Maximum likelihood estimation for Gaussians

Herman Kamper

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 of a probabilistic model given some observed data.

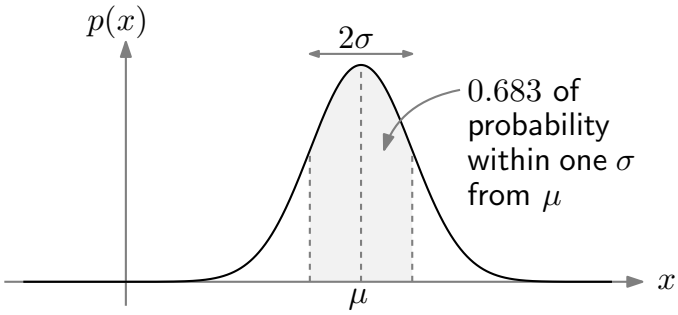
Maximum likelihood estimation is one of the most fundamental methods to set the parameters of a probabilistic model.

This note will look at the general principles of maximum likelihood estimation, and then apply this specifically to the Gaussian distribution. But the same steps can be followed to find the parameters of other probabilistic machine learning models.

The univariate Gaussian distribution

The univariate Gaussian has the following probability density function:

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



It is fully characterised by its mean μ and variance σ^2.

Parameter estimation for a Gaussian

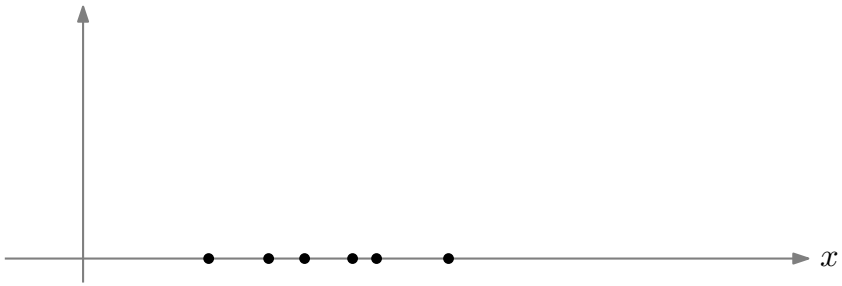
Given samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ from a univariate Gaussian with unknown mean and variance, is there a way (maybe with a "loss function") to find optimal estimates of the mean $\hat{\mu}$ and variance $\hat{\sigma}^2$?

We assume the samples are *independent and identically distributed* (IID), each a draw from the Gaussian $\mathcal{N}(x; \mu, \sigma^2)$. The *independent* in IID means that each sample is drawn separately: a particular sample doesn't tell us anything about another sample. The *identically distributed* means that, although samples are drawn separately, they all come from the same underlying distribution.

Remember, we do not know the mean or the variance of the distribution, we only get to see the samples.

Example

You observe the following data:



What would you guess is the mean of the data?

$$\text{sample mean} = \frac{1}{N} \sum_{n=1}^{N} x^{(n)}$$

We will see that the maximum likelihood estimate of the mean $\hat{\mu}$ turns out to be exactly this equation.

Maximum likelihood estimation for a univariate Gaussian

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

The joint density of the samples:

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

This is because each sample is independent (we can take the product of their densities to get the joint) and identically distributed (each density is a Gaussian with the same μ and σ^2).

Some settings for (μ, σ^2) will give high values for the data, others will give low values.

Idea: We choose the (μ, σ^2) that maximises the above, i.e.

$$\hat{\mu}, \hat{\sigma}^2 = \arg\max_{\mu, \sigma^2} \prod_{n=1}^{N} \mathcal{N}(x^{(n)}; \mu, \sigma^2)$$

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

This same terminology is used for any distribution, not just Gaussians.

Estimating the parameters

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

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

Taking the log helps us because the product then becomes a sum of logs, which is especially useful when there are exponentials involved, as we will see in a second.

Maximising the log likelihood is equivalent to maximising the original likelihood due to the monotonicity of the logarithm: if $a > b$, then $\log a > \log b$.

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

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

This is now our loss function.

