Saarland University

Summary

Probabilistic Machine Learning

Winter 2020/2021

Lecturer:
Prof. Dr. Isabel Valera

Documentation:
Christian Schmidt

Probabilistic Machine Learning: Lecture 1

Isabel Valera

1 Notations

We try to be consistent with notations throughout the course. Here we give a summary of the main notation we will use. These are the same as in Bishop (2006).

- $\mathbf{x} = (x_1, \ldots, x_D)^T$: a column D-dimensional vector, i.e. lower case bold Roman letter.
- $\mathbf{x}^T$: the row vector transposed of $\mathbf{x}$.
- $\mathbf{x}_1, \ldots, \mathbf{x}_N$: N samples of a D-dimensional vector.
- $\mathbf{X}$: data matrix with n-th row the row vector $\mathbf{x}_n^T$; i.e. the n, i entry of $\mathbf{X}$ is the i-th feature/dimension of the n-th observation $\mathbf{x}_n$.
- $\mathbf{x}$: data matrix for one-dimensional variables, i.e. this is a column vector whose n-th element is x_n.
- $\mathbf{x} \neq \mathbf{x}$: we use two different typefaces. $\mathbf{x}$ denotes a N-dim vector (i.e. N samples of one-dim variables); $\mathbf{x}$ denotes a D-dimensional vector (i.e. 1 sample of a D-dimensional variable).
- $\mathbb{E}_x [f(x, y)]$: expected value with respect to the random variable x of the function $f(x, y)$.
- $\mathbb{E} [x]$: expected value of x when there is no ambiguity as to which variable is being averaged over.

2 One variable

Suppose we have one random variable X, one-dimensional.
Assume that it is distributed according to some probability distribution, for instance Gaussian:

$$X \sim \mathcal{N}(\mu, \sigma^2) \quad , \tag{1}$$

where μ and σ^2 are the parameters.

Now, we observe N samples of this variable, $x_1, \ldots, x_N$, and assume they are independent and identically distributed as a Gaussian.

Question: given this sample, how do we estimate μ and σ?

2.1 Maximum Likelihood Estimation

One possibility is to find the pair (μ, σ) that maximizes the likelihood of observing the data. In general, the likelihood of the data $\mathbf{x}$ for a probability distribution of parameters θ is defined as $p(\mathbf{x}|\theta)$. Often it is mathematically more convenient to consider the logarithm of this quantity, also called log-likelihood. This is also useful numerically, as there is less risk of underflow when taking a sum instead of a product of a very small numbers. We denote the log-likelihood as $\mathcal{L}(\mu, \sigma^2)$ and we omit from the notation the explicit dependence on x. We recall here that the (log-) likelihood is a function of the parameters θ, in this case, μ and σ^2, as the observed data $x_1, \ldots, x_N$ are given.

Given that the logarithm is monotonically increasing, the maximum of these two coincide. Then we have:

$$\mu_{MLE}, \sigma^2_{MLE} = \underset{\mu,\sigma^2}{\operatorname{argmax}} \left\{ \mathcal{L}(\mu, \sigma^2) \right\} \tag{2}$$

We have N *independent* observations, denoted as a vector $\mathbf{x}$, so we can sum the log-likelihood of each of them:

$$\begin{aligned} \mathcal{L}(\mu, \sigma^2) = p(\mathbf{x}|\mu, \sigma^2) &= \sum_i^N \log\left(\mathcal{N}(x_i|\mu, \sigma^2)\right) && (3) \\ &= \sum_i^N \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp\left(-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma^2}\right)\right) && (4) \\ &= -\frac{1}{2}\sum_i^N \log(\sigma^2) - \frac{1}{2\sigma^2}\sum_i^N (x_i-\mu)^2 + const && (5) \end{aligned}$$

where the term *const* does not depend on the parameters. Now we want to extract the maximum of that expressions:

$$\begin{cases} \frac{\partial \mathcal{L}(\mu,\sigma^2)}{\partial \mu} \equiv 0 \\ \frac{\partial \mathcal{L}(\mu,\sigma^2)}{\partial \sigma^2} \equiv 0 \end{cases} \tag{6}$$

After some algebra, we can find the analytical expressions of the Maximum Likelihood estimation:

$$\begin{aligned} \mu_{MLE} &= \frac{1}{N}\sum_{n=1}^N x_n && (7) \\ \sigma^2_{MLE} &= \frac{1}{N}\sum_{n=1}^N (x_n - \mu_{ML})^2 && (8) \end{aligned}$$

These correspond to the sample *empirical* mean and variance respectively.

