Machine Learning 2 - Homework Week 1

Minh Ngo
MSc Artificial Intelligence
University of Amsterdam
nlminhtl@gmail.com

April 5, 2016

Collaborators: Ke Tran, Arthur Bražinskas, Riaan Zoetmulder

Problem 2

2.1

Likelihood of data is computing as following:

$$\begin{aligned}
p(X|\mu,\sigma^2) &= \prod_{n=1}^{N} p(x_n|\mu,\sigma^2) = \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(- \frac{(x_n-\mu)^2}{2\sigma^2} \right) \\
&= \left(\prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \right) \exp\left(- \sum_{n=1}^{N} \frac{(x_n-\mu)^2}{2\sigma^2} \right) \\
&= K_1 \exp\left(- \frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n^2 - 2x_n\mu + \mu^2) \right) \\
&= K_2 \exp\left(- \frac{1}{2\sigma^2} \sum_{n=1}^{N} (\mu^2 - 2x_n\mu) \right) \\
&= K_2 \exp\left(- \frac{N\mu^2 - 2\mu \sum_{n=1}^{N} x_n}{2\sigma^2} \right)
\end{aligned} \tag{1}$$

where:

$$
\begin{aligned}
K_1 &= \prod_{n=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \\
K_2 &= K_1 \exp\Bigg(-\frac{1}{2\sigma^2}\sum_{n=1}^{N} x_n^2 \Bigg)
\end{aligned}
\tag{2}
$$

2.2

Next posterior can be computed via the Bayes theorem:

$$
p(\mu|X,\sigma,\mu_0,\sigma_0^2) = \frac{p(X|\mu,\sigma^2)p(\mu|\mu_0,\sigma_0^2)}{p(X)}
\tag{3}
$$

2.3

$$
\begin{aligned}
p(\mu|X,\sigma,\mu_0,\sigma_0^2) &= \frac{K_2}{p(X)} \exp\Bigg(-\frac{N\mu^2 - 2\mu\sum_{n=1}^{N} x_n}{2\sigma^2}\Bigg) \frac{1}{\sigma\sqrt{2\pi}} \exp\Bigg(-\frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu^2 - 2\mu\mu_0}{2\sigma_0^2}\Bigg) \\
&= K_3 \exp\Bigg[-\mu^2\Bigg(\frac{N}{2\sigma^2} + \frac{1}{2\sigma_0^2}\Bigg) + \mu\Bigg(\frac{\sum_{n=1}^{N} x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\Bigg)\Bigg] \\
&= K_3 \exp\Bigg[-\mu^2\frac{1}{\frac{2\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}} - 2\mu\Bigg(\frac{\sum_{n=1}^{N} x_n}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\Bigg)\frac{\frac{\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}}{\frac{2\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}}\Bigg] \\
&= K_3 \exp\Bigg[-\mu^2\frac{1}{\frac{2\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}} - 2\mu\frac{\frac{\sigma_0^2\sum_{n=1}^{N} x_n + \mu_0\sigma^2}{N\sigma_0^2+\sigma^2}}{\frac{2\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}}\Bigg] \\
&= K_4 \exp\Bigg[-\frac{(\mu - \frac{\sigma_0^2\sum_{n=1}^{N} x_n + \mu_0\sigma^2}{N\sigma_0^2+\sigma^2})^2}{\frac{2\sigma^2\sigma_0^2}{N\sigma_0^2+\sigma^2}}\Bigg]
\end{aligned}
\tag{4}
$$

where:

$$
\begin{aligned}
K_3 &= \frac{K_2}{p(X)}\frac{1}{\sigma\sqrt{2\pi}} \exp\Bigg(-\frac{\mu_0^2}{2\sigma_0^2}\Bigg) \\
K_4 &= K_3 \exp\Bigg[\frac{(\sigma_0^2\sum_{n=1}^{N} x_n + \mu_0\sigma^2)^2}{2\sigma^2\sigma_0^2(N\sigma_0^2+\sigma^2)}\Bigg]
\end{aligned}
\tag{5}
$$

