Maximum likelihood estimation

For a Gaussian distribution

Herman Kamper

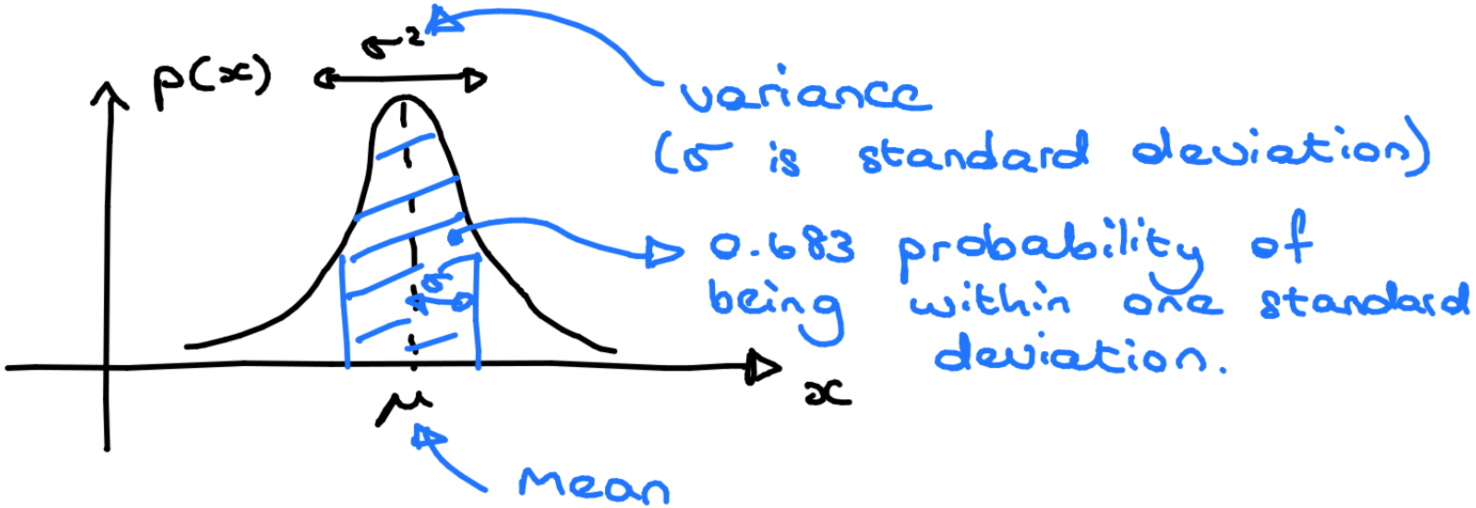
http://www.kamperh.com/

Probabilistic approaches in machine learning

- In many applications it is useful to deal with uncertainty
- Probability theory gives a principled way to do this
- Probabilistic perspective often useful for defining and combining loss functions
- Need a way to estimate the parameters in a probabilistic model
- **Maximum likelihood estimation** is one of the most fundamental methods

The Gaussian distribution

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

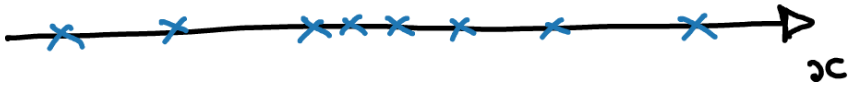


Maximum likelihood estimation (MLE)

Given samples $x^{(1)}, x^{(2)}, \ldots, x^{(N)}$ 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 $\mathcal{N}(x; \mu, \sigma^2)$.



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

MLE for univariate Gaussian

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

Joint density:

$$p(x^{(1)},\ldots,x^{(N)}) = \mathcal{N}(x^{(1)};\mu,\sigma^2) \times \mathcal{N}(x^{(2)};\mu,\sigma^2) \times \cdots \times \mathcal{N}(x^{(N)};\mu,\sigma^2)$$

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

Some settings of (μ,σ^2) will give high value on data, others low.

Idea: Choose (μ,σ^2) that maximises this, i.e.

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

Called the likelihood of the parameters.

This approach is therefore called

maximum likelihood estimation.

Terminology used for any distribution, not just Gaussians.

Estimating the parameters

Instead of maximizing likelihood directly, it is often easier to maximize log likelihood:

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

We like minimizing a loss, so let's minimize the negative log likelihood:

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

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

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

$$= -\sum_{n=1}^{N}\left[-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x^{(n)}-\mu)^2}{2\sigma^2}\right]$$

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

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

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

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

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

$$\frac{\partial J}{\partial \sigma^2} = \frac{N}{2}\left(\frac{1}{\sigma^2}\right) - \frac{1}{2\sigma^4}\sum_{n=1}^{N}(x^{(n)}-\mu)^2$$

$\frac{\partial J}{\partial \mu} = 0$: $\boxed{\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}}$

In Python: numpy.std

$\frac{\partial J}{\partial \sigma^2} = 0$: $\boxed{\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)}-\hat{\mu})^2}$

