Advanced Probabilistic Machine Learning and Applications

Martina Contisciani and Caterina De Bacco

October 25, 2021

Contents

1 Tutorial 1: Introduction to probabilistic ML

1.1 Exercise 1: Multivariate Gaussian

Given a dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}^\top$ in which the observations $\{\mathbf{x}_n\}$ are assumed to be drawn independently from a $K-$dimensional multivariate Gaussian distribution, i.e. $\mathbf{x}_n \sim \mathcal{N}_K(\mathbf{x}_n | \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)\ \forall n = 1, \dots, N$:

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

We are looking for the estimators ${\boldsymbol{\mu}_x}^{ML}, {\boldsymbol{\Sigma}_x}^{ML} = \arg\max_{\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x} p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ which are equivalent to ${\boldsymbol{\mu}_x}^{ML}, {\boldsymbol{\Sigma}_x}^{ML} = \arg\max_{\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x} \log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ since the logarithm is a monotonic increasing function.

Let's calculate $\log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$:

$$\begin{aligned}
\log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) &= \log \prod_{n=1}^{N} p(\mathbf{x}_n|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) \\
&= \sum_{n=1}^{N} \left[\log \left(\frac{1}{\sqrt{(2\pi)^K |\boldsymbol{\Sigma}_x|}} \right) - \frac{1}{2} (\mathbf{x}_n - \boldsymbol{\mu}_x)^\top {\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_x) \right] \\
&= -\frac{N}{2} \log((2\pi)^K |\boldsymbol{\Sigma}_x|) - \frac{1}{2} \sum_{n=1}^{N} \left[(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top {\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_x) \right] \\
&= C + \frac{N}{2} \log |{\boldsymbol{\Sigma}_x}^{-1}| - \frac{1}{2} \sum_{n=1}^{N} \left[(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top {\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_x) \right] \qquad (1) \\
&= C + \frac{N}{2} \log |{\boldsymbol{\Sigma}_x}^{-1}| - \frac{1}{2} \sum_{n=1}^{N} \operatorname{Tr} \left[(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top {\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_x) \right] \\
&= C + \frac{N}{2} \log |{\boldsymbol{\Sigma}_x}^{-1}| - \frac{1}{2} \sum_{n=1}^{N} \operatorname{Tr} \left[{\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top \right] \\
&= C + \frac{N}{2} \log |{\boldsymbol{\Sigma}_x}^{-1}| - \frac{1}{2} \operatorname{Tr} \left[{\boldsymbol{\Sigma}_x}^{-1} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top \right], \qquad (2)
\end{aligned}$$

where C is a constant. From equation (1) to equation (2) we have used three facts: i) a real number (1×1 matrix) is equal to its trace, ii) $\operatorname{Tr}[ABC] = \operatorname{Tr}[CAB] = \operatorname{Tr}[BCA]$, and iii) the trace is a linear function.

Now let's write the derivative of $\log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ w.r.t. $\boldsymbol{\mu}_x$ using equation (1):

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\mu}_x} \log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) &= -\frac{1}{2} \sum_{n=1}^{N} \frac{\partial}{\partial \boldsymbol{\mu}_x} \left[(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top {\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_x) \right] \\
&= -\frac{1}{2} \sum_{n=1}^{N} \left[-2{\boldsymbol{\Sigma}_x}^{-1} (\mathbf{x}_n - \boldsymbol{\mu}_x) \right] = {\boldsymbol{\Sigma}_x}^{-1} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x). \qquad (3)
\end{aligned}$$

Now let's write the derivative of $\log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ w.r.t. ${\boldsymbol{\Sigma}_x}^{-1}$ using equation (2):

$$\frac{\partial}{\partial {\boldsymbol{\Sigma}_x}^{-1}} \log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = \underbrace{\frac{N}{2} \frac{\partial \log |{\boldsymbol{\Sigma}_x}^{-1}|}{\partial {\boldsymbol{\Sigma}_x}^{-1}}}_{(a)} - \underbrace{\frac{1}{2} \frac{\partial \operatorname{Tr}}{\partial {\boldsymbol{\Sigma}_x}^{-1}} \left[{\boldsymbol{\Sigma}_x}^{-1} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top \right]}_{(b)} = (*)$$

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

$$\begin{aligned}
(b) &= \frac{1}{2} \left[\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top \right]^\top \quad \text{since } \frac{\partial \operatorname{Tr}(AB)}{\partial A} = B^\top \\
&= \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top \quad \text{since } \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top \text{ is symmetric}
\end{aligned}$$

and thus

$$(*) = (a) - (b) = \frac{N}{2} {\boldsymbol{\Sigma}_x}^\top - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top. \qquad (4)$$