Consequently:

$$\begin{aligned}
\mu_N &= \frac{\sigma_0^2 \sum_{n=1}^N x_n + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2} \\
\sigma_N^2 &= \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2}
\end{aligned} \tag{6}$$

2.5

From those results we can derive sequential update of μ_N and σ_N^2:

$$\begin{aligned}
\mu_N &= \frac{\sigma_0^2 \sum_{n=1}^N x_{n-1}}{N\sigma_0^2 + \sigma^2} + \frac{\sigma_0^2 x_n}{N\sigma_0^2 + \sigma^2} + \frac{\sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2} \\
&= \frac{(N-1)\sigma_0^2 + \sigma^2}{N\sigma_0^2 + \sigma^2} \mu_{N-1} + \frac{\sigma_0^2 x_n}{N\sigma_0^2 + \sigma^2}
\end{aligned} \tag{7}$$

$$\begin{aligned}
\sigma_N^2 &= \frac{\sigma^2 \sigma_0^2}{N\sigma_0^2 + \sigma^2} = \frac{1}{\dfrac{N}{\sigma^2} + \dfrac{1}{\sigma_0^2}} = \frac{1}{\dfrac{N-1}{\sigma^2} + \dfrac{1}{\sigma_0^2} + \dfrac{1}{\sigma^2}} \\
&= \frac{1}{\dfrac{1}{\sigma_{N-1}^2} + \dfrac{1}{\sigma^2}} = \frac{\sigma_{N-1}^2 \sigma^2}{\sigma^2 + \sigma_{N-1}^2}
\end{aligned} \tag{8}$$

2.6

The same result can be derived from the posterior distribution with completing the square:

$$\begin{aligned}
p(\mu|x_1, .., x_N) &= p(x_N|\mu, \sigma) p(\mu|x_1, .., x_{N-1}) \\
&= K_5 \exp\left(- \frac{(x_n - \mu)^2}{2\sigma^2} \right) \exp\left(- \frac{(\mu - \mu_{N-1})^2}{2\sigma_{N-1}^2} \right) \\
&= K_6 \exp\left(- \frac{1}{2}\mu^2 (\frac{1}{\sigma^2} + \frac{1}{\sigma_{N-1}^2}) + \mu(\frac{x_n}{\sigma^2} + \frac{\mu_{N-1}}{\sigma_{N-1}^2}) \right) \\
&= K_7 \exp\left(- \frac{1}{2} \frac{(\mu - (\frac{x_n}{\sigma^2} + \frac{\mu_{N-1}}{\sigma_{N-1}^2}) \frac{\sigma_{N-1}^2 \sigma^2}{\sigma^2 + \sigma_{N-1}^2})^2}{\frac{\sigma_{N-1}^2 \sigma^2}{\sigma^2 + \sigma_{N-1}^2}} \right)
\end{aligned} \tag{9}$$

where K_5, K_6, K_7 are constants. Consequently:

$$\begin{aligned}
\sigma_N^2 &= \frac{\sigma_{N-1}^2\sigma^2}{\sigma^2+\sigma_{N-1}^2} \\
\mu_N &= (\frac{x_n}{\sigma^2}+\frac{\mu_{N-1}}{\sigma_{N-1}^2})\frac{\sigma_{N-1}^2\sigma^2}{\sigma^2+\sigma_{N-1}^2} \\
&= \frac{x_n\sigma_{N-1}^2+\mu_{N-1}\sigma^2}{\sigma^2+\sigma_{N-1}^2}
\end{aligned} \tag{10}$$