Obs1: this was easy, because we could split the terms inside the sum as the samples were *independent* (and identically) distributed.
Obs2: from this estimation, we do not have idea of any measure on the uncertainty in estimating the parameters, i.e., μ_{MLE} and σ^2_{MLE} are point-estimates but no extra information was obtained from the calculations. In other words, we do not know how the ML estimate of the parameters compares with the true parameters, and thus how well the estimated parameters fit the overall true data distribution.

2.2 Posterior and Maximum a Posteriori estimates (MAP)

To address Obs2, we next treat μ, σ^2 also as random variables, each distributed according to some distribution. For instance, assume that:

$$p(\mu) = \mathcal{N}(\mu|, \mu_0, \sigma_0^2) \quad , \tag{9}$$

and keep σ^2 fixed. $p(\mu)$ is called the *prior* of the parameter μ. This distribution should be chosen in order to use some information (if any) that we know about the parameters. For instance, based on domain knowledge or by looking at the observed data; example: if the samples are all positive, perhaps it is more appropriate to set the prior as Gamma instead of Gaussian.
With this notion in mind and using Bayes' rule, we can now define the *Posterior* of μ as:

$$p(\mu|\mathbf{x}) = \frac{p(\mathbf{x}|\mu)\, p(\mu)}{p(\mathbf{x})} \propto p(\mathbf{x}|\mu)\, p(\mu) \quad , \tag{10}$$

where we made explicit the proportionality of the left-most term because in general the marginal likelihood $p(\mathbf{x})$, a.k.a. evidence, is not accessible (it involves complex calculations of integrals).

Hence, in practice, one estimates $p(\mu|\mathbf{x})$ by writing $p(\mathbf{x}|\mu)\,p(\mu)$ and attempting to then normalize the resulting expression. This is not generally possible analytically, but in this Gaussian example everything is doable. The result, after some algebra, is that the posterior is also Gaussian distributed, $p(\mu|\mathbf{x}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$, with parameters:

$$\mu_N = \frac{\sigma^2}{N\,\sigma_0^2 + \sigma^2}\,\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\,\mu_{MLE} \tag{11}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \tag{12}$$

Obs1: μ_N is a combination of the prior's mean μ_0 and the MLE's estimate μ_{MLE}. In particular, when $N = 0$ we recover the prior, i.e. $\mu_N = \mu_0$; on the opposite scenario $N \to \infty$, we get $\mu_N \to \mu_{MLE}$, i.e. the MLE estimate and the mean of the Posterior coincide.
Obs2: $\frac{1}{\sigma_N^2}$, also called *precision*, is additive; which means that for $N \to \infty$ the σ_N^2 decreases, i.e. the more data we observe, the more the posterior variance decreases and the posterior distribution becomes peaked around μ_{MLE}.
Obs3: for *finite* N, if we have $\sigma_0^2 \to \infty$ (prior with infinte variance, also called *non informative* prior), then also in this case $\mu_N \to \mu_{MLE}$, but this time with $\sigma_N^2 = \sigma^2/N$.

These analysis and results can be observed by playing around with the *jupyter* notebook included in Lecture 1 material section.[1]

Being able to characterize the posterior distribution allows not only to measure the variance of the estimated parameters (i.e. σ_N^2), but one can perform other tasks such as sampling from that distribution. However, if one is interested on a point-estimate from the posterior, one common choice is taking the value at the peak of this distribution (also called *mode*), i.e., the maximum a posteriori (MAP) estimate:

$$\mu_{MAP} = \underset{\mu}{\operatorname{argmax}}\, p(\mu|\mathbf{x}) \quad . \tag{13}$$

For the example considered above where the posterior is also Gaussian, we have $\mu_{MAP} \equiv \mu_N$. But this is not the case in general.

