadas en el rango $(y, y + \delta y)$ tal que $y = Ax$

Por tanto $P_y(y) = P_x(x)\left|\frac{dx_i}{dy_j}\right|$

Según el enunciado la transformación es $y = Ax$, A sería la matriz Jacobiana y $|A|$ su determinante. Teniendo eso en cuenta

$$P_x(x) = P_y(y)\left|\frac{dy_i}{dx_j}\right| = P_y(y)|A| \Rightarrow P_x(x)|A|^{-1} = P_y(y)$$

Teniendo esto en cuenta

$$H\{y\} = -\int P(y)\ln P(y) = -\int P(x)\ln\left(P(x)\cdot|A|^{-1}\right)dx = -\int P(x)\left(\ln P(x) - \ln|A|^{-1}\right)dx$$

$$= -\int P(x)\ln P(x) - \int P(x) - \ln|A|dx$$

$$= H\{x\} + \ln|A|\int P(x)$$

Como $\int P(x)$ evalúa a 1

$$H\{y\} = H\{x\} + \ln|A|$$

1.6 Ejercicio 1.33

Sabiendo que $H\{y \mid x\} = 0$ demostrar que $\forall x$ solo existe 1 y tal que $P(y \mid x) \neq 0$

Según (1.111) la entropia condicional $H\{y \mid x\} = -\int\int P(y,x)\ln P(y \mid x)dydx$

Sabemos que $P(x,y) = P(x \mid y)P(y) = P(y \mid x)P(x)$

Tomamos la forma discreta de la entropía

$$H\{y \mid x\} = -\sum_x\sum_y P(y)P(x \mid y)\ln P(y \mid x)$$

Si la entropía es 0 entonces para cada x solo existe un y en el cual la probabilidad $P(y \mid x) = 1$. Viendo la expresión de arriba vemos claramente que cuando la probabilidad es 1 el logaritmo es 0, en el resto de casos dado que $P(y)P(x \mid y)$ es 0 el resultado sigue siendo 0.

1.7 Ejercicio 1.39

Lo primero es obtener las probabilidades marginales de la tabla de probabilidad conjunta

$$P(x) = \sum_y P(x, y)$$

$$P(y) = \sum_x P(x, y)$$

Ahora, sabiendo que $P(x, y) = P(x)P(y \mid x) = P(y)P(x \mid y)$
Sacamos $P(y \mid x)$ y $P(x \mid y)$ tal que

$$P(y \mid x) = \frac{P(x, y)}{P(x)}$$

$$P(x \mid y) = \frac{P(x, y)}{P(y)}$$

Ahora ya podemos ir a por la entropía

$$H\{x\} = -\sum P(x_i) \ln P(x_i) = \ln 3 - \frac{2}{3} \ln 2$$

$$H\{y\} = -\sum P(y_i) \ln P(y_i) = \ln 3 - \frac{2}{3} \ln 2$$

$$H\{x \mid y\} = -\sum_x \sum_y P(x, y) \ln P(x \mid y) = \frac{2}{3} \ln 2$$

$$H\{y \mid x\} = -\sum_x \sum_y P(x, y) \ln P(y \mid x) = \frac{2}{3} \ln 2$$

$$H\{x, y\} = -\sum_x \sum_y P(x, y) \ln P(x, y) = \ln 3$$

$$I\{x, y\} = H\{x\}H\{x \mid y\} = \ln 3 - \frac{2}{3} \ln 2 - \frac{2}{3} \ln 2$$

2 Ejercicios del tema 2 del libro

2.1 Ejercicio 2.8

Demostrar que $E\{X\} = E_y\{E\{X \mid Y\}\}$

Partimos de esa expresión $E\{X\} = E_y\{E\{X \mid Y\}\}$

$$\sum_x x_i P(x_i) = \sum_y \left(\sum_x x P(x \mid y) \right) P(y)$$

$$\sum_x x P(x) = \sum_y \left(\sum_x x P(x \mid y) P(y) \right)$$

Sabiendo que $P(x \mid y)P(y) = P(x, y)$ tenemos

$$\sum_x x P(x) = \sum_y \left(\sum_x x P(x, y) \right)$$

Reordenando nos queda

$$\sum_x x P(x) = \sum_x x \sum_y P(x, y)$$

Esto marginaliza la probabilidad conjunta $P(x, y)$ y nos deja $P(x)$

$$\sum_x x P(x) = \sum_x x P(x)$$

Demostrar que $var[x] = E_y[var_x[x \mid y]] + var_y[E_x[x \mid y]]$

Expandimos con la definición de varianza

$$E_y\left[E_x\left[x^2 \mid y\right] - E_x\left[x \mid y\right]^2\right] + E_y\left[E_x\left[x \mid y\right]^2\right] - E_y\left[E_x\left[x \mid y\right]\right]^2$$

$$\underbrace{E_y\left[E_x\left[x^2 \mid y\right]\right]}_{E[x^2]} - \underbrace{E_y\left[E_x\left[x \mid y\right]^2\right] + E_y\left[E_x\left[x \mid y\right]^2\right]}_{0} - \underbrace{E_y\left[E_x\left[x \mid y\right]\right]^2}_{E[x]^2}$$

$$= E\left[x^2\right] - E\left[x\right]^2 = var\left[x\right]$$

2.2 Ejercicio 2.13

Evaluar KL(p||q) con $p(x) = N(x \mid \mu, \Sigma)$ y $q(x) = N(x \mid m, L)$ gausianas multivariantes.

