Introduction to probabilistic ML

Exercise 1: Multivariate Gaussian

1. Estimate μ_x and Σ_x using MLE.

Proof. We are looking for ${\mu_x}^{ML}, {\Sigma_x}^{ML} = \text{argmax}_{\mu_x,\Sigma_x}\, p(\mathbf{x}|\mu_x, \Sigma_x)$ which is equivalent to ${\mu_x}^{ML}, {\Sigma_x}^{ML} = \text{argmax}_{\mu_x,\Sigma_x} \log p(\mathbf{x}|\mu_x, \Sigma_x)$ since the logarithm is monotonic.

Let's calculate the form of $\log p(\mathbf{x}|\mu_x, \Sigma_x)$:

$$
\begin{aligned}
\log p(\mathbf{x}|\mu_x, \Sigma_x) &= \log \prod_{n=1}^{N} p(x_n|\mu_x, \Sigma_x) = \sum_{n=1}^{N} \log p(x_n|\mu_x, \Sigma_x) = \\
&= \sum_{n=1}^{N} \left[\log \left(\frac{1}{\sqrt{(2\pi)^K |\Sigma_x|}} \right) - \frac{1}{2}(x_n - \mu_x)^\top {\Sigma_x}^{-1}(x_n - \mu_x) \right] = \\
&= -\frac{N}{2} \log((2\pi)^K |\Sigma_x|) - \frac{1}{2} \sum_{n=1}^{N} \left[(x_n - \mu_x)^\top {\Sigma_x}^{-1}(x_n - \mu_x) \right] = \\
&= C + \frac{N}{2} \log |{\Sigma_x}^{-1}| - \frac{1}{2} \sum_{n=1}^{N} \left[(x_n - \mu_x)^\top {\Sigma_x}^{-1}(x_n - \mu_x) \right] = \qquad (1) \\
&= C + \frac{N}{2} \log |{\Sigma_x}^{-1}| - \frac{1}{2} \sum_{n=1}^{N} \mathrm{Tr} \left[(x_n - \mu_x)^\top {\Sigma_x}^{-1}(x_n - \mu_x) \right] = \\
&= C + \frac{N}{2} \log |{\Sigma_x}^{-1}| - \frac{1}{2} \sum_{n=1}^{N} \mathrm{Tr} \left[{\Sigma_x}^{-1}(x_n - \mu_x)(x_n - \mu_x)^\top \right] = \\
&= C + \frac{N}{2} \log |{\Sigma_x}^{-1}| - \frac{1}{2} \mathrm{Tr} \left[{\Sigma_x}^{-1} \sum_{n=1}^{N} (x_n - \mu_x)(x_n - \mu_x)^\top \right] \qquad (2)
\end{aligned}
$$

where C is a constant and from equation 1 to 2 we have used three facts: (1) a real number (1×1 matrix) is equal to its trace, (2) $\mathrm{Tr}[AB] = \mathrm{Tr}[BA]$, and (3) the trace is a linear function.

Now let's write the derivative of $\log p(\mathbf{x}|\mu_x, \Sigma_x)$ w.r.t. μ_x using the expression 1.

$$\begin{aligned}\frac{\partial}{\partial \mu_x} \log p(\mathbf{x}|\mu_x, \Sigma_x) &= -\frac{1}{2}\sum_{n=1}^{N} \frac{\partial}{\partial \mu_x}\left[(x_n-\mu_x)^\top {\Sigma_x}^{-1}(x_n-\mu_x)\right] = \\ &= -\frac{1}{2}\sum_{n=1}^{N}\left[-{\Sigma_x}^{-1}(x_n-\mu_x) - (x_n-\mu_x)^\top {\Sigma_x}^{-1}\right] = \\ &= -\frac{1}{2}\sum_{n=1}^{N}\left[-2{\Sigma_x}^{-1}(x_n-\mu_x)\right] = {\Sigma_x}^{-1}\sum_{n=1}^{N}(x_n-\mu_x)\end{aligned}$$

where we have used the fact that ${\Sigma_x}^{-1}$ is a full-rank symmetric positive-definite matrix. And we do the same w.r.t. ${\Sigma_x}^{-1}$ using the expression 2.