Obs1: given that the (usually hard-to-compute) term $p(\mathbf{x})$ in the denominator of the exact posterior does not depend on μ, to calculate μ_{MAP} one can simply consider the maximization over $p(\mathbf{x}|\mu)\,p(\mu)$. This quantity is easier to access because does not require the hard task of normalizing the posterior.

Obs2: All the results of this section can be generalized for D-dimensional variables.

3 More than one variable

Suppose now that we have more than one random variable. For simplicity, we consider the case of having two of them x_1 and x_2 and both of them one-dimensional.

Also in this case, assume that these are distributed according to some probability distribution, for instance Gaussian:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix} \right) \quad , \tag{14}$$

this is also called *Multivariate Gaussian* with parameters $\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$ and $\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{bmatrix}$.
Σ is called *covariance* matrix and to be valid it has to be *positive semi-definite* (all its eigenvalues should be nonnegative) and *symmetric*. In general, x_1 and x_2 are not *independent*, the way they depend to each other is regulated by the cross terms of the covariance matrix σ_{12}^2.

[1] https://cms.sic.saarland/pml/materials/index

3.1 Conditional and marginal distributions

When two variables are correlated with each other, a relevant question is how does one vary when the other takes a given fixed value. This answer is provided by the *conditional* distribution of x_1 *given* x_2:

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{p(x_2)} \tag{15}$$

The term in the denominator is the marginal probability of x_2 and is defined as:

$$p(x_2) = \int p(x_1, x_2)\, dx_1 \quad . \tag{16}$$

Given that this integral is not always easy to calculate, one can instead derive an expression for the conditional distribution rewriting as:

$$p(x_1|x_2) = \frac{p(x_1, x_2)}{Z_1} \tag{17}$$

where Z_1 is a normalization constant that one should find to make $p(x_1|x_2)$ a valid probability distribution. Usually finding Z_1 is easier than calculating the integral in (16). This can be done rewriting the Multivariate distribution isolating the terms depending on x_1:

$$(\mathbf{x} - \boldsymbol{\mu})^T \, \Sigma^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) \quad = \quad \frac{1}{\sigma_1^2} \left[x_1 - \left(\mu_1 + (x_2 - \mu_2) \frac{\sigma_1^2}{\sigma_{12}^2} \right) \right]^2 \tag{18}$$

This means that $p(x_1|x_2)$ is also a Gaussian, i.e. $\mathcal{N}(\mu_{x_1|x_2}, \sigma^2_{x_1|x_2})$, with parameters:

$$\mu_{x_1|x_2} \quad = \quad \mu_1 + (x_2 - \mu_2) \frac{\sigma_1^2}{\sigma_{12}^2} \tag{19}$$

$$\sigma^2_{x_1|x_2} \quad = \quad \sigma_1^2 \tag{20}$$

These calculations can be done explicitly, this is a good practice exercise for handling Gaussian distribution (e.g. completing the square). Similar thinking can be applied to derive the expression of the marginal $p(x_1)$, but in this case calculations are much more tedious! Fortunately, in the multivariate Gaussian case, to obtain the marginal distribution over a subset of multivariate normal random variables, one only needs to drop the irrelevant variables (the variables that one wants to marginalize out) from the mean vector and the covariance matrix. In our example, the result is that also $p(x_1)$ is a Gaussian distribution, namely $\mathcal{N}(x_1|\mu_1, \sigma_1^2)$.

You can find interactive plots visualizing these distributions inside the Lecture 1 material section.[2]

Obs1: the number of parameters for the 2-variable example is 2 for the mean plus 4 for the covariance matrix. More generally, it grows as $N + N(N+1)/2$, i.e. quadratically in the number of variables N.

Obs2: the Multivariate Gaussian has a single maximum (it is *unimodal*), therefore it may be too limited to fit data with many modes.

Obs3: if x_1 and x_2 come from different types of distributions, e.g. one is Gaussian and the other is Gamma, and they are correlated, there is no such thing as a Multivariate distribution with mixed types of variables. In these cases, how do we capture correlations? As an example, think of the variables (*salary*, *balance*). These are two correlated variables, if you increase your salary, then it is more likely that also your bank account's balance increases. However, one is always positive (salary), the other can go negative (balance).