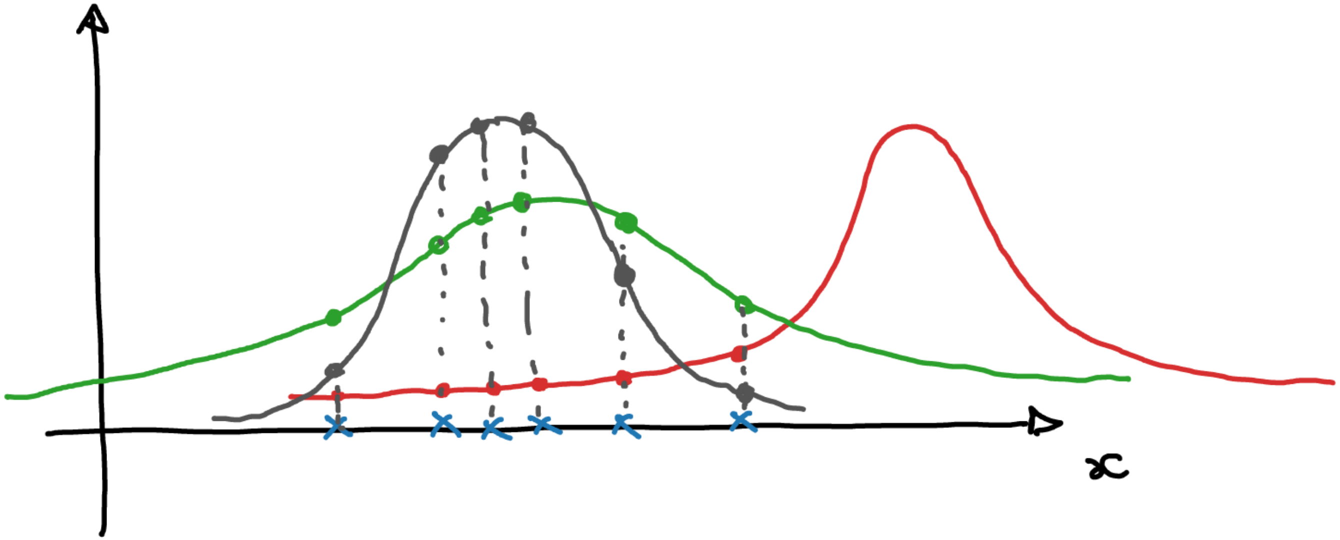
More about the likelihood

$$p(x^{(1)}, \ldots, x^{(N)}) = \mathcal{N}(x^{(1)}; \mu, \sigma^2) \times \mathcal{N}(x^{(2)}; \mu, \sigma^2) \times \ldots \times \mathcal{N}(x^{(n)}; \mu, \sigma^2)$$

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

For MLE: Think of the data $\{x^{(n)}\}_{n=1}^{N}$ as fixed.

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



MLE for the multivariate Gaussian

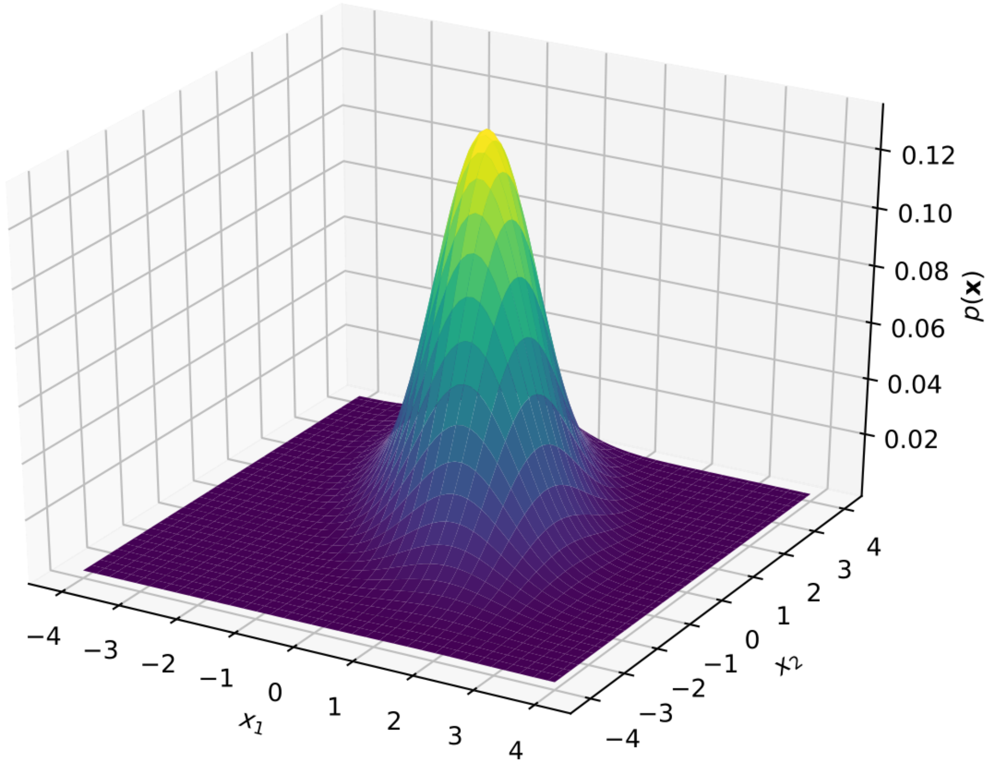
Given samples $\left\{\mathbf{x}^{(n)}\right\}_{n=1}^{N}$ from a multivariate Gaussian: $\mathbf{x} \in \mathbb{R}^D$

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

it can be shown in a similar way that the maximum likelihood estimates are:

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

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



Need vector and matrix derivatives to derive this