Machine Learning 2 — Homework 1

Maurice Frank
11650656
maurice.frank@posteo.de

September 9, 2019

Problem 1.

We have: $\boldsymbol{x} \in \mathbb{R}^n$, $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu_x}, \boldsymbol{\Sigma_x})$ and $\boldsymbol{z} \in \mathbb{R}^n$, $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}|\boldsymbol{\mu_z}, \boldsymbol{\Sigma_z})$. And:

$$
\begin{aligned}
\mathbf{y} &= \boldsymbol{x} + \boldsymbol{z} \\
&\implies \\
\mathbb{E}[\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x} + \boldsymbol{z}] \\
&= \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{z}] \\
&= \boldsymbol{\mu_x} + \boldsymbol{\mu_z} \\
\operatorname{cov}(\boldsymbol{y}) &= \mathbb{E}[(\boldsymbol{x} + \boldsymbol{y})^2] - (\mathbb{E}[\boldsymbol{x} + \boldsymbol{y}])^2 \\
&= \mathbb{E}[\boldsymbol{x}^2] + 2\mathbb{E}[\boldsymbol{x}\boldsymbol{y}] + \mathbb{E}[\boldsymbol{y}^2] - \mathbb{E}[\boldsymbol{x}]^2 - 2\mathbb{E}[\boldsymbol{x}]\mathbb{E}[\boldsymbol{y}] - \mathbb{E}[\boldsymbol{y}]^2 \\
&= \mathbb{E}[\boldsymbol{x}^2] - \mathbb{E}[\boldsymbol{x}]^2 + \mathbb{E}[\boldsymbol{y}^2] - \mathbb{E}[\boldsymbol{y}]^2 \\
&= \operatorname{cov}(\boldsymbol{x}) + \operatorname{cov}(\boldsymbol{y})
\end{aligned}
$$

Problem 2.

We have: $\boldsymbol{x} \in \mathbb{R}^D,\ p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and observe $\boldsymbol{\chi} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. Further it is $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathcal{N}\Sigma)$ and $\mu \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$.

1.

The likelihood is:

$$
\begin{aligned}
p(\boldsymbol{\chi}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \prod_{i=0}^{N} p(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&= \prod_{i=0}^{N} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&= \prod_{i=0}^{N} (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right)
\end{aligned}
$$

2.

Using Bayes we get the posterior:

$$
\begin{aligned}
p(\boldsymbol{\mu}|\boldsymbol{\chi},\boldsymbol{\Sigma},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) &= \frac{p(\boldsymbol{\chi}|\boldsymbol{\mu},\boldsymbol{\Sigma}) \cdot p(\boldsymbol{\mu}|\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)}{p(\boldsymbol{\chi})} \\
&\sim \prod_{i=0}^{N} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) \cdot \mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0)
\end{aligned}
$$

3.

$$
\begin{aligned}
p(\boldsymbol{\mu}|\boldsymbol{\chi},\boldsymbol{\Sigma},\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) &\sim \prod_{i=0}^{N} \mathcal{N}(\boldsymbol{x}_i|\boldsymbol{\mu},\boldsymbol{\Sigma}) \cdot \mathcal{N}(\boldsymbol{\mu}_0,\boldsymbol{\Sigma}_0) \\
&\sim \exp\left[-\frac{1}{2}(\boldsymbol{\mu}-\boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu}-\boldsymbol{\mu}_0) - \frac{1}{2}\sum_{i=0}^{N}(\boldsymbol{x}_i-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu})\right] \\
&\sim \exp\frac{1}{2}\left[-\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1} + N\cdot\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu} + 2\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=0}^{N}\boldsymbol{x}_i) - \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 - \sum_{i=0}^{N}\boldsymbol{x}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x}\right. \\
&\stackrel{\text{def}}{=} \exp\left[-\boldsymbol{\mu}^T\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu} + 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N - \boldsymbol{\mu}_N^T\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N\right] \\
&\implies \\
\Sigma_N^{-1} &= \boldsymbol{\Sigma}_0^{-1} + N\cdot\boldsymbol{\Sigma}^{-1} \\
&\wedge \\
2\boldsymbol{\mu}^T(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=0}^{N}\boldsymbol{x}_i) &= 2\boldsymbol{\mu}^T\boldsymbol{\Sigma}_N^{-1}\boldsymbol{\mu}_N \\
\mu_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=0}^{N}\boldsymbol{x}_i)
\end{aligned}
$$

With both parameters calculated we have the posterior:

$$\begin{aligned}
p(\boldsymbol{\mu}|\chi, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &= \mathcal{N}(\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N) \\
&= \mathcal{N}(\boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=0}^{N}\boldsymbol{x}_i), \frac{1}{\boldsymbol{\Sigma}_0^{-1} + N \cdot \boldsymbol{\Sigma}^{-1}})
\end{aligned}$$

4.