$$Kl(p||q) = \int p(x) \ln \frac{q(x)}{p(x)} dx$$

$$= \int p(x) \ln q(x) dx - \underbrace{\int p(x) \ln p(x) dx}_{=-H[x]}$$

Desarrollamos la integral de la izquierda porque la otra es solamente la entropía negativa

$$= \int p(x) \left(\frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |L| - \frac{1}{2} \left(x-m\right)^T L^{-1} \left(x-m\right) \right) dx$$

Dado que p(x) evalúa a 1 cuando integramos, nos queda

$$= \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |L| - \frac{1}{2} \int p(x) \left(x-m\right)^T \Sigma^{-1} \left(x-m\right) dx$$

$$= \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |L| - \frac{1}{2} \int p(x) \left(x^T L^{-1} x - x^T L^{-1} m - m^T L^{-1} x + m^T \Sigma^{-1} m \right) dx$$

$$= \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |L| - \frac{1}{2} \left(\underbrace{\int p(x) x^T L^{-1} x \, dx}_{E[x^T L^{-1} x]} - \underbrace{\int p(x) x^T L^{-1} m \, dx}_{E[x^T L^{-1} m]} - \underbrace{\int p(x) m^T L^{-1} x \, dx}_{E[m^T L^{-1} x]} + \underbrace{\int p(x) m^T \Sigma^{-1} m \, dx}_{E[m^T L^{-1} m]} \right)$$

$$= \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |L| - \frac{1}{2} \left(E[x^T L^{-1} x] - E[x^T L^{-1} m] - E[m^T L^{-1} x] + E[m^T L^{-1} m] \right)$$

Si lo juntamos con la entropía negativa de la otra integral

$$= \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |L| - \frac{1}{2} \left(E[x^T L^{-1} x] - E[x^T L^{-1} m] - E[m^T L^{-1} x] + E[m^T L^{-1} m] \right) \underbrace{- \frac{1}{2} \ln |\Sigma| - \frac{D}{2} \left(1 + \ln 2\pi\right)}_{-H[x]}$$

$$= \frac{1}{2} \left(\ln \frac{|L|}{|\Sigma|} - D - E[x^T L^{-1} x] - E[x^T L^{-1} m] - E[m^T L^{-1} x] + E[m^T L^{-1} m] \right)$$

Más facil. De (380) de *Matrix Cookbook* tenemos que

$$E \left[(x-m)^T A (x-m) \right] = (\mu - m)^T A (\mu - m) + Tr \left[A\Sigma \right]$$

Luego, usando esa identidad nos queda

$$\frac{1}{2} \ln \frac{|L|}{|\Sigma|} - \frac{D}{2} + (\mu - m)^T L^{-1} (\mu - m) + Tr \left[A\Sigma \right]$$

2.3 Ejercicio 2.15

Demostrar que la entropía para x $H[x] = \frac{1}{2} \ln |\Sigma| + \frac{D}{2} (1 + \ln 2\pi)$ tal que $x \sim N(x \mid \mu, \Sigma)$

$$H[x] = - \int p(x) \ln p(x) dx$$

$$- \int N(x \mid \mu, \Sigma) \frac{1}{2} \left(D \ln 2\pi + \ln |\Sigma| + \frac{1}{2} \left(x - \mu\right)^T \Sigma^{-1} (x - \mu) \right)$$

Sabiendo que

$$E \left[(x-m)^T A \left(x-m\right) \right] = (\mu - m)^T A \left(\mu - m\right) + Tr[A\Sigma]$$

Tenemos

$$= \frac{D}{2} \ln 2\pi + \frac{1}{2} \ln |\Sigma| + \frac{1}{2} \underbrace{Tr[\Sigma^{-1}\Sigma]}_{\sum^D \frac{1}{u_{ii}} u_{ii} = D}$$

$$= \frac{D}{2} \ln 2\pi + \ln |\Sigma| + \frac{1}{2} D = \frac{D}{2} \left(\ln 2\pi + 1\right) + \frac{1}{2} \ln |\Sigma|$$

2.4 Ejercicio 2.33

Teniendo en cuenta del ejercicio 3.32 que $p(x, y)$ en el exponente tiene $-\frac{1}{2}\left((x-\mu)^T \Lambda (x-\mu) + (y - Ax - b)^T L (y$

Sabiendo que $p(x, y) = p(x)p(y \mid x) = p(y)p(x \mid y)$, tomando logaritmos

$$\ln p(x, y) = \ln p(y) + \ln p(x \mid y)$$

$$\ln p(x \mid y) = \ln p(x, y) - \ln p(y)$$

$$\ln p(x, y) \propto -\frac{1}{2}\left((x-\mu)^T \Lambda (x-\mu) + (y - Ax - b)^T L (y - Ax - b)\right)$$

$$\ln p(y) \propto -\frac{1}{2}\left((y - A\mu + b)^T \left(L^{-1} + A\Lambda^{-1}A^T\right)^{-1} (y - A\mu + b)\right) \text{ por } (2.115)$$

$$\ln p(x \mid y) \propto -\frac{1}{2}\left((x-\mu)^T \Lambda (x-\mu) + (y - Ax - b)^T L (y - Ax - b) - (y - A\mu + b)^T \left(L^{-1} + A\Lambda^{-1}A^T\right)^{-1} (y - A$$

