Maximum Likelihood

Stefan Haufe

November 6, 2016

1 Recap

Let X be a random variable and x be a realization of that variable. For discrete X, the probability of X attaining x is denoted by $\Pr[X = x]$. The probability as a function of x is denoted $P(x)$, and we have $\sum_x P(x)x = 1$. For continuous X, the probability of X falling into the interval $[a, b]$ is denoted by $\Pr[a \leq X \leq b] = \int_a^b p(x)\,\mathrm{d}x$. Here, the *probability density function* (pdf) $p(x)$ assumes a similar role as $P(x)$ for discrete data, and we have $\int_x p(x)\,\mathrm{d}x = 1$. $p(x)$ and $P(x)$ are are referred to as *(probability) distributions*.

For two discrete random variables X and Y, the probability $\Pr[X = x \text{ and } Y = y]$ as a function of x and y is denoted by $P(x, y)$. For two jointly continuous variables X and Y, $p(x, y)$ denotes the corresponding joint pdf. $p(x, y)$ and $P(x, y)$ are referred to as the *joint distribution* of X and Y.

The following rules apply to both discrete and continuous random variables.

The *conditional distribution* $p(x|y)$ denotes the distribution of X given that $Y = y$. It is obtained by normalizing the joint distribution $p(x, y)$ by $p(y)$.

$$p(x|y) = \frac{p(x, y)}{p(y)} \,. \tag{1}$$

Theorem 1.1 (Law of total probability)**.** In order to obtain $p(y)$, we can apply the *law of total probability*.

$$p(y) = \int_x p(x, y)\,\mathrm{d}x = \int_x p(y|x)p(x)\,\mathrm{d}x \,. \tag{2}$$

$p(x)$ or $p(y)$ obtained that way are called *marginal distributions*.

Rearranging (1), the joint distribution of X and Y can be expressed as

$$p(x, y) = p(x|y)p(y) = p(y|x)p(x) \,. \tag{3}$$

Theorem 1.2 (Bayes' formula)**.** From the definition of conditional distributions above it follows that

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} \,. \tag{4}$$

Definition 1.1 (Independence)**.** Two random variables X and Y are independent, if $p(x|y) = p(x)$. In other words, if the realization $Y = y$ does not change the distribution of X. It follows that $p(y|x) = p(y)$, and that $p(x, y) = p(x)p(y)$ (i.e., the joint distribution factorizes). The latter is the formal definition of independence.

Definition 1.2 (Expected value)**.** The expected value of a random variable X is $\mathbb{E}[X] = \int_x xp(x)\,\mathrm{d}x$.

Functions $Y = f(X)$ of random variables are also random variables, for which all of the above holds.

2 Introduction to Maximum Likelihood

In the previous lecture, we have derived decision rules based on discriminant functions for classification tasks.

$$g(\mathbf{x}) = p(\omega_1|\mathbf{x}) - p(\omega_2|\mathbf{x}) > c \tag{5}$$

$$\Leftrightarrow (\mathbf{x}|\omega_1)p(\omega_1) - (\mathbf{x}|\omega_2)p(\omega_2) > c \tag{6}$$

These rules require knowledge of the *class-conditional distributions* $p(\mathbf{x}|\omega_i)$ and the *prior probabilities* $p(\omega_i)$. In the lecture, we have assumed that these distributions are known. In practice, they need to be estimated from data. This is a central machine learning problem.

In the fish factory example, we might have taken the effort to manually putdown lightness, weight and correct classification (into seabass or salmon) for a number of fishes in a logbook.

Mathematically, this gives rise to a sample $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where the *feature vectors* $\mathbf{x}_i \in \mathbb{R}^d$ contain measurements (e.g., weight and lightness), and the corresponding *labels* $y_i \in \{\omega_1, \ldots, \omega_c\}$ indicate the class membership (e.g., species).

We here assume that each sample has been drawn independently from the others, such that $p(\mathbf{x}_i, \mathbf{x}_j) = p(\mathbf{x}_i)p(\mathbf{x}_j)$ for any indices i, j. We further assume that any sample $\mathbf{x}_i$ for which $y_i = \omega_c$ is drawn from the same distribution $p(\mathbf{x}|\omega_c)$. Thus, the $\mathbf{x}_i$ characterizing samples of class ω_j are *i.i.d. (independent and identically distributed)* random variables.

Estimating the prior probabilities $p(\omega_j)$ is relatively easy as each is just a single number. The class-conditional distribution $p(\mathbf{x}|\omega_j)$ are however continuous functions supported in $\mathbb{R}^d$. We can simplify the estimation by assuming parametric distributions $p_{\boldsymbol{\theta}_j}(\mathbf{x}|\omega_j) = p(\mathbf{x}|\omega_j, \boldsymbol{\theta}_j)$, like the Gaussian (normal) distribution.