$$\frac{\partial}{\partial {\Sigma_x}^{-1}} \log p(\mathbf{x}|\mu_x, \Sigma_x) = \underbrace{\frac{N}{2}\frac{\partial \log |{\Sigma_x}^{-1}|}{\partial {\Sigma_x}^{-1}}}_{(3)} - \underbrace{\frac{1}{2}\frac{\partial\, \mathrm{Tr}}{\partial {\Sigma_x}^{-1}}\left[{\Sigma_x}^{-1}\sum_{n=1}^{N}(x_n-\mu_x)(x_n-\mu_x)^\top\right]}_{(4)} = (*)$$

$$(3) = \frac{N}{2}({\Sigma_x}^{-1})^{-\top} = \frac{N}{2}{\Sigma_x}^\top \quad \text{since } \frac{\partial \log |A|}{\partial A} = A^{-\top}$$

$$\begin{aligned}(4) &= \frac{1}{2}\left[\sum_{n=1}^{N}(x_n-\mu_x)(x_n-\mu_x)^\top\right]^\top \quad \text{since } \frac{\partial\, \mathrm{Tr}(AB)}{\partial A} = B^\top \\ &= \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu_x)(x_n-\mu_x)^\top \quad \text{since } (A+B)^\top = A^\top + B^\top \text{ and } (AB)^\top = B^\top A^\top\end{aligned}$$

Thus

$$(*) = (3) - (4) = \frac{N}{2}{\Sigma_x}^\top - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu_x)(x_n-\mu_x)^\top$$

Therefore we have the equation system:

$$\begin{cases}\partial_{\mu_x} \log p(\mathbf{x}) = 0 & \iff {\Sigma_x}^{-1}\sum_{n=1}^{N}(x_n-\mu_x) = 0 \\ \partial_{{\Sigma_x}^{-1}} \log p(\mathbf{x}) = 0 & \iff \frac{N}{2}{\Sigma_x}^\top - \frac{1}{2}\sum_{n=1}^{N}(x_n-\mu_x)(x_n-\mu_x)^\top = 0\end{cases}$$

The first equation can be readily solved since

$${\Sigma_x}^{-1}\sum_{n=1}^{N}(x_n-\mu_x) = 0 \iff \sum_{n=1}^{N} x_n - N\mu_x = 0 \iff \mu_x = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad (3)$$

and we can check that it is in fact a maximum

$$\frac{\partial^2}{\partial \mu_x} \log p(\mathbf{x}|\mu_x, \Sigma_x) = -N{\Sigma_x}^{-1} \prec 0 \quad \text{since } \Sigma_x \succ 0 \text{ (p.s.d.)} \tag{4}$$

so we can call ${\mu_x}^{ML} := \frac{1}{N}\sum_{n=1}^{N} x_n$ and substituting ${\mu_x}^{ML}$ in the second equation we have that

$$\frac{N}{2}{\Sigma_x}^\top - \frac{1}{2}\sum_{n=1}^{N}(x_n - {\mu_x}^{ML})(x_n - {\mu_x}^{ML})^\top = 0 \iff$$
$$\frac{N}{2}{\Sigma_x}^\top = \frac{1}{2}\sum_{n=1}^{N}(x_n - {\mu_x}^{ML})(x_n - {\mu_x}^{ML})^\top \iff$$
$${\Sigma_x}^\top = \Sigma_x = \frac{1}{N}\sum_{n=1}^{N}(x_n - {\mu_x}^{ML})(x_n - {\mu_x}^{ML})^\top$$

and again we can check that this is a maximum:

$$\frac{\partial^2}{\partial {\Sigma_x}^{-1}} \log(\mathbf{x}|\mu_x, \Sigma_x) = \frac{N}{2}\frac{\partial}{\partial {\Sigma_x}^{-1}}\left({\Sigma_x}^{-1}\right)^{-1} = -\frac{N}{2}{\Sigma_x}^2 \prec 0 \tag{5}$$

