Machine Learning II - Homework 1

Sindy Löwe
(11594969)

Collaborators:
Pascal Esser, Gautier Dagan

Problem 1

$$
\begin{aligned}
\mathbb{E}[\boldsymbol{y}] &= \mathbb{E}[\boldsymbol{x}+\boldsymbol{z}] \\
&= \mathbb{E}[\boldsymbol{x}] + \mathbb{E}[\boldsymbol{z}] \\
&= \boldsymbol{\mu_x} + \boldsymbol{\mu_z}
\end{aligned}
$$

$$
\begin{aligned}
\text{var}[\boldsymbol{y}] &= \mathbb{E}[(\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^2] \\
&= \mathbb{E}[(\boldsymbol{x}+\boldsymbol{z} - \mathbb{E}[\boldsymbol{x}+\boldsymbol{z}])^2] \\
&= \mathbb{E}[(\boldsymbol{x}+\boldsymbol{z} - \boldsymbol{\mu_x} - \boldsymbol{\mu_z})^2] \\
&= \mathbb{E}[\boldsymbol{x}^2 + \boldsymbol{xz} - \boldsymbol{x\mu_x} - \boldsymbol{x\mu_z} + \boldsymbol{xz} + \boldsymbol{z}^2 - \boldsymbol{z\mu_x} - \boldsymbol{z\mu_z} - \boldsymbol{x\mu_x} - \boldsymbol{z\mu_x} + \boldsymbol{\mu_x}^2 \\
&\quad + \boldsymbol{\mu_x\mu_z} - \boldsymbol{x\mu_x} - \boldsymbol{z\mu_z} + \boldsymbol{\mu_x\mu_z} + \boldsymbol{\mu_z}^2] \\
&= \mathbb{E}[(\boldsymbol{x}+\boldsymbol{\mu_x})^2] + \mathbb{E}[(\boldsymbol{z}+\boldsymbol{\mu_z})^2] + 2\,\mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu_x})(\boldsymbol{z}-\boldsymbol{\mu_z})] \\
&= \boldsymbol{\Sigma_x} + \boldsymbol{\Sigma_z} + 2\,\text{cov}(\boldsymbol{x},\boldsymbol{z})
\end{aligned}
$$

If $\boldsymbol{x} \perp\!\!\!\perp \boldsymbol{z}$, then $\text{cov}(\boldsymbol{x},\boldsymbol{z}) = 0$ and therefore, $\text{var}[\boldsymbol{y}] = \boldsymbol{\Sigma_x} + \boldsymbol{\Sigma_z}$.

Problem 2

1. The likelihood of the data $p(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})$ is:

$$
\begin{aligned}
p(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \prod_{n=1}^{N} \mathcal{N}(x_n|\boldsymbol{\mu},\boldsymbol{\Sigma}) \\
&= \prod_{n=1}^{N} (2\pi)^{-\frac{D}{2}} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(x_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(x_n-\boldsymbol{\mu})\right] \\
&= (2\pi)^{-\frac{D\cdot N}{2}} \det(\boldsymbol{\Sigma})^{-\frac{N}{2}} \exp\left[-\frac{1}{2}\sum_{n=1}^{N}(x_n-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(x_n-\boldsymbol{\mu})\right]
\end{aligned}
$$

2. The posterior of the data $p(\boldsymbol{\mu}|\boldsymbol{X},\boldsymbol{\Sigma},\boldsymbol{\mu_0},\boldsymbol{\Sigma_0})$ is:

$$
\begin{aligned}
p(\boldsymbol{\mu}|\boldsymbol{X},\boldsymbol{\Sigma},\boldsymbol{\mu_0},\boldsymbol{\Sigma_0}) &= \frac{p(\boldsymbol{X}|\boldsymbol{\mu},\boldsymbol{\Sigma})\cdot p(\boldsymbol{\mu}|\boldsymbol{\mu_0},\boldsymbol{\Sigma_0})}{p(\boldsymbol{X}|\boldsymbol{\Sigma},\boldsymbol{\mu_0},\boldsymbol{\Sigma_0})} \\
&= \frac{\prod_{n=1}^{N}\mathcal{N}(x_n|\boldsymbol{\mu},\boldsymbol{\Sigma})\cdot\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0},\boldsymbol{\Sigma_0})}{p(\boldsymbol{X}|\boldsymbol{\Sigma},\boldsymbol{\mu_0},\boldsymbol{\Sigma_0})}
\end{aligned}
$$

