Contents

1 Tutorial 1: Introduction to probabilistic ML

1.1 Exercise 1: Multivariate Gaussian

Given a data set $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}^\top$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a multivariate Gaussian distribution, i.e. $\mathbf{x}_1, \dots, \mathbf{x}_N \sim \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$:

1. Estimate the mean and covariance parameters $\boldsymbol{\mu}_x$ and $\boldsymbol{\Sigma}_x$, by maximum likelihood.

We are looking for the estimators ${\mu_x}^{ML}, {\Sigma_x}^{ML} = \arg\max_{\mu_x, \Sigma_x} p(\mathbf{x}|\mu_x, \Sigma_x)$ which are equivalent to ${\mu_x}^{ML}, {\Sigma_x}^{ML} = \arg\max_{\mu_x, \Sigma_x} \log p(\mathbf{x}|\mu_x, \Sigma_x)$ since the logarithm is an increasing function.

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

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

where C is a constant and from equation 1 to 2 we have used three facts: i) a real number (1×1 matrix) is equal to its trace, ii) $\operatorname{Tr}[AB] = \operatorname{Tr}[BA]$, and iii) 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 expression 1.

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

In order to pass from the first to the second equality we can compute each component of the derivative as follows:

$$\begin{aligned}
\frac{\partial}{\partial \mu_{x,k}} \log p(\mathbf{x}|\mu_x, \mathit{\Sigma}_x) &= -\frac{1}{2}\sum_{n=1}^{N} \frac{\partial}{\partial \mu_{x,k}} \sum_{i,j} \Big[(x_{n,i}-\mu_{x,i})^\top ({\mathit{\Sigma}_x}^{-1})_{ij}(x_{n,j}-\mu_{x,j})\Big] \\
&= \frac{1}{2}\sum_{n=1}^{N}\sum_{i,j}\Big[\delta_{ik}({\mathit{\Sigma}_x}^{-1})_{ij}(x_{n,j}-\mu_{x,j}) + \delta_{jk}({\mathit{\Sigma}_x}^{-1})_{ij}(x_{n,i}-\mu_{x,i})\Big] \\
&= \frac{1}{2}\sum_{n=1}^{N}\left[\sum_j ({\mathit{\Sigma}_x}^{-1})_{kj}(x_{n,j}-\mu_{x,j}) + \sum_i ({\mathit{\Sigma}_x}^{-1})_{ik}(x_{n,i}-\mu_{x,i})\right] \\
&= \frac{1}{2}\sum_{n=1}^{N}\left[\sum_j ({\mathit{\Sigma}_x}^{-1})_{kj}(x_{n,j}-\mu_{x,j}) + \sum_i ({\mathit{\Sigma}_x}^{-1})_{ki}(x_{n,i}-\mu_{x,i})\right]
\end{aligned}$$

where we invoked that ${\mathit{\Sigma}_x}^{-1}$ is a full-rank symmetric positive-definite matrix.

And we do the same w.r.t. ${\mathit{\Sigma}_x}^{-1}$ using expression 2.

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

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

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

and thus

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

Therefore we have the equation system:

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

The first equation can be readily solved since

$${\mathit{\Sigma}_x}^{-1}\sum_{n=1}^{N}(x_n-\mu_x) = 0 \iff \sum_{n=1}^{N}x_n - N\mu_x = 0 \iff \mu_x = \frac{1}{N}\sum_{n=1}^{N}x_n \tag{3}$$

and we can check that it is in fact a maximum

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

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

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

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. Assume the covariance matrix $\boldsymbol{\Sigma}_x$ to be known and a Gaussian prior over the mean parameter $\boldsymbol{\mu}_x$ with mean $\boldsymbol{\mu}_0$ and identity covariance matrix, i.e. $\mathcal{N}(\boldsymbol{\mu}_x|\boldsymbol{\mu}_0, \mathbf{I})$. Compute the distribution a posteriori of the mean parameter μ_x given the observed data $\mathbf{X}$, i.e. $p(\boldsymbol{\mu}_x|\mathbf{X}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_x)$, and its *Maximum a posteriori* (MAP) solution.

Using Bayes' theorem:

$$p(\mu_x|\mathbf{x}, \mu_0, \Sigma_0, \Sigma_x) = \frac{p(\mathbf{x}|\mu_x, \Sigma_x)p(\mu_x|\mu_0, \Sigma_0)}{p(\mathbf{x}|\mu_0, \Sigma_0, \Sigma_x)} \propto p(\mathbf{x}|\mu_x, \Sigma_x)p(\mu_x|\mu_0, \Sigma_0) \tag{6}$$

We are going to discover the form of the posterior distribution by trying to obtain a formula that we can recognize. If we do so, 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.

