Machine Learning Course - CS-433

Maximum Likelihood

Sept 29, 2020

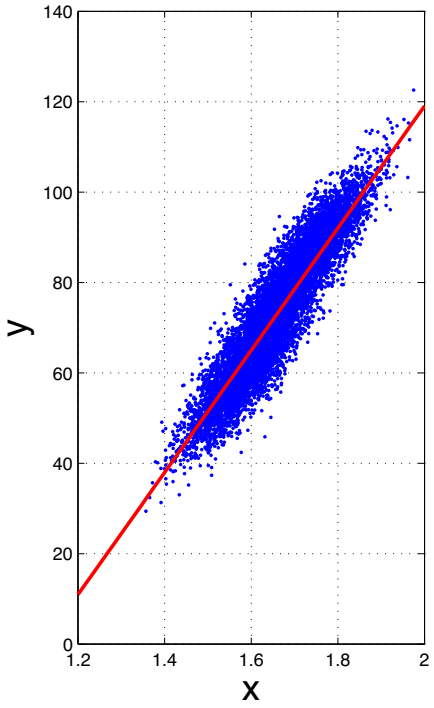
minor changes by Martin Jaggi 2020,2019,2018, minor changes by Rüdiger Urbanke 2017,
minor changes by Martin Jaggi 2016 ©Mohammad Emtiyaz Khan 2015

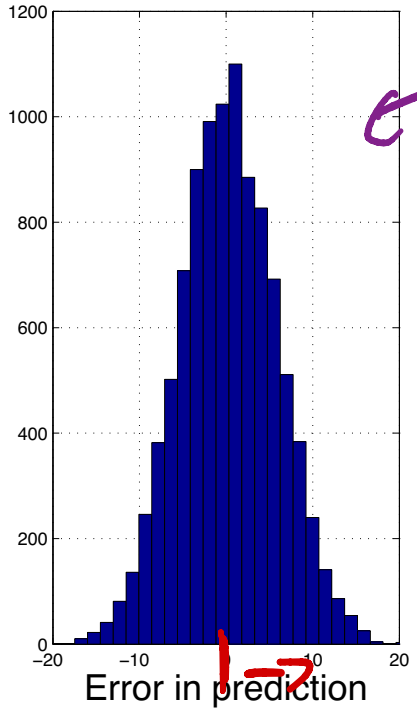
Last updated on: September 29, 2020

EPFL

Motivation

In the previous lecture we arrived at the least-squares problem in the following way: we postulated a particular cost function (square loss) and then, given data, found that model that minimizes this cost function. In the current lecture we will take an alternative route. The final answer will be the same, but our starting point will be probabilistic. In this way we find a second interpretation of the least-squares problem.





← Histogram of errors

$e_n = y_n - x_n^T w$

Gaussian distribution and independence

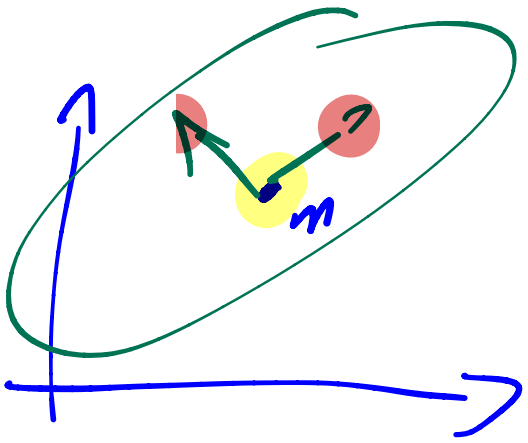
Recall the definition of a Gaussian random variable in $\mathbb{R}$ with mean μ and variance σ^2. It has a density of

$$p(y \,|\, \mu, \sigma^2) = \mathcal{N}(y \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right].$$

In a similar manner, the density of a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ (which must be a positive semi-definite matrix) is

$$\mathcal{N}(\mathbf{y} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^D \, det(\boldsymbol{\Sigma})}} \exp\left[-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right].$$

Also recall that two random variables X and Y are called *independent* when $p(x, y) = p(x)p(y)$.