These three observations show that a Multivariate Gaussian suffers to major problems: it might not capture complex situations like non-unimodal or mixed-type distributions. And it has too many parameters, growing quadratically with the number of variables.

[2] https://cms.sic.saarland/pml/materials/index

4 Latent variables and Generative models

Question: how do we capture correlations in such complex scenarios?

The answer is given by *latent* variables. The purpose of these is to capture correlations of a complicated distributions via simpler *conditional* distributions. This is formalized by a generalization of the key idea behind De Finetti's theorem, which states that the probability distribution of any infinite exchangeable sequence of Bernoulli random variables is a "mixture" of the probability distributions of independent and identically distributed sequences of Bernoulli random variables.

Consider two D-dimensional variables $\mathbf{x}_1, \mathbf{x}_2$ and a K-dimensional variable $\mathbf{z}$ (our *latent* variable). The we have:

$$p(\mathbf{x}_1, \mathbf{x}_2) = \int p(\mathbf{x}_1|\mathbf{z})\, p(\mathbf{x}_2|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z} \tag{21}$$

The quantity in the right-hand-side is a joint distribution, in general this can be complicated as generally $p(\mathbf{x}_1, \mathbf{x}_2) \neq p(\mathbf{x}_1)\, p(\mathbf{x}_2)$. The terms inside the integral are instead *factorized*, thanks to the introduction of the auxiliary latent variable $\mathbf{z}$. This is the key point of the theorem.

We can interpret $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z})$ as a model expressing the process that generated the observed data, i.e. we call this a *generative model*.

Obs1: even though we can factorize $p(\mathbf{x}_1, \mathbf{x}_2, \mathbf{z}) = p(\mathbf{x}_1|\mathbf{z})\, p(\mathbf{x}_2|\mathbf{z})\, p(\mathbf{z})$, we can still capture complex correlations between $\mathbf{x}_1, \mathbf{x}_2$ because when we marginalize out over $\mathbf{z}$ as in (21), we get a non-factorized joint $p(\mathbf{x}_1, \mathbf{x}_2)$.

4.1 Example: Gaussian Factor Analysis

An example of this is found in the following model (see Chapter 12 of Murphy (2012) for further details), where for simplicity we consider two one-dimensional variables:

$$p(x_1|z) = \mathcal{N}(x_1|w_1 z, \sigma_1^2) \tag{22}$$
$$p(x_2|z) = \mathcal{N}(x_2|w_2 z, \sigma_2^2) \tag{23}$$
$$p(z) = \mathcal{N}(z|\mu_0, \sigma_0^2) \tag{24}$$

Let's calculate the joint $p(x_1, x_2) = \int \mathcal{N}(x_1|w_1 z, \sigma_1^2)\, \mathcal{N}(x_2|w_2 z, \sigma_2^2)\, \mathcal{N}(z|\mu_0, \sigma_0^2)\, dz$. Without loss of generality, we can set and $\sigma_0^2 = 1$. Rewriting everything explicitly, we obtain that the marginal distribution of $(x_1, x_2)^T$ is a Multivariate Gaussian distribution with mean and covariance:

$$\boldsymbol{\mu} = (w_1\mu_0, w_2\mu_0)^T \tag{25}$$
$$\boldsymbol{\Sigma} = \begin{bmatrix} w_1^2 & w_1 w_2 \\ w_1 w_2 & w_2^2 \end{bmatrix} + \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \quad . \tag{26}$$

The expression for $\boldsymbol{\Sigma}$ highlights that we now have correlations between the two variables, thanks to the terms $w_1 w_2$ in the non-diagonal entries.

Obs1: even though x_1 and x_2 were *conditionally* independent given z, their joint distribution is not factorized anymore, they have indeed a non-diagonal covariance matrix Σ.

Obs2: In general models, the integral in $p(x_1, \ldots, x_D) = \int \prod_{d=1}^{D} p(x_d|\mathbf{z})\, p(\mathbf{z})\, d\mathbf{z}$ does not have closed-form solution.

References

C. M. Bishop, *Pattern recognition and machine learning* (Springer, 2006).

K. P. Murphy, *Machine learning: a probabilistic perspective* (MIT press, 2012).