annotated version

Machine Learning Course - CS-433

Maximum Likelihood

Oct 4, 2016

minor changes by Martin Jaggi 2016

Statistical View

Motivation

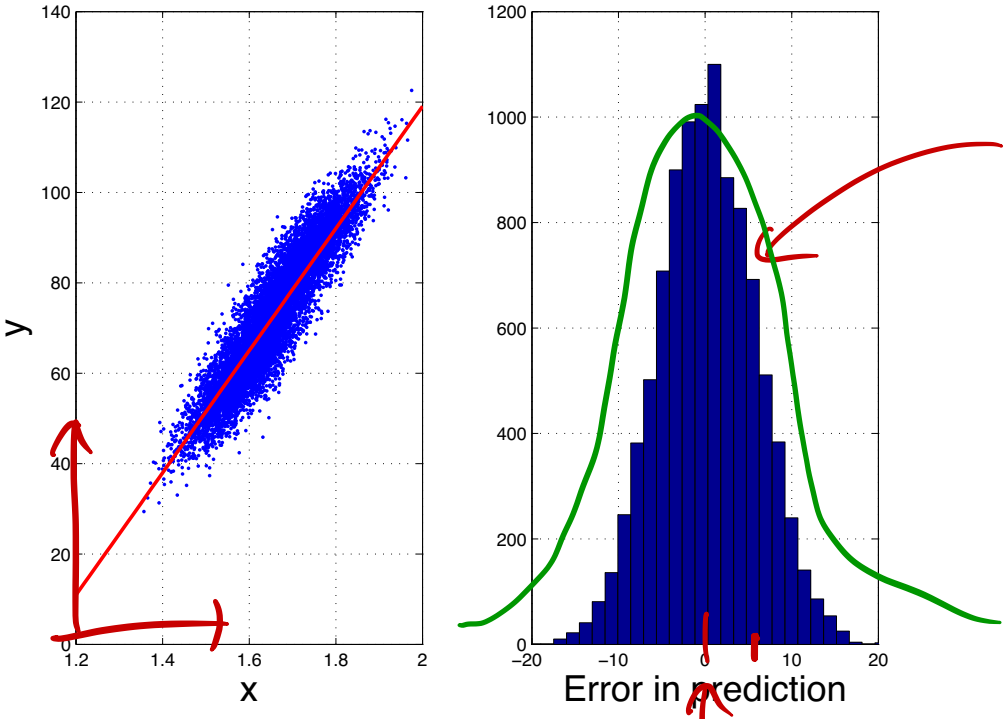
Many important questions remain unanswered: Under what conditions is least-squares optimal, and for what kind of data? How confident are we in our estimates? Will we be more confident with more data? If yes, how much more data? Does our 'confidence' *converge* with increasing data size?

Another one: Given new data, how to design a *good* cost function, e.g. in the presence of outliers?

We will answer these questions by assuming that our data is generated from a *probability distribution*.

more data (higher N)

$N = 1000$



histogram of errors

$y_n - x_n^T w$

zero

Gaussian distribution and independence

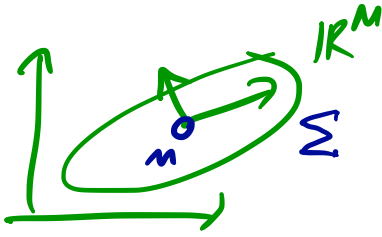
Gaussian random variable in $\mathbb{R}$ with mean μ and variance σ^2:

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

Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$ (p.s.d. matrix):

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

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 mistakes ϵ_n are Gaussian with mean 0 and variance σ^2 and are mutually independent.

i.i.d. = independent + identically distributed

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

Another way of expressing this:

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

This defines the likelihood of observing $\mathbf{y}$ given $\mathbf{X}$ and $\mathbf{w}$.

Defining cost with log-likelihood

The log-likelihood is simply the log of the likelihood.

$$\mathcal{L}_{lik}(\mathbf{w}) := \log\, p(\mathbf{y} \,|\, \mathbf{X}, \mathbf{w})$$

Compare this to MSE.

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

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

Maximum likelihood estimator (MLE)

Obviously, we have the following:

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

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).

There are many other advantages of this interpretation.

Properties of MLE

MLE is a sample approximation to the expected log-likelihood.

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

MLE is consistent under some conditions (check Wikipedia).

as $N \to \infty$

$$\mathbf{w}_{mle} \xrightarrow{①}{}^{p} \mathbf{w}_{true} \quad \text{in probability}$$

① $\mathbf{w}$ converges

MLE is asymptotically normal.

② distribution of $\mathbf{w}$ converges

$$(\mathbf{w}_{mle} - \mathbf{w}_{true}) \xrightarrow{②}{}^{d} \frac{1}{\sqrt{N}}\mathcal{N}(\mathbf{w}_{mle} \,|\, \mathbf{0}, \mathbf{F}^{-1}(\mathbf{w}_{true}))$$

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

We can replace the Gaussian distribution by the Laplace distribution.

$\varepsilon_n = y_n - \mathbf{x}_n^\top \mathbf{w}$

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

$\to \mathcal{L}_{lik} = MAE(\mathbf{w})$

Additional Notes

ToDo

1. Understand the big picture: why is the probabilistic view useful?
2. Get familiar with the notation $p(x \mid \theta)$.
3. Revise the Gaussian distribution (you must know the formula well). See Bishop Chapter 2 or Wikipedia for details.
4. Clearly understand the relationship between MSE and log-likelihood for least-squares.
5. Read the Wikipedia page for the definition of consistency and efficiency of an estimator.
6. Derive the MAE cost function using the log-likelihood framework with a Laplace distribution.