Therefore we have the equation system:

$$\begin{cases} \partial_{\boldsymbol{\mu}_x} \log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = 0 & \iff \boldsymbol{\Sigma}_x{}^{-1} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x) = 0 \\ \partial_{\boldsymbol{\Sigma}_x{}^{-1}} \log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = 0 & \iff \frac{N}{2} \boldsymbol{\Sigma}_x{}^{\top} - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x)(\mathbf{x}_n - \boldsymbol{\mu}_x)^{\top} = 0 \end{cases} \tag{5}$$

The first equation can be readily solved since

$$\boldsymbol{\Sigma}_x{}^{-1} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x) = 0 \iff \sum_{n=1}^{N} \mathbf{x}_n - N\boldsymbol{\mu}_x = 0 \iff \boldsymbol{\mu}_x = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n \tag{6}$$

and we can check that it is in fact a maximum

$$\frac{\partial^2}{\partial \boldsymbol{\mu}_x} \log p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = -N \boldsymbol{\Sigma}_x{}^{-1} \prec 0 \quad \text{since } \boldsymbol{\Sigma}_x{}^{-1} \succ 0 \tag{7}$$

so we have that $\boldsymbol{\mu}_x{}^{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$, and substituting $\boldsymbol{\mu}_x{}^{ML}$ in the second equation we have

$$\begin{aligned} &\frac{N}{2} \boldsymbol{\Sigma}_x{}^{\top} - \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x{}^{ML})(\mathbf{x}_n - \boldsymbol{\mu}_x{}^{ML})^{\top} = 0 \iff \\ &\frac{N}{2} \boldsymbol{\Sigma}_x{}^{\top} = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x{}^{ML})(\mathbf{x}_n - \boldsymbol{\mu}_x{}^{ML})^{\top} \iff \\ &\boldsymbol{\Sigma}_x{}^{\top} = \boldsymbol{\Sigma}_x = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x{}^{ML})(\mathbf{x}_n - \boldsymbol{\mu}_x{}^{ML})^{\top} \end{aligned} \tag{8}$$

and again we can check that this is a maximum:

$$\frac{\partial^2}{\partial \boldsymbol{\Sigma}_x{}^{-1}} \log(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) = \frac{N}{2} \frac{\partial}{\partial \boldsymbol{\Sigma}_x{}^{-1}} \left(\boldsymbol{\Sigma}_x{}^{-1}\right)^{-1} = -\frac{N}{2} \boldsymbol{\Sigma}_x{}^{2} \prec 0. \tag{9}$$

Finally, $\boldsymbol{\mu}_x{}^{ML} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$ and $\boldsymbol{\Sigma}_x{}^{ML} = \frac{1}{N} \sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_x{}^{ML})(\mathbf{x}_n - \boldsymbol{\mu}_x{}^{ML})^{\top}$.

2. Assume the covariance matrix $\boldsymbol{\Sigma}_x$ to be known and the existence of a multivariate Gaussian prior over the mean parameter $\boldsymbol{\mu}_x$ with mean $\boldsymbol{\mu}_0$ and identity covariance matrix, i.e. $\mathcal{N}_K(\boldsymbol{\mu}_x|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\Sigma}_0 = \mathbf{I}$. Compute the distribution a posteriori of the mean parameter $\boldsymbol{\mu}_x$ 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:

$$p(\boldsymbol{\mu}_x|\mathbf{x}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_x) = \frac{p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) p(\boldsymbol{\mu}_x|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)}{p(\mathbf{x}|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_x)} \propto p(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) p(\boldsymbol{\mu}_x|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). \tag{10}$$

We are going to discover the form of the posterior distribution by trying to obtain a formula that we can recognize. In particular, we are going to compute $\log p(\boldsymbol{\mu}_x|\mathbf{x}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_x)$ and try to obtain a quadratic form of $\boldsymbol{\mu}_x$ which is the form of Gaussian distributions.