3. For evaluating the posterior distribution, we can get rid of the divisor, since it does not depend on $\boldsymbol{\mu}$.

$$\begin{aligned}
p(\boldsymbol{\mu}|\boldsymbol{X}, \boldsymbol{\Sigma}, \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) &= \frac{\prod_{n=1}^{N} \mathcal{N}(x_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})}{p(\boldsymbol{X}|\boldsymbol{\Sigma}, \boldsymbol{\mu_0}, \boldsymbol{\Sigma_0})} \\
&\propto \prod_{n=1}^{N} \mathcal{N}(x_n|\boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_0}, \boldsymbol{\Sigma_0}) \\
&= (2\pi)^{-\frac{D \cdot N}{2}} \det(\boldsymbol{\Sigma})^{-\frac{N}{2}} \exp\left[-\frac{1}{2}\sum_{n=1}^{N}(x_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x_n - \boldsymbol{\mu})\right] \\
&\cdot (2\pi)^{-\frac{D}{2}} \det(\boldsymbol{\Sigma_0})^{-\frac{1}{2}} \exp\left[-\frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu_0})^T \boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu_0})\right] \\
&= (2\pi)^{\frac{-D(N+1)}{2}} \det(\boldsymbol{\Sigma})^{-\frac{N}{2}} \det(\boldsymbol{\Sigma_0})^{-\frac{1}{2}} \\
&\cdot \exp\left[-\frac{1}{2}\sum_{n=1}^{N}(x_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x_n - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu_0})^T \boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu_0})\right]
\end{aligned}$$

Thus, the posterior is proportional to an exponential of a quadratic form in $\boldsymbol{\mu}$. We can cast this exponential into the form of equation (2.71) in Bishop, in order to retrieve $\boldsymbol{\mu_N}$ and $\boldsymbol{\Sigma_N}$. Since $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma_0}$ are symmetric, we can use $\boldsymbol{a}^T\boldsymbol{C}\boldsymbol{b} = \boldsymbol{b}^T\boldsymbol{C}\boldsymbol{a}$.

$$\begin{aligned}
&-\frac{1}{2}\sum_{n=1}^{N}(x_n - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(x_n - \boldsymbol{\mu}) - \frac{1}{2}(\boldsymbol{\mu} - \boldsymbol{\mu_0})^T \boldsymbol{\Sigma_0}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu_0}) \\
&= \sum_{n=1}^{N} -\frac{1}{2} x_n \boldsymbol{\Sigma}^{-1} x_n + \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} x_n - \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \\
&- \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} - \frac{1}{2}\boldsymbol{\mu_0}^T \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} \\
&= -\frac{1}{2}\boldsymbol{\mu}^T(\boldsymbol{\Sigma_0}^{-1} + N\boldsymbol{\Sigma}^{-1})\boldsymbol{\mu} + \boldsymbol{\mu}^T(\boldsymbol{\Sigma}^{-1}\sum_{n=1}^{N} x_n + \boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0}) + \text{const}
\end{aligned}$$

Where const contains all terms that are independent of $\boldsymbol{\mu}$. This gives us:

$$\begin{aligned}
\boldsymbol{\Sigma_N} &= (\boldsymbol{\Sigma_0}^{-1} + N\boldsymbol{\Sigma}^{-1})^{-1} \\
\boldsymbol{\mu_N} &= (\boldsymbol{\Sigma_0}^{-1} + N\boldsymbol{\Sigma}^{-1})^{-1}(\boldsymbol{\Sigma_0}^{-1}\boldsymbol{\mu_0} + \boldsymbol{\Sigma}^{-1}N\boldsymbol{\mu}_{MLE})
\end{aligned}$$

Where $\boldsymbol{\mu}_{MLE} = \frac{1}{N}\sum_{n=1}^{N} x_n$. Therefore, we get the Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu_N}, \boldsymbol{\Sigma_N})$, since we have a normalizing factor consisting of a constant and the determinant and there is an exponential consisting of a quadratic and a linear term in $\boldsymbol{\mu}$ and a constant.