Finally, ${\mu_x}^{ML} := \frac{1}{N}\sum_{n=1}^{N} x_n$ and ${\Sigma_x}^{ML} := \frac{1}{N}\sum_{n=1}^{N}(x_n - {\mu_x}^{ML})(x_n - {\mu_x}^{ML})^\top$. □

2. Estimate μ_x using MAP assuming known μ_x and prior $\mu_x \sim \mathcal{N}(\mu_0, \Sigma_0)$.

   Proof. Using Bayes' theorem:

$$p(\mu_x|\mathbf{x}, \mu_0, \Sigma_0, \Sigma_x) = \frac{p(\mathbf{x}|\mu_x, \Sigma_x)p(\mu_x|\mu_0, \Sigma_0)}{p(\mathbf{x})} \propto p(\mathbf{x}|\mu_x, \Sigma_x)p(\mu_x|\mu_0, \Sigma_0) \tag{1}$$

   We are going to discover the form of the posterior distribution by trying to obtain a formula that we can recognize, if we achieve this calculating the normalizing constant is straight-forward. In particular, we are going to compute $\log p(\mu_x|\mathbf{x}, \mu_0, \Sigma_0, \Sigma_x)$ and try to obtain a quadratic form of μ_x which is the form that gaussian distributions have.

$$\log p(\mu_x|\mathbf{x}, \mu_0, \Sigma_0, \Sigma_x) = \log \mathcal{N}(\mathbf{x}|\mu_x, \Sigma_x) + \log \mathcal{N}(\mu_x|\mu_0, \Sigma_0) + C =$$
$$= -\frac{1}{2}\sum_{n=1}^{N}(x_n - \mu_x)^\top {\Sigma_x}^{-1}(x_n - \mu_x) - \frac{1}{2}(\mu_x - \mu_0)^\top {\Sigma_0}^{-1}(\mu_x - \mu_0) + C =$$
$$= \frac{1}{2}\left[\sum_{n=1}^{N}\left({\mu_x}^\top {\Sigma_x}^{-1}\mu_x - 2{\mu_x}^\top {\Sigma_x}^{-1} x_n\right) + {\mu_x}^\top {\Sigma_0}^{-1}\mu_x - 2{\mu_x}^\top {\Sigma_0}^{-1}\mu_0\right] + C =$$
$$= -\frac{1}{2}\left[{\mu_x}^\top\left(N{\Sigma_x}^{-1} + {\Sigma_0}^{-1}\right)\mu_x - 2{\mu_x}^\top\left({\Sigma_x}^{-1}\sum_{n=1}^{N} x_n + {\Sigma_0}^{-1}\mu_0\right)\right] + C \tag{2}$$

Now, we have to complete squares in equation 2. To do that we know that, if A is symmetric, $(x-y)^\top A(x-y) = x^\top Ax + y^\top Ay - 2x^\top Ay$. Comparing equation 2 with the previous formula we can call $x = \mu_x$ and $A = (N{\Sigma_x}^{-1} + {\Sigma_0}^{-1})$.

In order to find out who is y we have to make A appear in the expression $-2x^\top Ay$ of equation 2. We can easily achieve this by multiplying by AA^{-1}, making equation 2 to look like

$$(2) = -\frac{1}{2}\left[{\mu_x}^\top A\mu_x - 2{\mu_x}^\top A\left[A^{-1}\left({\Sigma_x}^{-1}\sum_{n=1}^{N} x_n + {\Sigma_0}^{-1}\mu_0\right)\right]\right] + C$$

and by calling $y = A^{-1}\left({\Sigma_x}^{-1}\sum_{n=1}^{N} x_n + {\Sigma_0}^{-1}\mu_0\right)$ we have that