La forma funcional que buscamos es cuadrática - $x^T Rx - x^T Rm - \ldots$ - por lo que vemos la distribución marginal p(y) no nos aporta más que una constante. Así pues expandimos la parte de la distribución conjunta

$$-\frac{1}{2}(x^T\Lambda x - x^T\Lambda\mu - \mu^T\Lambda x + \underbrace{\mu^T\Lambda\mu}_{const} + \underbrace{y^T y}_{const} - \underbrace{y^T LAx}_{x^T A^T Ly} - \underbrace{y^T Lb}_{const}$$

$$-x^T A^T Ly + x^T A^T LAx + x^T A^T Lb - \underbrace{b^T Ly}_{const} + b^T LAx + \underbrace{b^T b}_{const}$$

Vemos que, para obtener una forma cuadrática $(x-m)^T R(x-m) = x^T Rx - \underbrace{x^T Rm - m^T Rx}_{2x^T Rm} + const$ el primer término, que sería R se saca fácil factorizando

$$R = \left(\Lambda\mu + A^T LA\right)$$

Vemos que si cogemos x^T podemos sacar $x^T Rm$ reordenando

$$x^T Rm = x^T\left(\Lambda\mu + A^T Ly - A^T Lb\right) = x^T\left(\Lambda\mu + A^T L(y-b)\right)$$

Luego

$$Rm = \Lambda\mu + A^T L(y-b) \Rightarrow m = R^{-1}\left(\Lambda\mu + A^T L(y-b)\right)$$

2.5 Ejercicio 2.44

Hay que demostrar que la posteriori tiene la misma forma funcional que la prior y escribir la expresión para los parámetros de la posteriori.

Bien, tenemos para empezar una normal $N(x \mid \mu, \tau^{-1})$ y una prior conjugada dada por (2.154) $P(\mu, \lambda) = N\left(\mu \mid \mu_0, (\beta\lambda)^{-1}\right) Gam\left(\lambda \mid a, b\right)$

Teniendo en cuenta Bayes sabemos que

$$P(\mu, \lambda, x) = P(\mu, \lambda \mid x)P(x) = P(x \mid \mu, \lambda)P(\mu, \lambda)$$

La posteriori, que es la que necesitamos, será

$$P(\mu, \lambda \mid x) \propto P(x \mid \mu, \lambda)P(\mu, \lambda)$$

Por 2.152

$$P(x \mid \mu, \lambda) \propto \left[\lambda^{\frac{1}{2}} exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N exp\left(\lambda\mu\sum^N x_n - \frac{\lambda}{2}\sum^N x_n^2\right)$$

$$N\left(\mu \mid \mu_0, (\beta\lambda)^{-1}\right) \propto exp\left(-\frac{\beta\lambda}{2}(\mu-\mu_0)^2\right)$$

$$Gam\left(\lambda \mid a, b\right) \propto \lambda^{a-1} exp\left(-b\lambda\right)$$

Si lo juntamos todo

$$P(\mu, \lambda \mid x) \propto \left[\lambda^{\frac{1}{2}} exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N exp\left(-\frac{\beta\lambda}{2}(\mu-\mu_0)^2 + \lambda\mu\underbrace{\sum^N x_n}_{c} - \frac{\lambda}{2}\underbrace{\sum^N x_n^2}_{d} - b\lambda\right)\lambda^{a-1}$$

$$\propto \lambda^{\frac{N}{2}+a-1} exp\left(-\frac{\lambda\mu^2 N}{2}\right) exp\left(-\frac{\beta\lambda}{2}\left(\mu^2 - 2x\mu_0 + \mu_0^2\right) + \lambda\mu c - \frac{\lambda d}{2} - b\lambda\right)$$

$$\propto \lambda^{\frac{N}{2}+a-1} exp\left(-\frac{1}{2}\lambda\mu^2 N + \lambda\mu c - \frac{1}{2}\lambda d - b\lambda - \frac{1}{2}\beta\lambda\mu^2 + \beta\lambda\mu\mu_0 - \frac{1}{2}\beta\lambda\mu_0^2\right)$$

Separamos en 2 exponentes. Uno será la parte del Gamma y la otra la gausiana para obtener la forma funcional gauss-gamma.

$$\propto \lambda^{\frac{N}{2}+a-1} exp\left(-\underbrace{\left(b + \frac{1}{2}d + \frac{1}{2}\beta\mu_0^2\right)}_{\hat{\beta}}\lambda\right) exp\left(-\frac{1}{2}\lambda\mu^2 N + \lambda\mu c - \frac{1}{2}\beta\lambda\mu^2 + \beta\lambda\mu\mu_0\right)$$

Parece que ya tengo la parte de la gamma pero de momento no se cómo proceder para completar el cuadrado en el exponente de la derecha. He probado reordenando y agrupando los términos en μ^2 y μ pero me quedo sin el término independiente del cuadrado.