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:

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

Then take the partial derivatives with respect to μ:

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

Setting this to zero gives:

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

Then take the partial derivatives with respect to σ^2:

$$\begin{aligned}\frac{\partial J}{\partial \sigma^2} &= \sum_{n=1}^{N}\left[\frac{1}{2\sigma^2}-\frac{(x^{(n)}-\mu)^2}{2\sigma^4}\right] \\ &= \frac{N}{2\sigma^2}-\frac{1}{2\sigma^4}\sum_{n=1}^{N}(x^{(n)}-\mu)^2\end{aligned}$$

Setting this to zero and using the optimal mean gives:

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)}-\hat{\mu})^2$$

More about the likelihood

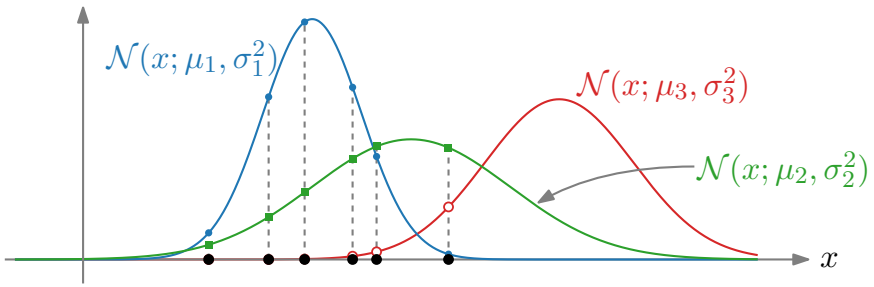
$$p\left(x^{(1)}, x^{(2)}, \ldots, x^{(N)}\right) = \mathcal{N}\left(x^{(1)}; \mu, \sigma^2\right) \cdot \mathcal{N}\left(x^{(2)}; \mu, \sigma^2\right) \cdots \mathcal{N}\left(x^{(N)}; \mu, \sigma^2\right)$$
$$= \prod_{n=1}^{N} \mathcal{N}\left(x^{(n)}; \mu, \sigma^2\right)$$

Negative log likelihood (NLL):

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

For maximum likelihood estimation, it is often useful to think of the data $\left\{x^{(n)}\right\}_{n=1}^{N}$ as fixed while the parameters are being wiggled. (This is also why we talk about the "likelihood of the parameters" rather than the "likelihood of the data".)

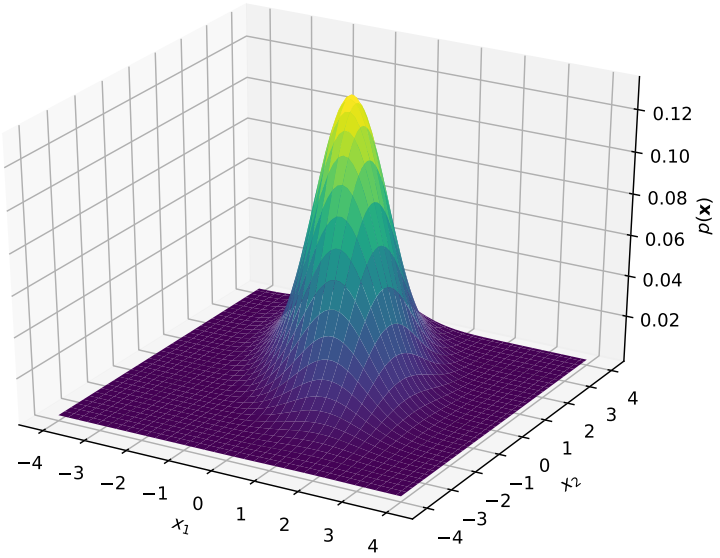
Given the data on the x-axis below, which of the three distributions would result in the highest likelihood, i.e. the smallest negative log likelihood?



Maximum likelihood estimation for a multivariate Gaussian

A multivariate Gaussian over vectors $\mathbf{x} \in \mathbb{R}^D$ are defined as:

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



Given samples $\left\{\mathbf{x}^{(n)}\right\}_{n=1}^N$ from a multivariate Gaussian, it can be shown in a similar way that the maximum likelihood estimates for the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ are given by:

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

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

(You need vector and matrix derivatives to proove these.)

Videos covered in this note

- Gaussians 1: Maximum likelihood estimation (20 min)