Let us focus on a single class ω, and let $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be the data sampled from that class. The function

$$p(\mathcal{D}|\omega, \boldsymbol{\theta}) = p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}) \tag{7}$$

is interpreted as the *likelihood* of the data given the distribution parameters $\boldsymbol{\theta}$. Note that it is not a probability distribution when treated as a function of $\mathcal{D}$.

The *maximum likelihood principle* selects the parameters $\boldsymbol{\theta}$ such that the likelihood of the observed data becomes maximal (i.e., are best supported by the data).

In practice, it is often more convenient to maximize the log-likelihood

$$l(\boldsymbol{\theta}) = \ln p(\mathcal{D}|\boldsymbol{\theta}) = \ln \prod_{i=1}^{n} p(\mathbf{x}_i|\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln p(\mathbf{x}_i|\boldsymbol{\theta}) \tag{8}$$

The ML estimate is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) \; . \tag{9}$$

How the solution can be found depends on the properties of $p(\mathbf{x}_i|\boldsymbol{\theta})$.

In general, if l is a smooth (say, twice differentiable) function, a necessary condition for a local maximum is that the gradient is zero. The gradient is the vector of partial derivatives

$$\boldsymbol{\nabla}_l(\boldsymbol{\theta}) = \left(\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p}\right)^\top \; . \tag{10}$$

To distinguish local maxima at $\boldsymbol{\nabla}_{\boldsymbol{\theta}} = \mathbf{0}$ from local minima, we need to check that the Hessian (Hesse matrix) at that point be negative definite (has all negative Eigenvalues). The Hessian is given by

$$H_l(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \partial \theta_p} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_1} & \cdots & \frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_p \partial \theta_p} \end{pmatrix} . \tag{11}$$

3 Multivariate Gaussian distribution

Assume that the samples are drawn from a multivariate Gaussian distribution, as in the linear/quadratic discriminant function examples in the previous lecture,

$$p(\mathbf{x}|\boldsymbol{\mu}, \Sigma) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] . \tag{12}$$

Notation: $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

Gaussian noise is reasonable under the central limit theorem (sum of independent noise sources), and often observed in practice.

The likelihood of the data is

$$p(\mathcal{D}|\boldsymbol{\mu}, \Sigma) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu})\right] . \tag{13}$$

The log-likelihood is

$$l(\boldsymbol{\mu}, \Sigma) = \sum_{i=1}^{n} -\frac{1}{2}\ln\left[(2\pi)^d|\Sigma|\right] - \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) \tag{14}$$

$$= -\frac{n}{2}\ln\left[(2\pi)^d|\Sigma|\right] - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) . \tag{15}$$

The partial derivative w.r.t. $\boldsymbol{\mu}$ is

$$\frac{\partial l(\boldsymbol{\mu}, \Sigma)}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x}_i-\boldsymbol{\mu}) . \tag{16}$$

At the maximum, it must hold

$$\sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x}_i-\hat{\boldsymbol{\mu}}) = 0 . \tag{17}$$

From which follows that

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i . \tag{18}$$

This is the sample mean, which is an expected result.

Note that $\hat{\boldsymbol{\mu}}$ is not a function of Σ. So we can estimate Σ in a second step. If both partial derivatives depend

on each other, we would have to solve a system of equations to obtain the solutions.

$$l(\Sigma) = -\frac{n}{2}\ln\left[(2\pi)^d|\Sigma|\right] - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top\Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) \tag{19}$$

$$= -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top\Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}) + c \tag{20}$$

$$= -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left\{(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top\Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})\right\} + c \tag{21}$$

$(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top\Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$ is a scalar, so it is identical to its trace.

$$= -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{Tr}\left\{\Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top\right\} + c \tag{22}$$

$\Sigma^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top$ is a $d \times d$ matrix, but the trace is the same as the scalar above due to the following property of traces: $\mathrm{Tr}(ABC) = \mathrm{Tr}(BCA) = \mathrm{Tr}(CAB)$.

$$= -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{Tr}\left\{\Sigma^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top\right\} + c \tag{23}$$

$$= -\frac{n}{2}\ln|\Sigma| - \frac{1}{2}\mathrm{Tr}\left\{\Sigma^{-1}A\right\} + c \tag{24}$$

The derivative is

$$\frac{\partial l(\Sigma)}{\partial\Sigma} = -\frac{n}{2}\Sigma^{-\top} - \frac{1}{2}\frac{\partial}{\partial\Sigma}\mathrm{Tr}\left\{\Sigma^{-1}A\right\} + c \tag{25}$$

$$= -\frac{n}{2}\Sigma^{-\top} + \frac{1}{2}\left(\Sigma^{-1}A\Sigma^{-1}\right)^\top \tag{26}$$

$$= -\frac{n}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}A\Sigma^{-1} \tag{27}$$

Setting to zero and rearranging leads to

$$\frac{n}{2}\hat{\Sigma}^{-1} = \frac{1}{2}\hat{\Sigma}^{-1}A\hat{\Sigma}^{-1} \tag{28}$$

$$n\hat{\Sigma}^{-1} = \hat{\Sigma}^{-1}A\hat{\Sigma}^{-1} \tag{29}$$

$$n = A\hat{\Sigma}^{-1} \tag{30}$$

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \tag{31}$$

This is the sample covariance matrix, which is also a reasonable result.