We are using the log posterior for the MAP estimate:

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\boldsymbol{\mu}|\chi, \boldsymbol{\Sigma}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) &= 0 \\
&= \frac{\partial}{\partial \boldsymbol{\mu}} - \frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=0}^{N}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \\
&\quad - \frac{D}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\
&= -\sum_{i=0}^{N}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) - \boldsymbol{\Sigma}_0^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_0) \\
&= -\boldsymbol{\Sigma}^{-1}\sum_{i=0}^{N}\boldsymbol{x}_i + N \cdot \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} - \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu} + \boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 \\
&\implies \\
\mu_{MAP} &= \frac{-\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \boldsymbol{\Sigma}^{-1}\sum_{i=0}^{N}\boldsymbol{x}_i}{N \cdot \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}_0^{-1}}
\end{aligned}$$

Problem 3.

We have a coin with head-proability of μ.

1.

For this Bernoulli experiment the MLE is:

$$\begin{aligned}
\mu_{ML} &= \frac{\sum_i^N x_i}{N} \\
&= 1
\end{aligned}$$

so the probability is 100 per cent.

2.

We assume $\mu \sim \text{Beta}(\mu|, a, b)$:

$$\begin{aligned}
p(\mu|\chi) &= p(\chi|\mu)p(\mu|a,b) \\
&= \prod_{i=0}^{N} \mu^{x_i}(1-\mu)^{(1-x_i)} \cdot \frac{\mu^{a-1}(1-\mu)^{b-1}}{B(a,b)} \\
&\implies \\
\ln p(\mu|\chi) &= \sum_{i=0}^{N} x_i \ln \mu + (1-x_i)\ln(1-\mu) + (a-1)\ln \mu + (b-1)\ln(1-\mu) + C \\
&\implies \\
\frac{\partial \ln p(\mu|\chi)}{\partial \mu} &= \sum_{i=0}^{N} \frac{x_i}{\mu} - \frac{1-x_i}{1-\mu} + \frac{a-1}{\mu} - \frac{b-1}{1-\mu} \\
&\implies \\
0 &= \frac{\sum x_i + a - 1}{\mu_{MAP}} - \frac{\sum(1-x_i) + b - 1}{1-\mu_{MAP}} \\
\frac{1}{\mu_{MAP}} &= \frac{\sum(1-x_i) + b - 1}{\sum x_i + a - 1} + 1 \\
\mu_{MAP} &= \frac{\sum x_i + a - 1}{N + a + b - 2}
\end{aligned}$$

for our case that is:

$$\mu_{MAP} = \frac{a+2}{a+b+1}$$

3.

Our posterior distribution is also a Beta distribution:

$$
\begin{aligned}
p(\mu|\chi, a, b) &= \frac{p(\chi|\mu) \cdot p(\mu|a,b)}{\int_0^1 p(\chi|\mu) \cdot p(\mu|a,b) d\mu} \\
&= \frac{\binom{m+l}{m}\mu^m(1-\mu)^l \cdot \frac{1}{B(a,b)}\mu^{a-1}(1-\mu)^{b-1}}{\int_0^1 \binom{m+l}{m}\mu^m(1-\mu)^l \cdot \frac{1}{B(a,b)}\mu^{a-1}(1-\mu)^{b-1} d\mu} \\
&= \frac{\mu^m(1-\mu)^l \cdot \mu^{a-1}(1-\mu)^{b-1}}{\int_0^1 \mu^m(1-\mu)^l \cdot \mu^{a-1}(1-\mu)^{b-1} d\mu} \\
&= \frac{\mu^{m+a-1}(1-\mu)^{l+b-1}}{\int_0^1 \mu^{m+a-1}(1-\mu)^{l+b-1} d\mu} \\
&= \frac{\mu^{m+a-1}(1-\mu)^{l+b-1}}{B(m+a, l+b)} \\
&= \text{Beta}(m+a, b+l)
\end{aligned}
$$

Our posterior distribution is a Beta distribution with the prev derived mean:

$$
\begin{aligned}
p(\mu|m, l, a, b) &= \text{Beta}(m+a, l+b) \\
&\implies \\
\mathbb{E}[p(\mu|m, l, a, b)] &= \frac{m+a}{m+l+a+b}
\end{aligned}
$$

For prior mean and ML estimate we have:

$$
\begin{aligned}
\mathbb{E}[\text{Beta}(\mu \,|\, \text{a,b})] &= \frac{a}{a+b} \\
\mu_{ML} &= \frac{m}{m+l}
\end{aligned}
$$

