

Machine Learning Course - CS-433

Maximum Likelihood

Oct 3, 2017

©Mohammad Emtiyaz Khan 2015

minor changes by Martin Jaggi 2016

small changes by Rüdiger Urbanke 2017

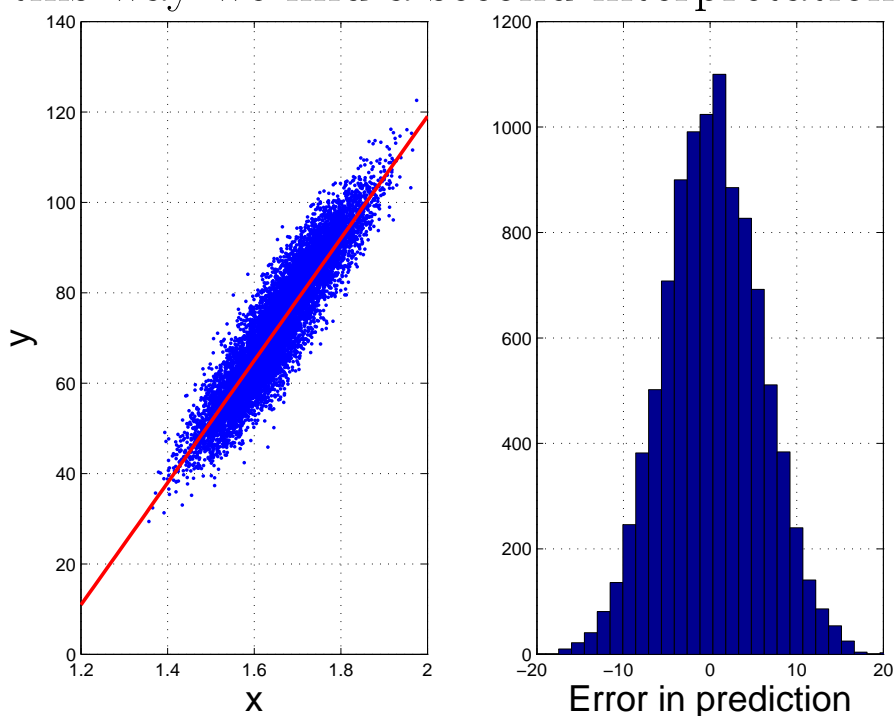
Last updated on: September 25, 2017



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

In the previous lecture we arrived at the least-squares (LS) 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 LS problem.



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

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

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

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

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

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 $\mathbf{y} = (y_1, \dots, y_N)$ given the input $\mathbf{x} = (x_1, \dots, x_N)$ 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. Not sur-

prisingly the resulting expression is called the [log-likelihood](#) (LL). As mentioned, we take this as our cost function that should be maximized. We get

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

Compare the LL to the MSE (mean squared error) that we discussed in our last lecture:

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

It is clear that maximizing the LL is equivalent to minimizing the MSE. Since we maximize the likelihood (log-likelihood), the model that we get when performing this maximization, is called the *maximum-likelihood estimate* (MLE). We will therefore write in the sequel the acronym MLE instead of LL. So we have

$$\arg \min_{\mathbf{w}} \mathcal{L}_{mse}(\mathbf{w}) = \arg \max_{\mathbf{w}} \mathcal{L}_{MLE}(\mathbf{w}).$$

This gives us an alternative interpretation of the LS problem. This interpretation has some advantages that we discuss now.

Properties of MLE

Note that the MLE cost function is a *sample* approximation to the *expected log-likelihood*:

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

Because of the above, one can show that under some conditions the MLE estimate is **consistent**, i.e., it will give us the correct model assuming that we have a sufficient amount of data.

$$\mathbf{w}_{MLE} \xrightarrow{p} \mathbf{w}_{TRUE} \quad \text{in probability}$$

The MLE is asymptotically normal,

$$(\mathbf{w}_{MLE} - \mathbf{w}_{TRUE}) \xrightarrow{d} \frac{1}{\sqrt{N}} \mathcal{N}(\mathbf{w}_{MLE} | \mathbf{0}, \mathbf{F}^{-1}(\mathbf{w}_{TRUE}))$$

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.

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

Another example

In the above derivation we assumed that the noise is Gaussian. But this is not the only possible choice. It is instructive to rederive the expression assuming instead a Laplace distribution:

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