We could leave an expression dependent on σ_{N-1}^2 as a sequential update. The further derivation is just for proving that results obtained here and in 2.5 are equal. Knowing that $\sigma_{N-1}^2 = \dfrac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}$ we can derive that:

$$\begin{aligned}
\mu_N &= \frac{x_n\dfrac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}+\mu_{N-1}\sigma^2}{\sigma^2+\dfrac{\sigma^2\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}} \\
&= \frac{x_n\dfrac{\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}+\mu_{N-1}}{1+\dfrac{\sigma_0^2}{(N-1)\sigma_0^2+\sigma^2}} \\
&= \frac{(N-1)\sigma_0^2+\sigma^2}{N\sigma_0^2+\sigma^2}\mu_{N-1}+\frac{\sigma_0^2x_n}{N\sigma_0^2+\sigma^2} \qquad q.e.d.
\end{aligned} \tag{11}$$

Problem 3

3.1

Likelihood of the data is:

$$
\begin{aligned}
p(X|\mu,\Sigma) = \prod_{n=1}^{N} p(x_n|\mu,\Sigma) &= C_1 \exp\Bigg(-\frac{1}{2}\sum_{n=1}^{N}(x_n-\mu)^T\Sigma^{-1}(x_n-\mu)\Bigg) \\
&= C_2 \exp\Bigg(-\frac{1}{2}\sum_{n=1}^{N}\mu^T\Sigma^{-1}\mu - 2x_n^T\S\sum_{n=1}^{N} igma^{-1}\mu\Bigg) \\
&= C_2 \exp\Bigg(-\frac{1}{2}(N\mu^T\Sigma^{-1}\mu - 2\sum_{n=1}^{N}x_n^T\Sigma^{-1}\mu)\Bigg)
\end{aligned} \tag{12}
$$

where C_1 and C_2 are constants.

3.2

We will use proportional sign to simplify derivation since a normalization term is not required to derive. Corresponding posterior distribution is:

$$
\begin{aligned}
p(\mu|X,\Sigma,\mu_0,\Sigma_0) &= \frac{p(\mu)p(X|\mu,\Sigma)}{p(X)} \\
&\propto p(\mu)\exp\Bigg(-\frac{1}{2}(N\mu^T\Sigma^{-1}\mu - 2\sum_{n=1}^{N}x_n^T\Sigma^{-1}\mu)\Bigg) \\
&\propto \exp\Bigg(-\frac{1}{2}(\mu^T\Sigma_0^{-1}\mu - 2\mu^T\Sigma_0^{-1}\mu_0)\Bigg)\exp\Bigg(-\frac{1}{2}(N\mu^T\Sigma^{-1}\mu - 2\sum_{n=1}^{N}x_n^T\Sigma^{-1}\mu)\Bigg) \\
&\propto \exp\Bigg[-\frac{1}{2}\Bigg(\mu^T(\Sigma_0^{-1}+N\Sigma^{-1})\mu - 2\mu^T(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{n=1}^{N}x_n\Bigg)\Bigg]
\end{aligned} \tag{13}
$$

3.3

$$
\begin{aligned}
p(\mu|X,\Sigma,\mu_0,\Sigma_0) &\propto \exp\Bigg[-\frac{1}{2}\Bigg(\mu^T\overbrace{(\Sigma_0^{-1}+N\Sigma^{-1})}^{\Sigma_N^{-1}}\mu - 2\mu^T(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{n=1}^{N}x_n\Bigg)\Bigg] \\
&\propto \exp\Bigg[-\frac{1}{2}\Bigg(\mu^T\Sigma_N^{-1}\mu - 2\mu^T\Sigma_N^{-1}\underbrace{\Sigma_N(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{n=1}^{N}x_n}_{\mu_N}\Bigg)\Bigg]
\end{aligned} \tag{14}
$$

Now we can complete a square:

$$p(\mu|X,\Sigma,\mu_0,\Sigma_0) \propto \exp\Bigg[-\frac{1}{2}(\mu-\mu_N)^T\Sigma_N^{-1}(\mu-\mu_N)\Bigg] \qquad q.e.d. \tag{15}$$

3.4

$$\Sigma_N^{-1} = \Sigma_0^{-1} + N\Sigma^{-1} \tag{16}$$

$$\mu_N = \Sigma_N(\Sigma_0^{-1}\mu_0 + \Sigma^{-1}\sum_{n=1}^{N} x_n) \tag{17}$$