We can form an linear combination to give the posterior mean:

$$
\begin{aligned}
\mathbb{E}[p(\mu|m, l, a, b)] &= \frac{a}{m+l+a+b} + \frac{m}{m+l+a+b} \\
&= \frac{a+b}{a+b}\frac{a}{m+l+a+b} + \frac{m+l}{m+l}\frac{m}{m+l+a+b} \\
&= \mathbb{E}[\text{Beta}(\mu \,|\, \text{a,b})]\frac{a+b}{m+l+a+b} + \mu_{ML}\frac{m+l}{m+l+a+b}
\end{aligned}
$$

Given that $a+b > 0$ and $m+l > 0$ and thus both coefficient in $(0,1)$ this shows that the posterior mean is between those two estimates.

Problem 4.

Distributions in the exponential family have the form:

$$
p(\boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\phi}(\boldsymbol{\eta})\boldsymbol{T}(\boldsymbol{x}) - \boldsymbol{A}(\boldsymbol{\eta})\right)
$$

1.

(i)

$$\begin{aligned}
\mathrm{Pois}(k|\lambda) &= \frac{\lambda^k \exp -\lambda}{k!} \\
&= \frac{1}{k!} \cdot \exp \log \lambda^k \cdot \exp -\lambda \\
&= \frac{1}{k!} \cdot \exp\left(k \log \lambda - \lambda\right) \\
&\Longrightarrow \\
h(x) &= \frac{1}{k!} \\
\phi(\eta) &= \log \lambda \\
T(x) &= k && \text{(sufficient statistic)} \\
A(\eta) &= \lambda
\end{aligned}$$

(ii)

$$\begin{aligned}
\mathrm{Gam}(\tau|a,b) &= \frac{1}{\Gamma(a)} b^a \tau^{a-1} e^{-b\tau} \\
&= \exp\left(a \log b - \log \Gamma(a)\right) \cdot \exp\left((a-1)\log \tau\right) \cdot \exp\left(-b\tau\right) \\
&= \exp\left[(a-1)\log \tau - b\tau - (\log \Gamma(a) - a \log b)\right] \\
&\Longrightarrow \\
h(x) &= 1 \\
\phi(\eta) &= \begin{bmatrix} a-1 \\ -b \end{bmatrix} \\
T(x) &= \begin{bmatrix} \log \tau \\ \tau \end{bmatrix} && \text{(sufficient statistic)} \\
A(\eta) &= \log \Gamma(a) - a \log(b)
\end{aligned}$$

(iii)

$$\mathrm{Cauchy}(x|\gamma,\mu) = \frac{1}{\pi\gamma} \frac{1}{1 + \left(\frac{x-\mu}{\gamma}\right)^2}$$

Because of the $1+$ in the denominator the Cauchy distribution can not be factorized into a canonical form of the exponential family. Thus it is not part of it.

(iv)

$$
\begin{aligned}
\text{vonMises}(x|\kappa,\mu) &= \frac{1}{2\pi I_0(\kappa)} \exp\left(\kappa \cos\left(x-\mu\right)\right) \\
&= \exp\left(\cos\left(x-\mu\right)\kappa\right) \cdot \exp\left(-\log\left(2\pi I_0(\kappa)\right)\right) \\
&= \exp\left(\kappa(\cos\mu\cos x + \sin\mu\sin x) - \log\left(2\pi I_0(\kappa)\right)\right) \\
&\implies \\
h(x) &= 1 \\
\phi(\eta) &= \begin{bmatrix} \kappa\cos\mu \\ \kappa\sin\mu \end{bmatrix} \\
T(x) &= \begin{bmatrix} \cos x \\ \sin x \end{bmatrix} && \text{(sufficient statistic)} \\
A(\eta) &= \log\left(2\pi I_0(\kappa)\right)
\end{aligned}
$$

2.

(i)

$$
\begin{aligned}
\mathbb{E}[\text{Pois}(\text{k} \mid \lambda)] &= \sum_{k=0}^{\infty} k e^{-\lambda} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} k \cdot \frac{\lambda \cdot \lambda^{(k-1)}}{k \cdot (k-1)!} \\
&= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^{(k-1)}}{(k-1)!} \\
&= \lambda e^{-\lambda} \exp\lambda \\
&= \lambda
\end{aligned}
$$

$$
\begin{aligned}
\mathbb{E}[(\text{Pois}(k|\lambda) - \lambda)^2] &= e^{-\lambda} \sum_{k=0}^{\infty} (k-\lambda)^2 \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} (k^2 - 2k\lambda + \lambda^2) \frac{\lambda^k}{k!} \\
&= e^{-\lambda}((\lambda^2+\lambda)e^{\lambda} - 2\lambda^2 e^{\lambda} + \lambda^2 e^{\lambda}) \\
&= \lambda
\end{aligned}
$$