4. We get the maximum a posterior solution for $\boldsymbol{\mu}$ by taking the derivative of the Normal distribution and setting it to zero. Since the exponent of the Gaussian is the only part that is dependent on $\boldsymbol{\mu}$, we can directly work with this term to facilitate the solution (using 2.71 from Bishop):

$$\begin{aligned}
\frac{\partial \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)}{\partial \boldsymbol{\mu}} &\propto \frac{\partial ln\, \mathcal{N}(\boldsymbol{\mu}|\boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N)}{\partial \boldsymbol{\mu}} \\
&\propto \frac{\partial - \frac{1}{2}\boldsymbol{\mu}^T \boldsymbol{\Sigma_N}^{-1}\boldsymbol{\mu} + \boldsymbol{\mu}^T \boldsymbol{\Sigma_N}^{-1}\boldsymbol{\mu_N}}{\partial \boldsymbol{\mu}} \\
&= -\boldsymbol{\mu}^T \boldsymbol{\Sigma_N}^{-1} + (\boldsymbol{\Sigma_N}^{-1}\boldsymbol{\mu_N})^T
\end{aligned}$$

Setting this term to zero and solving for $\boldsymbol{\mu}$ gives us the maximum a posteriori solution:

$$\begin{aligned} 0 &= -\boldsymbol{\mu}^T \boldsymbol{\Sigma_N}^{-1} + (\boldsymbol{\Sigma_N}^{-1}\boldsymbol{\mu_N})^T \\ \boldsymbol{\mu}^T &= (\boldsymbol{\Sigma_N}^{-1}\boldsymbol{\mu_N})^T \boldsymbol{\Sigma_N} \\ \boldsymbol{\mu} &= \boldsymbol{\Sigma_N}\boldsymbol{\Sigma_N}^{-1}\boldsymbol{\mu_N} \\ \boldsymbol{\mu}_{MAP} &= \boldsymbol{\mu_N} \end{aligned}$$

Problem 3

1. We denote the number of observations of $x = 1$ (heads) by m and the total number of observations with N. This gives us the MLE estimation:

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

Therefore, MLE assigns a probability of one for the next toss to come up with head.

2. When using a Beta-distribution, the probability that the coin comes up with head in the 4th toss can be calculated with equation (2.20) from Bishop:

$$p(x = 1|D) = \frac{m + a}{m + a + l + b} = \frac{3 + a}{3 + a + b}$$

Where l is the number of coin tosses coming up tails ($l = N - m$).

3. According to Bishop (2.8), (2.15), (2.19) and (2.20):

$$\begin{aligned} \mu_{MLE} &= \frac{m}{m + l} \\ \mu_{prior} &= \frac{a}{a + b} \\ p(x = 1|D) &= \frac{m + a}{m + a + l + b} = \mathbb{E}[\mu|D] \end{aligned}$$

By rearranging the equation, we get:

$$\begin{aligned} \mathbb{E}[\mu|D] &= \frac{m + a}{m + a + l + b} \\ &= \frac{m + l}{m + a + l + b}\frac{m}{m + l} + \frac{a + b}{m + a + l + b}\frac{a}{a + b} \\ &= (1 - \lambda)\frac{m}{m + l} + \lambda\frac{a}{a + b} \\ &= (1 - \lambda)\mu_{MLE} + \lambda\mu_{prior} \end{aligned}$$

Where $0 \leq \lambda \leq 1$. Therefore, the posterior mean is a mixture of μ_{MLE} and μ_{prior}, which implies that it lies in between these two values.

Problem 4

1. We cast members of an exponential family into the form:

$$p(x|\eta) = h(x)\exp[\eta^T\, T(x) - A(\eta)]$$

and denote the sufficient statistics with T.