Problem 4

4.1

$$\mathcal{N}(x|a,A) = \frac{1}{(2\pi)^{D/2}|A|^{1/2}}\exp(-\frac{1}{2}(x-a)^TA^{-1}(x-a)) \tag{18}$$

$$\mathcal{N}(x|b,B) = \frac{1}{(2\pi)^{D/2}|B|^{1/2}}\exp(-\frac{1}{2}(x-b)^TB^{-1}(x-b)) \tag{19}$$

$$\begin{aligned}
\mathcal{N}(x|a,A)\mathcal{N}(x|b,B) &= \frac{1}{(2\pi)^{D}|AB|^{1/2}}\exp\Bigg(-\frac{1}{2}(a^TA^{-1}a + b^TB^{-1}b)\Bigg)\times \\
&\times \exp\Bigg[-\frac{1}{2}\Bigg(x^T(A^{-1}+B^{-1})x + 2x^T(B^{-1}b + A^{-1}a)\Bigg)\Bigg] \\
&= C_1\exp\Bigg[-\frac{1}{2}\Bigg(x^TC^{-1}x + 2x^T(B^{-1}b + A^{-1}a)\Bigg)\Bigg] \\
&= C_1\exp\Bigg[-\frac{1}{2}\Bigg(x^TC^{-1}x + 2x^TC^{-1}C(B^{-1}b + A^{-1}a)\Bigg)\Bigg] \\
&= C_1\exp\Bigg[-\frac{1}{2}(x-c)^TC^{-1}(x-c)\Bigg]\exp\Bigg[\frac{1}{2}c^TC^{-1}c\Bigg] \\
&= C_2\exp\Bigg[-\frac{1}{2}(x-c)^TC^{-1}(x-c)\Bigg] \\
&= C_2\frac{(2\pi)^{D/2}}{|C^{-1}|^{1/2}}\frac{1}{(2\pi)^{D/2}|C|^{1/2}}\exp\Bigg[-\frac{1}{2}(x-c)^TC^{-1}(x-c)\Bigg] \\
&= K^{-1}\mathcal{N}(x|c,C) \qquad q.e.d.
\end{aligned} \tag{20}$$

where:

$$\begin{aligned}
C &= (A^{-1} + B^{-1})^{-1} \\
c &= C(A^{-1}a + B^{-1}b) \\
C_1 &= \frac{1}{(2\pi)^D |AB|^{1/2}} \exp\left(-\frac{1}{2}(a^T A^{-1} a + b^T B^{-1} b) \right) \\
C_2 &= C_1 \exp\left(\frac{1}{2} c^T C^{-1} c \right) \\
K^{-1} &= C_2 \frac{(2\pi)^{D/2}}{|C^{-1}|^{1/2}}
\end{aligned} \tag{21}$$

4.2

Matrix Cookbook (156):

$$\begin{aligned}
C &= (A^{-1} + B^{-1})^{-1} = (A^{-1} + IB^{-1}I^T)^{-1} = \\
&= A - AI(B + I^T AI)^{-1} I^T A = A - A(B + A)^{-1} A
\end{aligned} \tag{22}$$

Applying the same trick by setting $A^{-1} = IA^{-1}I^T$ and then using the Woodbury Matrix Identity formula we get:

$$C = (B^{-1} + IA^{-1}I^T)^{-1} = B - B(A + B)^{-1} B \qquad q.e.d. \tag{23}$$

4.3

$$K^{-1} = \overbrace{\frac{1}{(2\pi)^{D/2}} \frac{1}{|ABC^{-1}|^{1/2}}}^{E_1} \exp\left(-\frac{1}{2}(\overbrace{a^T A^{-1} a + b^T B^{-1} b - c^T C^{-1} c}^{E_2}) \right) \tag{24}$$