Tomaré el término $\frac{1}{2}\beta\mu_0^2\lambda$ del exponente izquierdo y lo usaré para tener el término independiente.

$$\propto \lambda^{\frac{N}{2}+a-1} exp\left(-\underbrace{\left(b + \frac{1}{2}d\right)}_{\hat{\beta}}\lambda\right) exp\left(-\frac{1}{2}\lambda\mu^2(N+\beta) + +\lambda\mu\left(\mu_0(1+\beta)+c\right) + \frac{1}{2}\beta\mu_0^2\lambda\right)$$

Para completar el cuadrado sabemos que

$$c_2\mu^2 + c_1\mu + c_0 = -\frac{1}{2\sigma^2}(x-\theta)^2 + d$$

Pues

$$\theta = -\frac{c_1}{2c_2}, \sigma^2 = -\frac{1}{2c_2}, d = -\frac{c_1^2}{4c_2}$$

$$\propto \lambda^{\frac{N}{2}+a-1} exp\left(-\underbrace{\left(b + \frac{1}{2}d\right)}_{\hat{\beta}}\lambda\right) exp\left(-\frac{1}{2\left(-\frac{1}{2}\lambda(N+\beta)\right)}\left(\mu - \underbrace{\frac{\lambda\left(\mu_0(1+\beta)+c\right)}{2\left(-\frac{1}{2}\lambda(N+\beta)\right)}}_{\hat{\mu}_0}\right)^2 - \frac{\left(\lambda\left(\mu_0(1+\beta)+c\right)\right)^2}{4\left(-\frac{1}{2}\lambda(N+\beta)\right)}\right)$$

Por tanto tenemos 2 exponentes, uno con la forma del gamma y otro con la forma funcional de la gaussiana quedándonos una posteriori funcionalmente similar a la prior, que es lo que queríamos demostrar.

2.6 Ejercicio 2.45

Comprobar que la distribución Wishart es una conjugada a priori para la matriz de precisión de una gausiana multivariante.

Para ello tomamos la función de verosimilitud para Λ dado un data set

$$\prod^{N} N(x_i \mid \mu, \Lambda^{-1}) \propto |\Lambda|^{\frac{1}{2}} exp\left(-\frac{1}{2}\sum^{N}(x_i-\mu)^T \Lambda (x_i-\mu)\right)$$

$$\propto |\Lambda|^{\frac{1}{2}} exp\left(-\frac{1}{2}Tr\left[\Lambda M\right]\right)$$

siendo M la matriz $\sum^N (x_i-\mu)(x_i-\mu)^T$

Ahora, si nos fijamos en (2.155) vemos que la forma funcional es la misma y por tanto, si lo multiplicamos por una prior Wishart tendremos una Wishart a posteriori.

Demostración

Ya tenemos la likelihood, por lo tanto tomamos una prior wishart con la definición de (2.155) y la multiplicamos por la likelihood

$$W(\Lambda \mid W, v)\prod^{N} N(x_i \mid \mu, \Lambda^{-1}) \propto B|\Lambda|^{\frac{v-D-1}{2}}\Lambda^{\frac{1}{2}} exp\left(-\frac{1}{2}\left(Tr[\Lambda M]+Tr[W^{-1}\Lambda]\right)\right)$$

Como $Tr[A]+Tr[B]=Tr[A+B]$

$$W(\Lambda \mid W, v)\prod^{N} N(x_i \mid \mu, \Lambda^{-1}) \propto B|\Lambda|^{\frac{v-D-1}{2}}\Lambda^{\frac{1}{2}} exp\left(-\frac{1}{2}\left(Tr[\Lambda M+W^{-1}\Lambda]\right)\right)$$

Que sigue teniendo una forma funcional Wishart dado que en el exponente seguimos teniendo una dependencia funcional de Λ

3 Ejercicios del tema 3 del libro

3.1 Ejercicio 3.2

Demostrar que $\Phi(\Phi^T\Phi)^{-1}\Phi^T$ toma cualquier vector v y lo proyecta en el subespacio vectorial formado por los vectores columna de Φ

Usa este resultado para demostrar que la solución de mínimos cuadrados corresponde a la proyección ortogonal del vector t en el sistema generador Φ

Tomamos el vector $v' = \Phi(\Phi^T\Phi)^{-1}\Phi^T v$

Si usamos $(\Phi^T\Phi)^{-1}\Phi^T v = \gamma$ vemos que $v' = \Phi\gamma$.

Vemos que esto usa Φ como un sistema generador del subespacio vectorial, por lo que producirá un vector que es combinacion de los vectores columna $\varphi_j \in \Phi$ de modo que

$$\Phi\gamma = \varphi_1\gamma_1 + \ldots + \varphi_m\gamma_m$$

Siendo γ_i el componente i-ésimo del vector γ y φ_i la columna i-ésima de la matriz Φ.

Si comparamos esto con el método de mínimos cuadrados vemos que tomamos $y = \Phi w_{ML} = \Phi(\Phi^T\Phi)^{-1}\Phi^T t$ por lo que es la proyección de t en el espacio vectorial generado por las columnas de Φ.