$$\begin{aligned}
&\log p(\boldsymbol{\mu}_x|\mathbf{x}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_x) = \log \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x) + \log \mathcal{N}(\boldsymbol{\mu}_x|\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) + C = \\
&= -\frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu}_x)^\top {\boldsymbol{\Sigma}_x}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}_x) - \frac{1}{2}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_0)^\top {\boldsymbol{\Sigma}_0}^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_0) + C = \\
&= -\frac{1}{2}\left[\sum_{n=1}^{N}\left({\boldsymbol{\mu}_x}^\top {\boldsymbol{\Sigma}_x}^{-1}\boldsymbol{\mu}_x - 2{\boldsymbol{\mu}_x}^\top {\boldsymbol{\Sigma}_x}^{-1}\mathbf{x}_n\right) + {\boldsymbol{\mu}_x}^\top {\boldsymbol{\Sigma}_0}^{-1}\boldsymbol{\mu}_x - 2{\boldsymbol{\mu}_x}^\top {\boldsymbol{\Sigma}_0}^{-1}\boldsymbol{\mu}_0\right] + C = \\
&= -\frac{1}{2}\left[{\boldsymbol{\mu}_x}^\top\left(N{\boldsymbol{\Sigma}_x}^{-1} + {\boldsymbol{\Sigma}_0}^{-1}\right)\boldsymbol{\mu}_x - 2{\boldsymbol{\mu}_x}^\top\left({\boldsymbol{\Sigma}_x}^{-1}\sum_{n=1}^{N}\mathbf{x}_n + {\boldsymbol{\Sigma}_0}^{-1}\boldsymbol{\mu}_0\right)\right] + C. \qquad (11)
\end{aligned}$$

Now, we have to complete the squares in equation (11). To do that we know that, if A is symmetric, $(\mathbf{x}-\mathbf{y})^\top \mathbf{A}(\mathbf{x}-\mathbf{y}) = \mathbf{x}^\top\mathbf{A}\mathbf{x} + \mathbf{y}^\top\mathbf{A}\mathbf{y} - 2\mathbf{x}^\top\mathbf{A}\mathbf{y}$. Comparing equation (11) with the previous formula we can call $\mathbf{x} = \boldsymbol{\mu}_x$ and $\mathbf{A} = (N{\boldsymbol{\Sigma}_x}^{-1} + {\boldsymbol{\Sigma}_0}^{-1})$.

In order to find out who is $\mathbf{y}$ we have to make $\mathbf{A}$ appear in the expression $-2\mathbf{x}^\top\mathbf{A}\mathbf{y}$ of equation (11). We can easily achieve this multiplying by $\mathbf{A}\mathbf{A}^{-1}$, making equation (11) like

$$(c) = -\frac{1}{2}\left[{\boldsymbol{\mu}_x}^\top\mathbf{A}\boldsymbol{\mu}_x - 2{\boldsymbol{\mu}_x}^\top\mathbf{A}\left[\mathbf{A}^{-1}\left({\boldsymbol{\Sigma}_x}^{-1}\sum_{n=1}^{N}\mathbf{x}_n + {\boldsymbol{\Sigma}_0}^{-1}\boldsymbol{\mu}_0\right)\right]\right] + C \qquad (12)$$

and by calling $\mathbf{y} = \mathbf{A}^{-1}\left({\boldsymbol{\Sigma}_x}^{-1}\sum_{n=1}^{N}\mathbf{x}_n + {\boldsymbol{\Sigma}_0}^{-1}\boldsymbol{\mu}_0\right)$ we have that

$$(c) = -\frac{1}{2}(\boldsymbol{\mu}_x - \mathbf{y})^\top\mathbf{A}(\boldsymbol{\mu}_x - \mathbf{y}) + C. \qquad (13)$$

Now, if $\boldsymbol{\mu}_x$ had a multivariate Normal posterior distribution, i.e., $\boldsymbol{\mu}_x|\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$, then $\log p(\boldsymbol{\mu}_x|\mathbf{x})$ would be of the form

$$\log p(\boldsymbol{\mu}_x|\mathbf{x}) = -\frac{1}{2}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_1)^\top{\boldsymbol{\Sigma}_1}^{-1}(\boldsymbol{\mu}_x - \boldsymbol{\mu}_1) + C \qquad (14)$$

which implies, by comparing the two expressions, that the posterior distribution of $\boldsymbol{\mu}_x$ is a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_1 = \mathbf{y}$ and covariance $\boldsymbol{\Sigma}_1 = \mathbf{A}^{-1}$.

Finally, we need to compute the MAP estimate of $\boldsymbol{\mu}_x$ given $\mathbf{x}$. This estimator is defined as ${\boldsymbol{\mu}_x}^{MAP} := \arg\max_{\boldsymbol{\mu}_x} p(\boldsymbol{\mu}_x|\mathbf{x}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0, \boldsymbol{\Sigma}_x)$ which, making similar calculations as the ones done in the previous section, can be proved to be the mean of the multivariate Normal distribution, that is, ${\boldsymbol{\mu}_x}^{MAP} = \boldsymbol{\mu}_1 = \mathbf{y}$.