(ii)

$$\begin{aligned}
\mathbb{E}[\mathrm{Gam}(\tau|a,b)] &= \frac{b^a}{\Gamma(a)} \cdot \int_0^\infty \tau \cdot \tau^{a-1} e^{-b\tau} d\tau \\
&= \frac{b^a}{\Gamma(a)} \cdot \int_0^\infty \tau^a e^{-b\tau} d\tau \\
&= \frac{b^a}{\Gamma(a)} \cdot \int_0^\infty (\frac{\tau}{b})^a e^{-\tau} \frac{1}{b} d\frac{\tau}{b} \\
&= \frac{b^a}{\Gamma(a)} \cdot \frac{1}{b^{a+1}} \int_0^\infty \tau^a e^{-\tau} d\frac{\tau}{b} \\
&= \frac{1}{\Gamma(a)} \cdot \frac{1}{b} \cdot \Gamma(a+1) \\
&= \frac{1}{\Gamma(a)} \cdot \frac{1}{b} \cdot a \cdot \Gamma(a) \\
&= \frac{a}{b}
\end{aligned}$$

$$\begin{aligned}
\mathrm{Var}(\mathrm{Gam}(\tau|a,b)) &= \mathbb{E}[(\mathrm{Gam}(\tau|a,b) - \frac{a}{b})^2] \\
&= \frac{b^a}{\Gamma(a)} \cdot \int_0^\infty (\tau - \frac{a}{b})^2 \cdot \tau^{a-1} e^{-b\tau} d\tau \\
&= \frac{b^a}{\Gamma(a)} \cdot \Bigg[\int_0^\infty \tau^{a+2-1} e^{-\tau} \frac{1}{b^{a+2}} d\frac{\tau}{b} - 2\frac{a}{b} \int_0^\infty \tau^{a+1-1} e^{-\tau} \frac{1}{b^{a+1}} d\frac{\tau}{b} \\
&\qquad + (\frac{a}{b})^2 \int_0^\infty \tau^{a-1} e^{-\tau} \frac{1}{b^a} d\frac{\tau}{b} \Bigg] \\
&= \frac{b^a}{\Gamma(a)} \cdot \left[\frac{\Gamma(a+2)}{b^{a+2}} - 2\frac{a \cdot \Gamma(a+1)}{b \cdot b^{a+1}} + \frac{a^2 \cdot \Gamma(a)}{b^2 \cdot b^a} \right] \\
&= \frac{b^a}{\Gamma(a)} \cdot \left[\frac{(a^2+a)\Gamma(a)}{b^{a+2}} - 2\frac{a^2 \cdot \Gamma(a)}{b^{a+2}} + \frac{a^2 \cdot \Gamma(a)}{b^{a+2}} \right] \\
&= \frac{a}{b^2}
\end{aligned}$$

3.

Every member of the exponential family has a conjugate prior as such also the Poisson distribution. The conjugate prior to a Poisson is a Gamma which we can show by computing the posterior:

$$
\begin{aligned}
p(\lambda|k,a,b) &= \frac{p(k|\lambda)\cdot p(\lambda|a,b)}{\int p(k)} \\
&= \frac{\frac{\lambda^k e^{-\lambda}}{k!}\frac{1}{\Gamma(a)}b^a\lambda^{a-1}e^{-b\lambda}}{\int \frac{\lambda^{*k}e^{-\lambda^*}}{k!}\frac{1}{\Gamma(a)}b^a\lambda^{*a-1}e^{-b\lambda^*}d\lambda^*} \\
&= \frac{\lambda^{a+k-1}e^{-(b+1)\lambda}}{\int \lambda^{*a+k-1}e^{-(b+1)\lambda^*}d\lambda^*} \\
&= \frac{\lambda^{a+k-1}e^{-(b+1)\lambda}}{\int \left(\frac{\lambda^*}{b+1}\right)^{a+k-1}e^{-\lambda^*}\frac{1}{b+1}d\frac{\lambda^*}{b+1}} \\
&= \frac{\lambda^{a+k-1}e^{-(b+1)\lambda}}{\frac{1}{(b+1)^{a+k}}\int \lambda^{*a+k-1}e^{-\lambda^*}d\frac{\lambda^*}{b+1}} \\
&= \frac{(b+1)^{a+k}}{\Gamma(a+k)}\cdot\lambda^{a+k-1}e^{-(b+1)\lambda} \\
&= \mathrm{Gamma}(a+k,b+1)
\end{aligned}
$$