A partir de aquí tomamos t y lo descomponemos en 2 vectores, uno dentro del subespacio vectorial y el otro fuera, esto es, uno es y y el otro es t según la figura 3.2 de PRML de forma que

$$t = y + f \to t - y = f$$

Hay que demostrar que el vector $f = t - y$ es ortogonal a la matriz Φ. Sabiendo que φ_j es la columna j de Φ

$$(y-t)^T \varphi_j = (\Phi w_{ML} - t)\,\varphi_j = \left(\Phi(\Phi^T\Phi)^{-1}\Phi^T t - t\right)\varphi_j$$

Sacando el factor común

$$t^T\left(\Phi\underbrace{(\Phi^T\Phi)^{-1}\Phi^T}_{\Phi^\dagger} - I\right)\varphi_j$$

Vemos que para cualquier φ_j el resultado es 0 dado que las propiedades de la matriz inversa son aplicables para la semi-inversa de Moore-Penrose. La condición de ortogonalidad se cumple $\forall\varphi_j$ por lo que $y - t$ es ortogonal a Φ.

3.2 Ejercicio 3.7

Verificar el resultado (3.49) usando la técnica de completar el cuadrado para la distribución a posteriori de los parámetros w en el que m_N y S_N vienen dadas por (3.50) y (3.51)

Por Bayes sabemos que

$$P(w \mid t) = aP(t \mid w)P(w)$$

La prior $P(w \mid t) = N(w \mid m_0, S_0)$ y la verosimilitud $P(t \mid w) = \prod^N N(t \mid w\Phi, b^{-1})$

$P(w \mid t)$ será proporcional al exponente de $P(t \mid w)$ por el exponente de $P(w)$. De 3.10 y 3.48 tenemos

$$P(w \mid t) \propto exp\left(-\frac{b}{2}(t - \Phi w)^T(t - \Phi w)\right) exp\left(-\frac{1}{2}(w - m_0)^T S_0^{-1}(w - m_0)\right)$$

Tenemos que conseguir una forma cuadrática en el exponente tal que $w^T\Lambda w - w^T\Lambda\mu - \mu^T\Lambda w + const$

$$\propto exp\left(-\frac{1}{2}\left(bt^Tt - bt^T\Phi w - bw^T\Phi^Tt + bw^T\Phi^T\Phi w + w^TS_0^{-1}w - w^TS_0^{-1}m_0 - m_0^TS_0^{-1}w + m_0^TS_0^{-1}m_0\right)\right)$$

Reordenando y agrupando los términos para conseguir la forma funcional que queremos

$$\propto exp\left(-\frac{1}{2}\left(w^T\left(S_0^{-1} + b\Phi^T\Phi\right)w - w^T\underbrace{\left(b\Phi^Tt + S_0^{-1}m_0\right)}_{\Lambda\mu} - \left(bt^T\Phi + m_0^TS_0^{-1}\right)w + const\right)\right)$$

De aquí vemos que

$$\Lambda = S_0^{-1} + b\Phi^T\Phi$$

$$\Lambda\mu = b\Phi^Tt + S_0^{-1}m_0$$

$$\mu = \Lambda^{-1}\left(b\Phi^Tt + S_0^{-1}m_0\right)$$

De donde observamos que $S_N = \Lambda$ y $m_N = \mu$

3.3 Ejercicio 3.11

Haciendo uso de la identidad

$$\left(M + vv^T\right)^{-1} = \frac{\left(M^{-1}v\right)\left(v^TM^{-1}\right)}{1 + v^TM^{-1}v}$$

demostrar que la varianza σ_N^2 satisface

$$\sigma_N^2 \leq \sigma_{N+1}^2$$

Teniendo en cuenta que

$$\sigma_{n+1}^2 = \frac{1}{\beta} + \phi(x)^TS_{n+1}\phi(x)$$

Sabiendo que

$$S_{n+1}^{-1} = S_n^{-1} + \beta\phi_{n+1}\phi_{n+1}^T$$

Si lo sustituimos en la expresión anterior

$$\sigma_{n+1}^2 = \frac{1}{\beta} + \phi(x)^T\left(S_n^{-1} + \beta\phi_{n+1}\phi_{n+1}^T\right)^{-1}\phi(x)$$