$$(2) = -\frac{1}{2}(\mu_x - y)^\top A(\mu_x - y) + C$$

Now, if μ_x had a normal posterior distribution, i.e., $\mu_x|\mathbf{x} \sim \mathcal{N}(\mu_1, \Sigma_1)$, then $\log p(\mu_x|x)$ would be of the form

$$\log p(\mu_x|\mathbf{x}) = -\frac{1}{2}(\mu_x - \mu_1)^\top {\Sigma_1}^{-1}(\mu_x - \mu_1) + C$$

which implies, by comparing the two expressions, that the posterior distribution of μ_x is a Gaussian distribution with mean $\mu_1 = y$ and covariance $\Sigma_1 = A^{-1}$.

Finally, we need to compute the MAP estimate of μ_x given $\mathbf{x}$. This estimator is defined as ${\mu_x}^{MAP} := \mathrm{argmax}_{\mu_x}\, p(\mu_x|\mathbf{x})$ which, making similar calculations as the ones done in the previous section, can be proved to be the mean of the normal distribution, that is, ${\mu_x}^{MAP} = \mu_1 = y$. □

Exercise 2: Categorical distribution

1. Estimate π using ML.

Proof. We have to solve the problem (note that we use the shorthand $\pi = \{\pi_k\}_{k=1}^{K}$)

$$\pi^{ML} := \operatorname*{argmax}_{\pi} p(\mathbf{x}|\pi) \qquad \text{subject to } \sum_{k=1}^{K} \pi_k = 1 \tag{1}$$

which is equivalent to solving

$$\pi^{ML} := \operatorname*{argmax}_{\pi} \log p(\mathbf{x}|\pi) \qquad \text{subject to } \sum_{k=1}^{K} \pi_k = 1 \tag{2}$$

and using Lagrange multipliers this is equivalent to solving

$$\pi^{ML} := \underset{\pi}{\operatorname{argmax}} \left[\log p(\mathbf{x}|\pi) - \lambda \left(\sum_{k=1}^{K} \pi_k - 1 \right) \right] \tag{3}$$

where λ is a sufficiently large real positive number.

Let's write down the form of the log-likelihood:

$$p(\mathbf{x}|\pi) = \prod_{n=1}^{N} p(x_n|\pi) = \prod_{n=1}^{N}\prod_{k=1}^{K} \pi_k^{[x_n=k]} \quad \text{where } [x=k] = \begin{cases} 1 & \text{if } x = k \\ 0 & \text{otherwise} \end{cases}$$

$$\log p(\mathbf{x}|\pi) = \sum_{n=1}^{N}\sum_{k=1}^{K} \log\left(\pi_k^{[x_n=k]}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} [x_n = k] \log \pi_k \tag{4}$$

Now we have to solve the system

$$\begin{cases} \partial_{\pi_1} \log p(\mathbf{x}|\pi) = 0 \\ \partial_{\pi_2} \log p(\mathbf{x}|\pi) = 0 \\ \dots \\ \partial_{\pi_K} \log p(\mathbf{x}|\pi) = 0 \end{cases} \tag{5}$$

Therefore, let us solve this equation for every $l \in \{1, 2, ..., K\}$

$$\begin{aligned} \frac{\partial \log p(\mathbf{x}|\pi)}{\partial \pi_l} &= \sum_{n=1}^{N}\sum_{k=1}^{K} \frac{\partial\left([x_n = k]\log \pi_k\right)}{\partial \pi_l} - \lambda \frac{\partial\left(\sum_{k=1}^{K} \pi_k - 1\right)}{\partial \pi_l} = \\ &= \sum_{n=1}^{N} \frac{[x_n = l]}{\pi_l} - \lambda = 0 \iff \pi_l = \frac{1}{\lambda}\sum_{n=1}^{N}[x_n = l] = \frac{1}{\lambda} n_l \end{aligned}$$

where n_l represents how many x_n in $\mathbf{x}$ have the value l. Note that this is indeed a maximum since

$$\frac{\partial^2}{\partial \pi_l} \log p(\mathbf{x}|\pi) = -\frac{n_l}{\pi_l^2} < 0$$

assuming that every class has a non-zero probability of happening (that is, it has been observed at least once).