1.2 Exercise 2: Categorical distribution

> Given a dataset $\mathbf{X} = \{x_1, \ldots, x_N\}^\top$ in which the observations $x_n \in \{1, \ldots, K\}$ are assumed to be drawn independently from a Categorical distribution, i.e. $x_n \sim Categorical(x_n|\pi_1, \ldots, \pi_K)\ \forall n = 1, \ldots, N$:
> 1. Estimate the parameters, i.e. the category probabilities $\{\pi_k\}$ by *maximum likelihood* (ML).

We have to solve the problem

$$\boldsymbol{\pi}^{ML} := \arg\max_{\boldsymbol{\pi}} p(\mathbf{x}|\boldsymbol{\pi}) \qquad \text{subject to } \sum_{k=1}^{K} \pi_k = 1 \tag{15}$$

which is equivalent to solving

$$\boldsymbol{\pi}^{ML} := \arg\max_{\boldsymbol{\pi}} \log p(\mathbf{x}|\boldsymbol{\pi}) \qquad \text{subject to } \sum_{k=1}^{K} \pi_k = 1 \tag{16}$$

and using Lagrange multipliers this is equivalent to solving

$$\boldsymbol{\pi}^{ML} := \arg\max_{\boldsymbol{\pi}} \left[\log p(\mathbf{x}|\boldsymbol{\pi}) - \lambda \left(\sum_{k=1}^{K} \pi_k - 1\right)\right] \tag{17}$$

where λ is a sufficiently large real positive number.

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

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

$$\log p(\mathbf{x}|\boldsymbol{\pi}) = \sum_{n=1}^{N}\sum_{k=1}^{K} \log\left(\pi_k^{[x_n=k]}\right) = \sum_{n=1}^{N}\sum_{k=1}^{K} [x_n = k] \log \pi_k. \tag{18}$$

Now we have to solve the system

$$\begin{cases} \partial_{\pi_1}\left[\log p(\mathbf{x}|\boldsymbol{\pi}) - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right] = 0 \\ \partial_{\pi_2}\left[\log p(\mathbf{x}|\boldsymbol{\pi}) - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right] = 0 \\ \ldots \\ \partial_{\pi_K}\left[\log p(\mathbf{x}|\boldsymbol{\pi}) - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)\right] = 0 \end{cases} \tag{19}$$

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

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

where n_k represents how many x_n in $\mathbf{x}$ have the category k. Note that this is indeed a maximum since

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

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 derive λ substituting $\boldsymbol{\pi}^{ML}(\lambda)$ on the restriction over $\boldsymbol{\pi}$:

$$\sum_{k=1}^{K} \pi_k^{ML}(\lambda) = \frac{1}{\lambda}\sum_{k=1}^{K} n_k = 1 \iff \lambda = \sum_{k=1} n_k = N. \tag{20}$$

Therefore, the maximum likelihood estimator of π_k is

$$\pi_k^{ML} = \frac{1}{N}\sum_{n=1}^{N}[x_n = k] = \frac{n_k}{N}. \tag{21}$$

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

We assume a prior

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = Dirichlet(\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})}\prod_{k=1}^{K}\pi_k^{\alpha_k - 1}. \tag{22}$$

Using Bayes' theorem we have that

$$\begin{aligned} p(\boldsymbol{\pi}|\mathbf{x}, \boldsymbol{\alpha}) \propto p(\mathbf{x}|\boldsymbol{\pi})p(\boldsymbol{\pi}|\boldsymbol{\alpha}) \propto \prod_{n=1}^{N}\prod_{k=1}^{K}\pi_k^{[x_n=k]}\prod_{k=1}^{K}\pi_k^{\alpha_k-1} = \\ = \prod_{k=1}^{K}\pi_k^{\sum_{n=1}^{N}[x_n=k]+\alpha_k-1} = \prod_{k=1}^{K}\pi_k^{n_k+\alpha_k-1}. \end{aligned} \tag{23}$$

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

2 Q&A

Question 1

Does the covariance matrix of a multivariate Gaussian have to be positive definite or semi-positive definite?

Answer 1

The covariance matrix $\boldsymbol{\Sigma}_x$ of a multivariate Gaussian distribution has to be positive definite for the density to exist. However, if it is positive semi-definite then we end up in the degenerate case, which implies the usage of the generalized inverse and the pseudo-determinant.

Question 2

Why do not we use another Lagrange multiplier to encode the constraint $\pi_k \geq 0$ in the pseudo-likelihood of the Categorical distribution?

Answer 2

If $\pi_k < 0$ then the log-likelihood doesn't exist, so we don't need an additional constraint to check this.