$$\begin{aligned}
&\log p(\mu_x|\mathbf{x}, \mu_0, \Sigma_0, \Sigma_x) = \log \mathcal{N}(\mathbf{x}|\mu_x, \Sigma_x) + \log \mathcal{N}(\mu_x|\mu_0, \Sigma_0) + C = \\
&= -\frac{1}{2}\sum_{n=1}^{N}(x_n - \mu_x)^{\top}{\Sigma_x}^{-1}(x_n - \mu_x) - \frac{1}{2}(\mu_x - \mu_0)^{\top}{\Sigma_0}^{-1}(\mu_x - \mu_0) + C = \\
&= -\frac{1}{2}\left[\sum_{n=1}^{N}\left({\mu_x}^{\top}{\Sigma_x}^{-1}\mu_x - 2{\mu_x}^{\top}{\Sigma_x}^{-1}x_n\right) + {\mu_x}^{\top}{\Sigma_0}^{-1}\mu_x - 2{\mu_x}^{\top}{\Sigma_0}^{-1}\mu_0\right] + C = \\
&= -\frac{1}{2}\left[{\mu_x}^{\top}\left(N{\Sigma_x}^{-1} + {\Sigma_0}^{-1}\right)\mu_x - 2{\mu_x}^{\top}\left({\Sigma_x}^{-1}\sum_{n=1}^{N}x_n + {\Sigma_0}^{-1}\mu_0\right)\right] + C
\end{aligned} \tag{7}$$

Now, we have to complete squares in equation 7. 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 7 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 7. We can easily achieve this multiplying by AA^{-1}, making equation 7 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 $y = A^{-1}\left(\mathit{\Sigma}_x{}^{-1}\sum_{n=1}^{N} x_n + \mathit{\Sigma}_0{}^{-1}\mu_0\right)$ 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., $\mu_x|\mathbf{x} \sim \mathcal{N}(\mu_1, \mathit{\Sigma}_1)$, then $\log p(\mu_x|x)$ would be of the form

$$\log p(\mu_x|\mathbf{x}) = -\frac{1}{2}(\mu_x - \mu_1)^{\top} \mathit{\Sigma}_1{}^{-1}(\mu_x - \mu_1) + C$$

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 $\mathit{\Sigma}_1 = A^{-1}$.

Finally, we need to compute the MAP estimate of μ_x given $\mathbf{x}$. This estimator is defined as $\mu_x{}^{MAP} := \arg\max_{\mu_x} p(\mu_x|\mathbf{x})$ 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, $\mu_x{}^{MAP} = \mu_1 = y$.

1.2 Exercise 2: Categorical distribution

Given a data set $\mathbf{X} = \{x_1, \dots, x_N\}^\top$ in which the observations $x_n \in \{1, \dots, k\}$ are assumed to be drawn independently from a Categorical distribution, i.e., $x_1, \dots, x_N \sim Categorical(x|\pi_1, \dots, \pi_k)$:

1. Estimate the parameters, i.e., the category probabilities $\{\pi_k\}$ by maximum likelihood.

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

$$\pi^{ML} := \arg\max_\pi p(\mathbf{x}|\pi) \qquad \text{subject to } \sum_{k=1}^K \pi_k = 1 \tag{8}$$

which is equivalent to solving

$$\pi^{ML} := \arg\max_\pi \log p(\mathbf{x}|\pi) \qquad \text{subject to } \sum_{k=1}^K \pi_k = 1 \tag{9}$$

and using Lagrange multipliers this is equivalent to solving

$$\pi^{ML} := \arg\max_\pi \left[\log p(\mathbf{x}|\pi) - \lambda \left(\sum_{k=1}^K \pi_k - 1\right)\right] \tag{10}$$

where λ is a sufficiently large real positive number.

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

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

$$\log p(\mathbf{x}|\pi) = \sum_{n=1}^N \sum_{k=1}^K \log\left(\pi_k^{[x_n=k]}\right) = \sum_{n=1}^N \sum_{k=1}^K [x_n = k] \log \pi_k \tag{11}$$

Now we have to solve the system

$$\begin{cases} \partial_{\pi_1}\left[\log p(\mathbf{x}|\pi) - \lambda\left(\sum_{k=1}^K \pi_k - 1\right)\right] = 0 \\ \partial_{\pi_2}\left[\log p(\mathbf{x}|\pi) - \lambda\left(\sum_{k=1}^K \pi_k - 1\right)\right] = 0 \\ \dots \\ \partial_{\pi_K}\left[\log p(\mathbf{x}|\pi) - \lambda\left(\sum_{k=1}^K \pi_k - 1\right)\right] = 0 \end{cases} \tag{12}$$

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:

$$\sum_{k=1}^{K} \pi_k^{ML}(\lambda) = \frac{1}{\lambda}\sum_{k=1}^{K} n_k = 1 \iff \lambda = \sum_{k=1} n_k = N \tag{13}$$

Therefore, the maximum likelihood estimator of π_k is

$$\pi_k^{ML} = \frac{1}{N}\sum_{n=1}^{N} [x_n = k] = \frac{n_k}{N} \tag{14}$$

2. Assume a Dirichlet prior over the category probabilities $\{\pi_k\}$ with hyperparameter $\alpha = (\alpha_1, ..., \alpha_K)$, i.e., $\pi_1, \ldots, \pi_k \sim Dirichlet(\pi_1, \ldots, \pi_k|\alpha)$. Compute the distribution a posteriori of the category probabilities $\{\pi_k\}$ given the observed data $\mathbf{X}$, i.e., $p(\pi_1, \ldots, \pi_k|\mathbf{X}, \alpha)$.