We have a set of solutions $\pi_k^{ML}(\lambda) = n_k/\lambda$, one per each value of λ. In order to solve the problem we solve λ substituting $\pi^{ML}(\lambda)$ on the restriction over π:

$$\sum_{k=1}^{K} \pi_k^{ML}(\lambda) = \frac{1}{\lambda}\sum_{k=1}^{K} n_k = 1 \iff \lambda = \sum_{k=1} n_k = N \tag{6}$$

Therefore, the maximum likelihood estimator of π_k is

$$\pi_k^{ML} = \frac{1}{N}\sum_{n=1}^{N}[x_n = k] = \frac{n_k}{N} \tag{7}$$

$\square$

2. Calculate the posterior $p(\pi|\mathbf{x})$ assuming a prior $\pi \sim Dirichlet(\alpha)$.

Proof. We assume a prior

$$p(\pi|\alpha) = Dirichlet(\alpha) = \frac{1}{B(\alpha)}\prod_{k=1}^{K}\pi_k^{\alpha_k - 1} \tag{1}$$

Using Bayes' theorem we have that

$$p(\pi|\mathbf{x},\alpha) \propto p(\mathbf{x}|\pi)p(\pi|\alpha) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{[x_n=k]}\prod_{k=1}^{K}\pi_k^{\alpha_k-1} =$$
$$= \prod_{k=1}^{K}\pi_k^{\sum_{n=1}^{N}[x_n=k]+\alpha_k-1} = \prod_{k=1}^{K}\pi_k^{n_k+\alpha_k-1}$$

And, since it has the same form as a Dirichet distribution up to the normalization constant, we know that $\pi|\mathbf{x} \sim Dirichlet(n_1 + \alpha_1, n_2 + \alpha_2, \ldots, n_K + \alpha_K)$. $\square$

Exercise 3: Graphical models

1. Write the graphical model corresponding to the generative model

$$p(\{x_n, z_n\}_{n=1}^N, \{\pi_k, \mu_k\}_{k=1}^K) = \prod_{n=1}^N p(x_n|z_n, \{\mu_k\}_{k=1}^K, \sigma_x) p(z_n|\{\pi_k\}_{k=1}^K) p(\{\pi_k\}_{k=1}^K|\alpha) \quad (1)$$

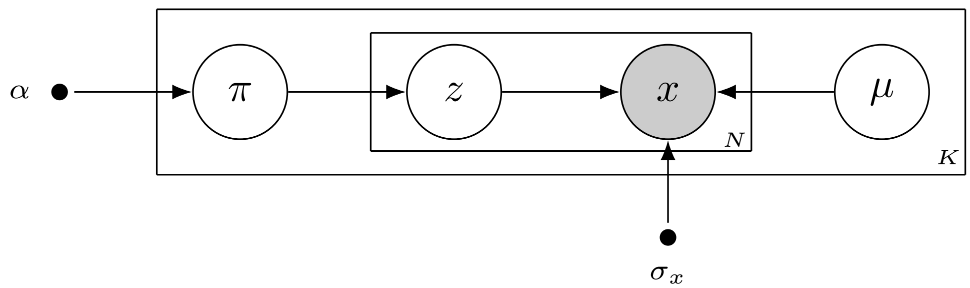

Figure 1: Graphical model of exercise 3.1

2. Write down the generative model of the graphical model.

$$p(\{\omega_n, z_n\}_{n=1}^N, \{\theta_m\}_{m=1}^M, \alpha, \beta) = \prod_{n=1}^N p(\omega_n|z_n, \beta) p(z_n|\{\theta_m\}_{m=1}^M) p(\{\theta_m\}_{m=1}^M|\alpha) p(\beta)$$