

# Maximum likelihood estimation

Herman Kamper

2023-02

# Probabilistic approaches in machine learning

In many machine learning problems it is useful to have a way to deal with uncertainty.

Probability theory gives us a principled way to do this.

A probabilistic perspective is also often useful for defining and combining loss functions.

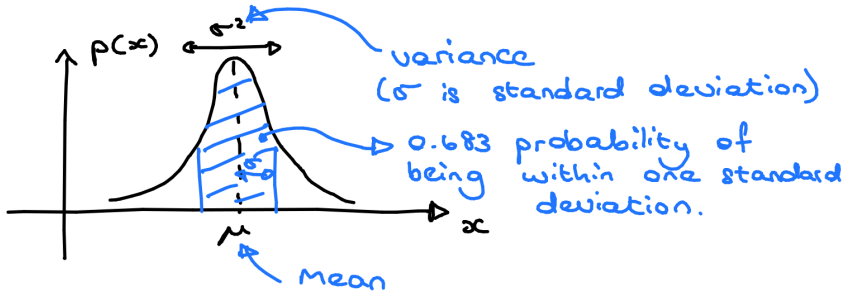
But to be able to follow a probabilistic approach, we need a way to estimate the parameters in a probabilistic model.

**Maximum likelihood estimation** is one of the most fundamental methods to set the parameters in a probabilistic model.

In this note we look at estimating the parameters of a Gaussian distribution. But we will see that this same approach can be used in many other machine learning models, with the steps proceeding exactly as we do here.

# The Gaussian distribution

$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



# Maximum likelihood estimation for a univariate Gaussian

Given samples  $x^{(1)}, x^{(2)}, \dots, x^{(N)}$  from a univariate Gaussian with unknown mean and variance, could we devise a way (maybe with a “loss function”) to find optimal estimates of the mean  $\hat{\mu}$  and variance  $\hat{\sigma}^2$ ?

How would these estimates compare with the sample mean and variance?

We assume the samples are *independent and identically distributed* (IID), each a draw from the Gaussian  $\mathcal{N}(x; \mu, \sigma^2)$ . (Remember, we do not know the mean or the variance, we only get to see the samples.)

## Example

# MLE for a univariate Gaussian

We are given IID samples  $\{x^{(n)}\}_{n=1}^N$ , each a draw from  $\mathcal{N}(x; \mu, \sigma^2)$ .

The joint density of the samples:

Idea: We choose the  $(\mu, \sigma^2)$  that maximises the above, i.e.

This is called the *likelihood* of the parameters. The overall approach is therefore called *maximum likelihood estimation*.

# Estimating the parameters

Instead of maximising the likelihood directly, it is often easier to maximise the log likelihood:

This is because the log of the product becomes the sum of logs, and we will see in a second why this is useful.

I also like minimising loss functions (instead of maximising things), so let us minimise the *negative log likelihood*:

**Strategy:** Set  $\frac{\partial J}{\partial \mu} = 0$  and  $\frac{\partial J}{\partial \sigma^2} = 0$  and solve jointly to find  $\hat{\mu}$  and  $\hat{\sigma}^2$ .

First we write out the negative log likelihood a bit more:

$$\begin{aligned} J(\mu, \sigma^2) &= - \sum_{n=1}^N \log \left[ \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{(x^{(n)} - \mu)^2}{2\sigma^2} \right\} \right] \\ &= \\ &= \end{aligned}$$

Then take the partial derivatives with respect to  $\mu$ :

Then take the partial derivatives with respect to  $\sigma^2$ :

And set the partial derivatives equal to zero:

$$\frac{\partial J}{\partial \mu} = 0 :$$

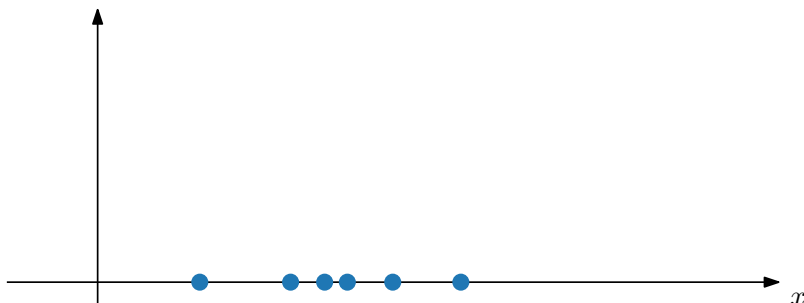
$$\frac{\partial J}{\partial \sigma^2} = 0 :$$

## More about the likelihood

$$\begin{aligned} p\left(x^{(1)}, x^{(2)}, \dots, x^{(N)}\right) &= \mathcal{N}\left(x^{(1)}; \mu, \sigma^2\right) \cdot \mathcal{N}\left(x^{(2)}; \mu, \sigma^2\right) \cdots \mathcal{N}\left(x^{(N)}; \mu, \sigma^2\right) \\ &= \prod_{n=1}^N \mathcal{N}\left(x^{(n)}; \mu, \sigma^2\right) \end{aligned}$$

Negative log likelihood (NLL):

$$J(\mu, \sigma^2) = -\log \prod_{n=1}^N \mathcal{N}\left(x^{(n)}; \mu, \sigma^2\right) = -\sum_{n=1}^N \log \mathcal{N}\left(x^{(n)}; \mu, \sigma^2\right)$$





# MLE for a multivariate Gaussian

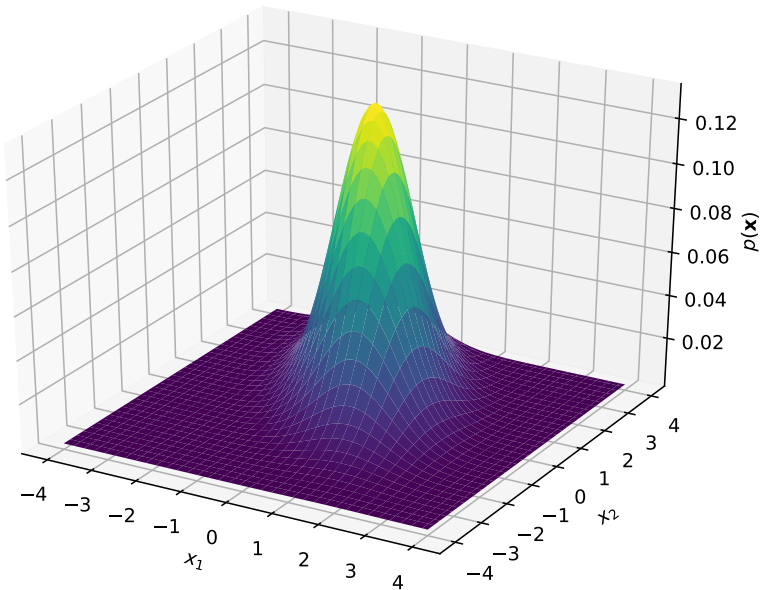
Given samples  $\{\mathbf{x}^{(n)}\}_{n=1}^N$  from a multivariate Gaussian:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

then it can be shown in a similar way that the MLEs are:

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}^{(n)}$$

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}^{(n)} - \hat{\boldsymbol{\mu}})(\mathbf{x}^{(n)} - \hat{\boldsymbol{\mu}})^\top$$



## Videos covered in this note

- [Gaussians 1: Maximum likelihood estimation](#) (20 min)