$$\begin{aligned}
E_1 &= \frac{1}{(2\pi)^{D/2}} \frac{1}{|AB(A^{-1} + B^{-1})|^{1/2}} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|ABA^{-1} + ABB^{-1}|^{1/2}} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|AA^{-1}B + ABB^{-1}|^{1/2}} \\
&= \frac{1}{(2\pi)^{D/2}} \frac{1}{|B + A|^{1/2}}
\end{aligned} \tag{25}$$

$$
\begin{aligned}
E_2 &= a^T A^{-1} a + b^T B^{-1} b - C(A^{-1}a + B^{-1}b)^T C^{-1} C(A^1 a + B^{-1} b) \\
&= a^T A^{-1} a + b^T B^{-1} b - (A^{-1}a + B^{-1}b)^T C(A^{-1}a + B^{-1}b) \\
&= a^T A^{-1} a + b^T B^{-1} b - a^T A^{-1} C A^{-1} a - \\
&- a^T A^{-1} C B^{-1} b - b^T B^{-1} C A^{-1} a - b^T B^{-1} C B^{-1} b \\
&= a^T A^{-1} a + b^T B^{-1} b - \\
&- a^T A^{-1}(A - A(A+B)^{-1}A)A^{-1}a - \\
&- a^T A^{-1}(A - A(A+B)^{-1}A)B^{-1}b - \\
&- b^T B^{-1}(A - A(A+B)^{-1}A)A^{-1}a - \\
&- b^T B^{-1}(B - B(A+B)^{-1}B)B^{-1}b \\
&= a^T A^{-1} a + b^T B^{-1} b - a^T A^{-1} a + a^T (A+B)^{-1} a - a^T B^{-1} b + a^T (A+B)^{-1} A B^{-1} b - \\
&- b^T B^{-1} a + b^T B^{-1} A (A+B)^{-1} a - b^T B^{-1} b + b^T (A+B)^{-1} b \\
&= a^T (A+B)^{-1} a + b^T (A+B)^{-1} b - 2a^T B^{-1} b + 2a^T (A+B)^{-1} A B^{-1} b \\
&= a^T (A+B)^{-1} a + b^T (A+B)^{-1} b - 2a^T (I - (A+B)^{-1} A) B^{-1} b \\
&= a^T (A+B)^{-1} a + b^T (A+B)^{-1} b - 2a^T (B^{-1} - (A+B)^{-1} A B^{-1}) b
\end{aligned}
\tag{26}
$$

Woodbury matrix identity:

$$
\begin{aligned}
B^{-1} - (A+B)^{-1} A B^{-1} &= B^{-1} - B^{-1} B (A + B B^{-1} B)^{-1} A B^{-1} \\
&= (B + B B^{-1} A)^{-1} = (B + A)^{-1}
\end{aligned}
\tag{27}
$$

Consequently:

$$
E_2 = (a-b)^T (A+B)^{-1} (a-b) \tag{28}
$$

Combining everything together:

$$
K^{-1} = \frac{1}{(2\pi)^{D/2}} \frac{1}{|A+B|^{1/2}} \exp\left(-\frac{1}{2}(a-b)^T (A+B)^{-1} (a-b) \right) = \mathcal{N}(a|b, A+B) \qquad q.e.d. \tag{29}
$$

Problem 5

5.1

Bishop (2.8):

$$\mu_{MLE} = \frac{m}{N} = \frac{3}{3} = 1 \tag{30}$$

where m - number of times a head is observed, N - number of observation.