We assume a prior

$$p(\pi|\alpha) = Dirichlet(\alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \tag{15}$$

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 $\pi|\mathbf{x} \sim Dirichlet(n_1 + \alpha_1, n_2 + \alpha_2, \ldots, n_K + \alpha_K)$.

2 Q&A

Question 1

On page 3 [of the lecture notes] I am confused about the meaning of the following statements, "For the example considered above of a Gaussian Posterior, we have $\mu_{MAP} \equiv \mu_N$" - This is obvious. "But this is not the case in general." - I am not sure what μ_N means in general. Just the mean of the posterior (of arbitrary shape)?

Answer 1

A: The comment is indeed suggesting that, in general, the maximum of the posterior is not attained at its mean. Yes, μ_N in this context can be thought as the mean of the posterior distribution.

Question 2

"Obs1: if the Prior is uniform, then $\mu_{MAP} \equiv \mu_N$" - I am not sure what a uniform prior means in this context. Clearly a uniform distribution over a fixed interval will clip the likelihood. A uniform prior over the whole space doesn't seem to make sense. Maybe we can talk about the limit in case of uniform distributions centered at the mean of the likelihood.

Answer 2

There is a mistake in Obs. 1, of pag.3 in the notes of lecture 1. There, we wanted to point out that for a uniform prior we get $\mu_{MAP} \equiv \mu_{MLE}$, since in such case the posterior is identical to the likelihood multiplied by a proportionality constant (the prior). The extreme case of a non-informative Gaussian (as I believe also pointed out by a student during the tutorial) can be thought as: $\lim_{\sigma_0^2 \to +\infty} \mathcal{N}(\mu_0, \sigma_0^2)$, practically $\mathcal{N}(\mu_0, \sigma_0^2)$ with $\sigma_0^2 \gg 1$. Thus, it is immediate to observe from Eq. (11) that $\mu_N \sim \mu_{MLE}$, and because the Gaussian has maximum at its mean, that $\mu_{MAP} \sim \mu_{MLE}$. However, note that having $\mu_{MAP} = \mu_{MLE}$ for a uniform prior is a general fact, since such prior enters as a multiplication factor in Eq. (10), implying the likelihood is be identical (minus a multiplicative constant given by the prior and the marginal) to the posterior.

Question 3

On page 4, "Given that this integral is not always easy to calculate, one can instead derive an expression for the conditional distribution rewriting as: p(x1|x2) = p(x1,x2)/Z1 .... Usually finding Z1 is easier than calculating the integral in (16)." I am not sure what this means. My guess (as discussed) would be that we would usually assume the joint factorises into a likelihood and prior of known distributions, and see if we can combine the factors into another known distribution - from which the normalisation constant would be evident (similar to the exercises today). Am I missing something here?

Answer 3

You are not missing anything, the sentence is indeed suggesting to what you mentioned. Because likelihood and priors are posited, you can select a combination of the two which leads to a posterior with a nice form. This is what happens for instance with conjugate priors, priors that combined with the likelihood lead to a posterior of the same shape, e.g. Poisson-Gamma.

Question 4

3. On page 5, section 4 - the expression of the joint p(x1,x2) (x1, x2 as d-dimensional) given in eqn 21 seems to be derived from De Finetti's theorem. In particular, the distinction that seems important to me is that we are not simply assuming the existence of a latent variable z (k-dimensional) such that x1, x2 are conditionally independent given z, but the theorem guarantees the existence of such a latent variable z.
Regarding this it would be helpful if you could point me to a reference for the theorem. In particular I would be interested in knowing under what conditions this holds true, what is the dimensionality of z, and maybe some interpretation of z (maybe with respect to the data generating process?). Of course I don't expect you to explain all of this to me, I should be able to figure it out from the reference! Just writing it down so you know what I'm looking for.

Answer 4

For the the main necessary ingredients to prove the theorem, and for a self-contained formal proof, you can look up here[a]. The famous paper here[b] has a brief application/mention of this see Sec 3.1. Another nice, but more formal reference is this[c].

[a] https://arxiv.org/pdf/1809.00882.pdf
[b] https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf
[c] https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6847223

3 Errata

- The goal of ex. 2 point 2 was indeed to compute *both* I mentioned the a posteriori distribution of the mean parameter μ_x, and its MAP estimator (during the tutorial I mentioned only the latter). This is exactly what has been done during the tutorial. Keep in mind that the a posteriori distribution has to be normalized, and in this exercise the normalization constant "comes for free" after observing that the a posteriori distribution is Gaussian.
- There is a mistake in Obs. 1, pag. 3 in the note of lecture 1. There we wanted to point out that for a uniform prior $\mu_{MAP} \equiv \mu_{MLE}$, since in such case the posterior is identical to the likelihood multiplied by a proportionality constant (the prior).