Usando la identidad indicada en el enunciado tenemos

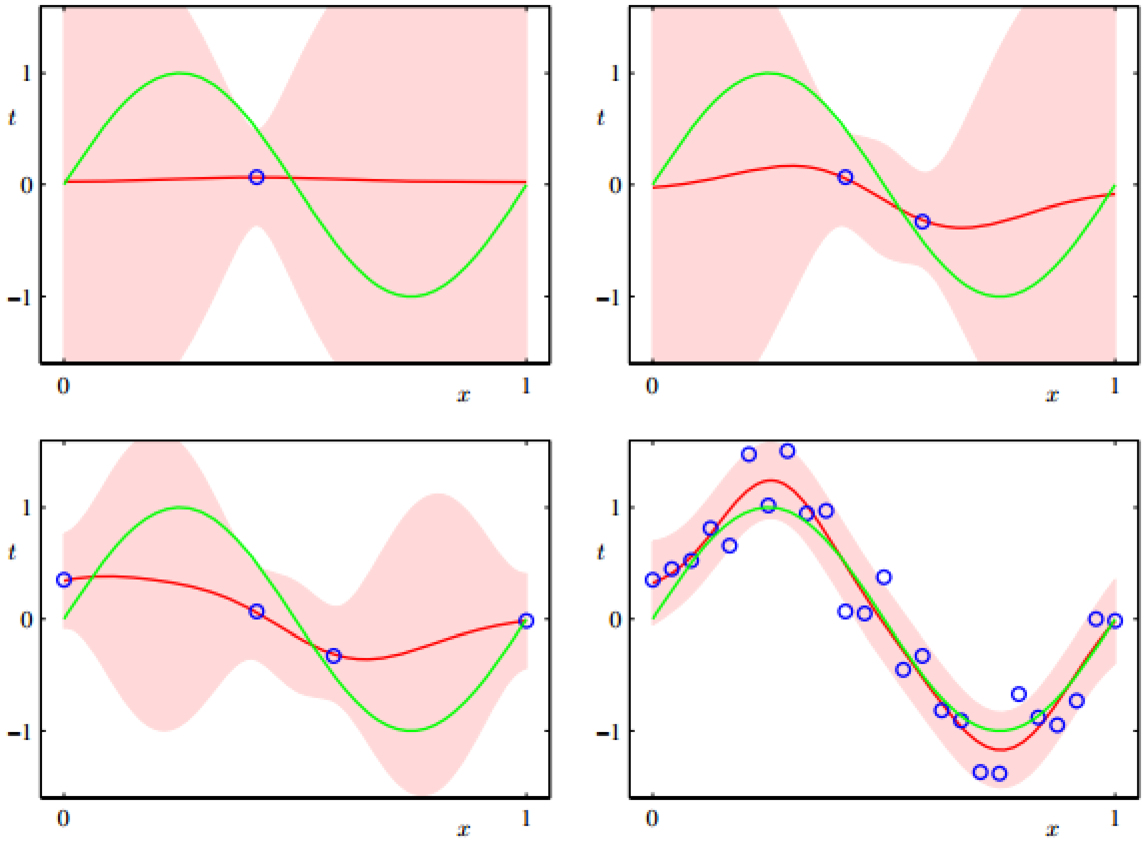
$$\sigma_{n+1}^2 = \frac{1}{\beta} + \phi(x)^T\left(S_n - \frac{\beta S_n\phi_{n+1}\phi_{n+1}^TS_n}{1 + \beta\phi_{n+1}^TS_n\phi_{n+1}}\right)\phi(x)$$

$$= \underbrace{\frac{1}{\beta} + \phi(x)^TS_n\phi(x)}_{\sigma_n^2} - \frac{\phi(x)^T\beta S_n\phi_{n+1}\phi_{n+1}^TS_n\phi(x)}{1 + \beta\phi_{n+1}^TS_n\phi_{n+1}}$$

$$= \sigma_n^2 - \frac{\phi(x)^T\beta S_n\phi_{n+1}\phi_{n+1}^TS_n\phi(x)}{1 + \beta\phi_{n+1}^TS_n\phi_{n+1}}$$

Ahora, dado que la matriz de covarianzas es positiva semidefinida $v^T S v \geq 0 \, \forall v$ sabemos que tanto en el numerador como en el denominador tendremos reales positivos. Esto implica que $\sigma_{n+1}^2 \leq \sigma_n^2$. Por cierto, la notación puede llevar a confusión, cuando escribimos ϕ_{n+1} nos referimos, tal como se hace en el libro de soluciones, a $\phi(x_{n+1})$.

Creo que con esto podemos sacar una medida de la incertidumbre asociada en cada punto, que sería lo que vemos en la figura 3.8, que pegamos aquí



3.4 Ejercicio 3.16

Para empezar tenemos que ver el prior para los parámetros $P(w)$. Lo identificamos con (2.113)

$$P(w) = N(w \mid 0, a^{-1}I)$$

Ahora identificamos (2.114) $P(t \mid w)$. En este caso la media serian los valores que obtendríamos con Φw_{ML} , por lo que identificamos $Ax + b = \Phi w$. ¿Cuál sería la incertidumbre? L^{-1}? Sería, en principio, el ruido en los datos que es lo que representa β^{-1}.

$$N(t \mid \Phi w, \beta^{-1}I)$$

Por tanto ya hemos identificado las variables necesarias. En nuestro caso sería

$$\begin{cases} x = w \\ y = t \\ A = \Phi \\ \mu = 0 \\ \Lambda^{-1} = a^{-1}I_m & \text{siendo m el número de parámetros} \\ L^{-1} = \beta^{-1}I_n & \text{siendo n el número de puntos del data set} \end{cases}$$

De (2.115) vemos que

$$p(t \mid a, \beta) = N(t, \underbrace{0}_{\mu}, \underbrace{\beta^{-1} I_n + a^{-1}\Phi\Phi^T}_{L^{-1}+A\Lambda^{-1}A^T})$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}} \frac{1}{|\beta^{-1} I_n + a^{-1}\Phi\Phi^T|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(t-0)^T \left(\beta^{-1} I_n + a^{-1}\Phi\Phi^T\right)^{-1} (t-0)\right)$$

Tomamos el logaritmo

$$\ln p(t \mid a, \beta) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|\beta^{-1} I_n + a^{-1}\Phi\Phi^T| - \frac{1}{2}t^T\left(\beta^{-1} I_n + a^{-1}\Phi\Phi^T\right)^{-1} t$$