A probabilistic model for least-squares

We assume that our data is generated by the model,

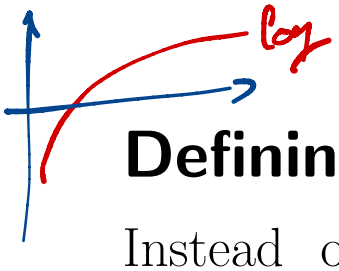
$$y_n = \mathbf{x}_n^\top \mathbf{w} + \epsilon_n,$$

where the ϵ_n (the noise) is a zero-mean Gaussian random variable with variance σ^2 and the noise that is added to the various samples is independent of each other, and independent of the input. Note that the model $\mathbf{w}$ is unknown.

Therefore, given N samples, the likelihood of the data vector $\mathbf{y} = (y_1, \cdots, y_N)$ given the input $\mathbf{X}$ (each row is one input) and the model $\mathbf{w}$ is equal to

$$p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = \prod_{n=1}^{N} p(y_n \mid \mathbf{x}_n, \mathbf{w}) = \prod_{n=1}^{N} \mathcal{N}(y_n \mid \mathbf{x}_n^\top \mathbf{w}, \sigma^2).$$

The probabilistic view point is that we should maximize this likelihood over the choice of model $\mathbf{w}$. I.e., the "best" model is the one that maximizes this likelihood.



Defining cost with log-likelihood

Instead of maximizing the likelihood, we can take the logarithm of the likelihood and maximize it instead. Expression is called the log-likelihood (LL).

$$\mathcal{L}_{\text{LL}}(\mathbf{w}) := \log p(\mathbf{y} \mid \mathbf{X}, \mathbf{w}) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{w})^2 + \text{const.}$$

Compare the LL to the MSE (mean squared error)

$$\mathcal{L}_{\text{LL}}(\mathbf{w}) = -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{w})^2 + \text{const}$$

$$\mathcal{L}_{\text{MSE}}(\mathbf{w}) = \frac{1}{2N} \sum_{n=1}^{N} (y_n - \mathbf{x}_n^\top \mathbf{w})^2$$

Maximum-likelihood estimator (MLE)

It is clear that maximizing the LL is equivalent to minimizing the MSE:

$$\arg\min_{\mathbf{w}} \mathcal{L}_{\text{MSE}}(\mathbf{w}) = \arg\max_{\mathbf{w}} \mathcal{L}_{\text{LL}}(\mathbf{w}).$$

This gives us another way to design cost functions.

MLE can also be interpreted as finding the model under which the observed data is most likely to have been generated from (probabilistically). This interpretation has some advantages that we discuss now.

Properties of MLE

MLE is a *sample* approximation to the *expected log-likelihood*:

$$\mathcal{L}_{\text{LL}}(\mathbf{w}) \approx \mathbb{E}_{p(y,\mathbf{x})}\big[\log p(y \mid \mathbf{x}, \mathbf{w})\big]$$

MLE is consistent, i.e., it will give us the correct model assuming that we have a sufficient amount of data. (can be proven under some weak conditions)

$$\mathbf{w}_{\text{MLE}} \longrightarrow^{p} \mathbf{w}_{\text{true}} \quad \text{in probability}$$

The MLE is asymptotically normal, i.e.,

$$(\mathbf{w}_{\text{MLE}} - \mathbf{w}_{\text{true}}) \longrightarrow^{d} \frac{1}{\sqrt{N}}\mathcal{N}(\mathbf{w}_{\text{MLE}} \mid \mathbf{0}, \mathbf{F}^{-1}(\mathbf{w}_{\text{true}}))$$

where $\mathbf{F}(\mathbf{w}) = -\mathbb{E}_{p(\mathbf{y})}\left[\frac{\partial^2 \mathcal{L}}{\partial \mathbf{w} \partial \mathbf{w}^\top}\right]$ is the Fisher information.

MLE is efficient, i.e. it achieves the Cramer-Rao lower bound.

$$\text{Covariance}(\mathbf{w}_{\text{MLE}}) = \mathbf{F}^{-1}(\mathbf{w}_{\text{true}})$$

Another example

We can replace Gaussian distribution by a Laplace distribution.

$$p(y_n \,|\, \mathbf{x}_n, \mathbf{w}) = \frac{1}{2b} e^{-\frac{1}{b}|y_n - \mathbf{x}_n^\top \mathbf{w}|}$$