(i)

$$Pois(k|\lambda) = \frac{1}{k!}\, exp[ln(\lambda) \cdot k - \lambda]$$

$$\begin{aligned}
h(k) &= \frac{1}{k!} \\
\eta &= ln(\lambda) \Rightarrow \lambda = exp(\eta) \\
T(k) &= k \\
A(\eta) &= exp(\eta)
\end{aligned}$$

(ii)

$$\begin{aligned}
Gam(\tau|a,b) &= \frac{1}{\Gamma(a)}\, b^a\, \tau^{a-1}\, exp(-b\,\tau) \\
&= \tau^{-1}\, exp[-b\,\tau + a\,ln(\tau) + a\,ln(b) - ln(\Gamma(a))]
\end{aligned}$$

$$\begin{aligned}
h(\tau) &= \tau^{-1} \\
\eta &= \begin{pmatrix} -b \\ a \end{pmatrix} \Rightarrow b = -\eta_1, a = \eta_2 \\
T(\tau) &= \begin{pmatrix} \tau \\ ln(\tau) \end{pmatrix} \\
A(\eta) &= -a\,ln(b) + ln(\Gamma(a)) \\
&= -\eta_2\,ln(-\eta_1) + ln(\Gamma(\eta_2))
\end{aligned}$$

(iii) The Cauchy distribution is not an exponential family. Since it contains a factor that consists of a sum of both types of involved variables, it cannot be factorized and brought into the form of exp(scalar product).

(iv)

$$\begin{aligned}
vonMises(x|\kappa,\mu) &= \frac{1}{2\pi}\, exp[\kappa\, sin(\mu)\, sin(x) + \kappa\, cos(\mu)\, cos(x) - ln(\boldsymbol{I}_0(\kappa))] \\
h(x) &= \frac{1}{2\pi} \\
\eta &= \begin{pmatrix} \kappa\, sin(\mu) \\ \kappa\, cos(\mu) \end{pmatrix} \Rightarrow \kappa = \sqrt{\eta_1^2 + \eta_2^2}\ \text{(using } sin^2 + cos^2 = 1\text{)} \\
T(x) &= \begin{pmatrix} sin(x) \\ cos(x) \end{pmatrix} \\
A(\eta) &= ln(\boldsymbol{I}_0(\kappa)) \\
&= ln(\boldsymbol{I}_0(\sqrt{\eta_1^2 + \eta_2^2}))
\end{aligned}$$

2. (i)

$$\begin{aligned}
\mathbb{E}[k] &= \frac{\partial A}{\partial \eta} = exp(\eta) = \lambda \\
Var[k] &= \frac{\partial^2 A}{\partial^2 \eta} = exp(\eta) = \lambda
\end{aligned}$$

(ii)

$$\begin{aligned}
\mathbb{E}[k] &= \frac{\partial A}{\partial \eta_1} = \frac{\eta_2}{-\eta_1} = \frac{a}{b} \\
Var[k] &= \frac{\partial^2 A}{\partial^2 \eta_1} = -\frac{\eta_2}{\eta_1^2} = \frac{a}{b^2}
\end{aligned}$$

3. The conjugate prior for an exponential family in the form

$$p(x|\eta) = h(x)\,\text{exp}[\eta^T\,T(x) - A(\eta)]$$

is of the form:

$$p(\eta|\tau,\nu) \propto \text{exp}[\eta^T\tau - \nu A(\eta)]$$

We can find the conjugate prior of the Poisson distribution, by plugging in $A(\eta)$ and replacing variables with the values from the Poisson distribution (i.e. $\tau = a, \nu = b$ and $\lambda = exp(\eta)$):

$$\begin{aligned}
p(\eta|\tau,\nu) &\propto \text{exp}[\eta^T\tau - \nu A(\eta)] \\
&= \text{exp}[\eta^T\tau - \nu\,\text{exp}(\eta)] \\
&= \text{exp}[ln(\lambda)\;a - b\lambda] \\
&= \lambda^{-1}\,\text{exp}[(a+1)ln(\lambda) - b\lambda]
\end{aligned}$$

This is equivalent to the Gamma distribution as shown in task 4.1.(ii) (without the terms independent of λ). Therefore, we can postulate that the Poisson distribution has a conjugate prior and it takes the form of the Gamma distribution $Gam(\lambda|a+1,b)$.