Para el determinante hay que usar C.14 que dice que $|I_n + AB^T| = |I_m + A^T B|$ siendo A y B matrices de NxM. Sabiendo que $det(cA) = c^N det(A)$

$$|\beta^{-1} I_n + a^{-1}\Phi\Phi^T| = \beta^{-N}|I_n + a^{-1}\beta^N\Phi\Phi^T|$$

Ahora usamos C.14

$$= \beta^{-N}|I_m + a^{-1}\beta^N\Phi^T\Phi| = \beta^{-N}a^{-M}|aI_m + \beta\Phi^T\Phi|$$

Si nos fijamos en (3.81) vemos que podemos sustituir $|aI_m + \beta\Phi^T\Phi|$ por $|A|$. Esto hace que nos quede la expresión

$$\ln p(t \mid a, \beta) = -\frac{N}{2}\ln 2\pi - \frac{1}{2}\ln\left(a^{-M}\beta^{-N}|A|\right) - \frac{1}{2}t^T\left(\beta^{-1} I_n + a^{-1}\Phi\Phi^T\right)^{-1} t$$

$$= -\frac{N}{2}\ln\beta - \frac{M}{2}a - \frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|A| - \frac{1}{2}t^T\left(\beta^{-1} I_n + a^{-1}\Phi\Phi^T\right)^{-1} t$$

Ahora vamos a usar la identidad de C.7 para la parte que nos queda, la cual nos dice que $\left(A + BD^{-1}C\right)^{-1} = A^{-1} - A^{-1}B\left(D + CA^{-1}B\right)^{-1} CA^{-1}$, tomando $D^{-1} = \alpha^{-1}I_m$, $A = \beta^{-1}I_n$, $B = \Phi$, $C = \Phi^T$ y nos queda

$$\beta I_n - \beta I_n \Phi \underbrace{\left(\alpha I_m + \Phi^T\beta I_n\Phi\right)^{-1}}_{=A(3.81)} \Phi^T\beta I_n$$

$$\beta I_n - \beta\Phi A^{-1}\Phi^T\beta I_n$$

Si lo juntamos

$$-\frac{\beta}{2}t^T t - \frac{\beta^2}{2}t^T\Phi A^{-1}\Phi^T I_n$$

Usando (3.84) $m_N = \beta A^{-1}\Phi^T t$

$$-\frac{\beta}{2}t^T t + \frac{1}{2}m_N^T A m_N$$

Del ejercicio (3.18) sabemos que esto último

$$-\frac{\beta}{2}t^T t + \frac{1}{2}m_N^T A m_N = \frac{\beta}{2}||t - \Phi m_N||^2 + \frac{\alpha}{2}m_N^T m_N$$

si lo juntamos todo

$$\ln p(t \mid a, \beta) = -\frac{N}{2}\ln\beta - \frac{M}{2}a - \frac{N}{2}\ln 2\pi - \frac{1}{2}\ln|A| - \underbrace{\frac{\beta}{2}||t - \Phi m_N||^2 + \frac{\alpha}{2}m_N^T m_N}_{E(m_N)}$$

que es la expresión (3.86)

4 Ejercicios tema 4 del libro

4.1 Ejercicio 4.1

Sea convex hull

$$x = \sum_n a_n x_n \mid \left(\sum_n a_n = 1 \wedge a_n \geq 1 \right)$$

Consideremos $\{y_n\}$ con su correspondiente 'convex hull'.

$$y = \sum_n b_n y_n \mid \left(\sum_n b_n = 1 \wedge b_n \geq 1 \right)$$

Por definición, los 2 conjuntos de puntos son linealmente separables si existe $\hat{w}$ y un escalar w_0 tal que $\forall x_n \forall y_m \left(\hat{w}^T x_n + w_0 > 0 \wedge \hat{w}^T y_m + w_0 < 0 \right)$

Demostrar que si sus convex null intersectan entonces los 2 conjuntos de puntos no pueden ser linealmente separables y si son linealmente separables sus convex null no pueden intersectarse.

Si intersectan, entonces existe un punto z en el que : $z \in \sum_i a_i x_i \wedge z \in \sum_j b_j y_j$. De la condición de que sean linealmente separables sabemos que $p(x) = \hat{w}^T x + w_0 > 0 \forall x$, dado que a_i es no negativo y $a_i \in [0,1]$ $\hat{w}^T a_i x_i + w_0 > 0$

Para y, lo mismo $q(y) = \hat{w}^T b_j y + w_0 < 0 \forall y$

De ello sabemos que

$$\sum_i \left(\hat{w}^T a_i x_i \right) + w_0 > 0$$

y

$$\sum_j \left(\hat{w}^T b_j y_j \right) + w_0 < 0$$

Como $\sum a_i = 1$ y $\sum b_j = 1$ pues

$$\sum_i a_i \left(\hat{w}^T x_i + w_0 \right) > 0$$

$$\sum_j b_j \left(\hat{w}^T y_j + w_0 \right) < 0$$

Dado que en algun punto tienen que intersectar, $p(x) = q(y)$, luego

$$\sum_i a_i \left(\hat{w}^T x_i + w_0 \right) = \sum_j b_j \left(\hat{w}^T y_j + w_0 \right)$$

Esto sería una contradicción dado que tienen que cumplir simultaneamente ser mayor y menor que 0.