5.2

Bishop (2.18), (2.19), (2.20), [http://math.stackexchange.com/users/22857/martin argerami]:

$$p(\mu|m, l, a, b) = \frac{\Gamma(m + a + l + b)}{\Gamma(m + a)\Gamma(l + b)}\mu^{m+a-1}(1 - \mu)^{l+b-1} \tag{31}$$

$$\begin{aligned} p(x = 1|\mathcal{D}) &= \int_0^1 p(x = 1|\mu)p(\mu|\mathcal{D})d\mu = \int_0^1 \mu p(\mu|\mathcal{D})d\mu = E[\mu|\mathcal{D}] \\ &= \frac{\Gamma(m + a + l + b)}{\Gamma(m + a)\Gamma(l + b)} \int_0^1 \mu\mu^{m+a-1}(1 - \mu)^{l+b-1}d\mu \\ &= \frac{\Gamma(m + a + l + b)}{\Gamma(m + a)\Gamma(l + b)} \frac{\Gamma(m + 1)\Gamma(l + b - 1)}{\Gamma(l + b + m + a)} \\ &= \frac{m + a}{m + a + l + b} \end{aligned} \tag{32}$$

where l = N - m.
Consequently, μ_{MAP} in our case will be:

$$\mu_{MAP} = \frac{3 + a}{3 + a + b} \tag{33}$$

5.3

$$E[\mu] = \frac{a}{a + b} \tag{34}$$

$$\mu_{MLE} = \frac{m}{m + l} \tag{35}$$

$$\mu_{MAP} = \frac{a + m}{a + b + m + l} \tag{36}$$

To prove that μ_{MLE} lies in between $E[\mu]$ and μ_{MAP} we have to prove that:

$$\lambda E[\mu] + (1-\lambda)\mu_{MLE} = \mu_{MAP} \tag{37}$$

for some $0 < \lambda < 1$.

$$\begin{aligned}
\mu_{MAP} &= \frac{m}{m+a+l+b} + \frac{a}{m+a+l+b} \\
&= \frac{m+l}{m+l}\frac{m}{m+a+l+b} + \frac{a+b}{a+b}\frac{a}{m+a+l+b} \\
&= \frac{m+l}{m+a+l+b}\frac{m}{m+l} + \frac{a+b}{m+a+l+b}\frac{a}{a+b} \\
&= \underbrace{\frac{m+l}{m+a+l+b}}_{1-\lambda}\mu_{MLE} + \underbrace{\frac{a+b}{m+a+l+b}}_{\lambda}E[\mu] \\
&= \lambda E[\mu] + (1-\lambda)\mu_{MLE} \qquad q.e.d.
\end{aligned} \tag{38}$$

λ is between 0 and 1 because of the fact that a,b,m,l > 0.

Problem 6

The Student's T distribution is the following:

$$St(x|\mu,\Sigma,\nu) = \frac{\Gamma(\dfrac{\nu+p}{2})}{\Gamma(\dfrac{\nu}{2})(\nu\pi)^{\dfrac{p}{2}}\Sigma^{\dfrac{1}{2}}(1+\dfrac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu))^{\dfrac{\nu+p}{2}}} \tag{39}$$

$$E[X] = \int_{-\infty}^{\infty} \frac{Cx}{[1+\dfrac{1}{\nu}(x-\mu)^T\Sigma^{-1}(x-\mu)]^{\dfrac{\nu+p}{2}}}dx \tag{40}$$

where C is normalization constant.
Replacing $x = z + \mu$:

$$
\begin{aligned}
E[X] &= C\int_{-\infty}^{\infty} \frac{(z+\mu)}{[1+\frac{1}{\nu}z^T\Sigma^{-1}z]^{\frac{\nu+p}{2}}}dz \\
&= \underbrace{C\int_{-\infty}^{\infty} \frac{z}{[1+\frac{1}{\nu}z^T\Sigma^{-1}z]^{\frac{\nu+p}{2}}}dz}_{0} + \mu \underbrace{C\int_{-\infty}^{\infty} \frac{1}{[1+\frac{1}{\nu}z^T\Sigma^{-1}z]^{\frac{\nu+p}{2}}}dz}_{1} = \mu
\end{aligned} \tag{41}
$$

References

M. A. (http://math.stackexchange.com/users/22857/martin argerami). How to evaluate this integral? (relating to binomial). Mathematics Stack Exchange. URL `http://math.stackexchange.com/q/122302`. URL:http://math.stackexchange.com/q/122302 (version: 2015-01-02).