4.2 Ejercicio 4.10

Considerando el modelo de clasificación del ejercicio 4.9 y suponiendo que las densidades clase-condicionales vienen dadas por gausianas con covarianza compartida tal que

$$p(\phi \mid C_k) = N(\phi \mid \mu_k, \Sigma)$$

Demostrar que la solución de máxima verosimilitud para la media para la clase C_k viene dada por

$$\mu_k = \frac{1}{N_k} \sum (t_{nk} \phi_n)$$

que representa la media de los vectores de 'features' asignados a la clase C_k.

De forma similar demostrar que la solución de máxima verosimilitud para la covarianza compartida viene dada por

$$\Sigma = \sum_k^K \left(\frac{N_k}{N} S_k \right)$$

donde

$$S_k = \frac{1}{N_k} \sum_n \left(t_{nk} (\phi_n - \mu_k) (\phi_n - \mu_k)^T \right)$$

Por tanto tenemos un modelo de K clases, siendo ϕ el vector de 'features', un data set $\{\phi_i, t_i\}$, $t_i = [t_1, \ldots, t_k]$. El esquema de codificación es '1-of-K' por lo que $t_{ij} = I_{jk} \longleftrightarrow$ el patron i pertenece a la clase k.

$\pi_k = \dfrac{N_k}{N}$

Tomamos $p(\phi \mid C_k) = N(\phi \mid \mu_k, \Sigma)$ y usamos la definición de la gausiana multivariable para obtener

$$p(\phi \mid C_k) = \frac{1}{(2\pi)^{\frac{D}{2}}} \frac{1}{|\Sigma|^{\frac{1}{2}}} exp \left(-\frac{1}{2} (\phi - \mu_k)^T \Sigma^{-1} (\phi - \mu_k) \right)$$

La función de verosimilitud viene dada por

$$p(t \mid \pi_k) = \prod_n \prod_k (p(\phi_n \mid C_k) \pi_k)^{t_{nk}}$$

tomando el logaritmo

$$\ln p(t \mid \pi_k) = \sum_n \sum_k t_{nk} (\ln p(\phi_n \mid C_k) + \ln \pi_k)$$

Sustituyendo

$$\ln p(t \mid \pi_k) = \sum_n \sum_k t_{nk} (\ln N(\phi_n \mid \mu_k, \Sigma) + \ln \pi_k)$$

$$= \sum_n \sum_k t_{nk} \left(-\frac{D}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma| - \frac{1}{2} \left((\phi_n - \mu_k)^T \Sigma^{-1} (\phi_n - \mu_k) \right) + \ln \pi_k \right)$$

Sabiendo que la derivada de la forma cuadrática

$$\frac{\partial}{\partial s} (x - s)^T W (x - s) = -2W (x - s)$$

Derivando respecto a μ_k e igualando a 0 tenemos - usando

$$0 = \sum_n t_{nk} \Sigma^{-1} (\phi_n - \mu_k)$$

Reordenamos sabiendo que $\sum_n t_{nk} = N_k$

$$\sum_n t_{nk}\phi_n = \mu_k \sum_n t_{nk}$$
$$\frac{1}{N_k}\sum_n t_{nk}\phi_n = \mu_k$$

Que es la respuesta que buscábamos.
Para Σ procedemos de la misma forma. Derivamos respecto a Σ^{-1} e igualamos a 0.
Reescribiendo

$$(\phi_n - \mu_k)^T \Sigma^{-1} (\phi_n - \mu_k) = Tr\left[\Sigma^{-1} (\phi_n - \mu_k)^T (\phi_n - \mu_k)\right]$$

Tenemos

$$\sum_n \sum_k t_{nk} \left(-\frac{1}{2}\ln|\Sigma| - \frac{1}{2}Tr\left[\Sigma^{-1} (\phi_n - \mu_k)^T (\phi_n - \mu_k)\right] + \ln \pi_k\right) = 0$$
$$-\frac{1}{2}\sum_n \sum_k t_{nk} \left(\ln|\Sigma| + Tr\left[\Sigma^{-1} (\phi_n - \mu_k)^T (\phi_n - \mu_k)\right] + \ln \pi_k\right) = 0$$

Ahora podemos usar C.28 para $\ln|\Sigma|$ y C.24

$$-\frac{1}{2}\sum_n \sum_k t_{nk} \left(-\Sigma + (\phi_n - \mu_k)(\phi_n - \mu_k)^T\right) = 0$$
$$\frac{1}{2}\Sigma \underbrace{\sum_n \underbrace{\sum_k (t_{nk})}_{1}}_{N} - \frac{1}{2}\sum_n \sum_k t_{nk} \left((\phi_n - \mu_k)(\phi_n - \mu_k)^T\right) = 0$$
$$\frac{N}{2}\Sigma - \frac{1}{2}\sum_n \sum_k t_{nk} \left((\phi_n - \mu_k)(\phi_n - \mu_k)^T\right) = 0$$
$$\Sigma = \frac{1}{N}\sum_n \sum_k t_{nk} \left((\phi_n - \mu_k)(\phi_n - \mu_k)^T\right)$$

Que es la misma expresión que buscamos para la matriz de covarianza compartida, multiplicando y dividiendo por N_k.