We now have all ingredients for devising a Bayes-optimal classifier under the assumption of Gaussian class-conditional distributions.

4 Bias of the ML (co)variance estimate

Are these good estimators? What is a good estimator?

Quality criterion: mean squared deviation from the true parameter.

$$\mathrm{MSE}[\hat{\theta}] = \mathbb{E}\left[(\hat{\theta} - \theta)^2\right] \tag{32}$$

$$= \mathbb{E}\left[\left(\hat{\theta} - \mathbb{E}[\hat{\theta}]\right)^2\right] + \left(\mathbb{E}[\hat{\theta}] - \theta\right)^2 \tag{33}$$

$$= \mathrm{Var}(\hat{\theta}) + \mathrm{Bias}(\hat{\theta}, \theta)^2 \ , \tag{34}$$

where the expectation is taken over repeated independent experiments.

It can be shown that the ML estimate for the mean is unbiased: $\mathbb{E}[\hat{\mu}] = \mu$, and that $\mathrm{Var}(\hat{\mu}) = {}^{\sigma^2}/_{n}$.

However, the ML (co)variance estimate turns out to be biased:

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-1}{n}\sigma^2 \ . \tag{35}$$

(similar for $\hat{\Sigma}$)

The corrected estimator

$$\hat{\Sigma} = \frac{n}{n-1}\hat{\Sigma}_{\mathrm{ML}} = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top \tag{36}$$

is unbiased.

Note: both estimators are *consistent* (converge to the true value for $n \to \infty$, and variance goes to zero for $n \to \infty$). In most practical cases, this is sufficient.

5 Linear regression

So far, we have mainly talked about the classification rules. Another important problem in supervised machine learning is regression, i.e., function approximation.

Multiple nonlinear regression equation:

$$y_i = f(\mathbf{x}_i) + \epsilon_i \ . \tag{37}$$

Multiple linear regression equation:

$$y_i = \mathbf{x}_i^\top \boldsymbol{\beta} + \epsilon_i \ . \tag{38}$$

Here, $\mathbf{x}$ are the called independent variables (regressors, predictors), and y is the dependent variable (response, target). The regression problem is to estimate the optimal linear function (defined by the parameters $\boldsymbol{\beta}$) from data $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $y_i \in \mathbb{R}$.

Can we estimate $\boldsymbol{\beta}$ using maximum likelihood?

We can if we recognize that the noise terms ϵ_i are realizations of a random variable. For similar reasons as in the classification case, it is reasonable to assume i.i.d. Gaussian distributed noise:

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{39}$$

$$p(\epsilon|\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top\boldsymbol{\beta}\right)^2}{2\sigma^2}\right) \tag{40}$$

The data likelihood is as a function of $\boldsymbol{\beta}$ and σ^2 is

$$p(\mathcal{D}|\boldsymbol{\beta}, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{2\sigma^2}\right) . \tag{41}$$

$$\begin{aligned}
p(\mathcal{D}|\boldsymbol{\beta}, \sigma^2) &= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_{i=1}^{n} \exp\left(-\frac{\left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2}{2\sigma^2}\right) && (42)\\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2\right) && (43)\\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} ||\mathbf{y} - X\boldsymbol{\beta}||^2\right) && (44)\\
&= \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \exp\left(-\frac{1}{2\sigma^2} \left(\mathbf{y} - X\boldsymbol{\beta}\right)^\top \left(\mathbf{y} - X\boldsymbol{\beta}\right)\right) && (45)
\end{aligned}$$

Where we summarized data and noise into the vectors $\mathbf{y} = (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \in \mathbb{R}^n$, and the matrix $X = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^\top \in \mathbb{R}^{n \times d}$.

The log-likelihood is

$$l(\boldsymbol{\beta}, \sigma^2) = -n \ln(\sigma) - \frac{1}{2\sigma^2} \left(\mathbf{y} - X\boldsymbol{\beta}\right)^\top \left(\mathbf{y} - X\boldsymbol{\beta}\right) + c . \tag{46}$$

The derivative w.r.t. $\boldsymbol{\beta}$ is

$$\begin{aligned}
\frac{\partial l(\boldsymbol{\beta}, \sigma^2)}{\boldsymbol{\beta}} &= \frac{1}{\sigma^2} X^\top \left(\mathbf{y} - X\boldsymbol{\beta}\right) && (47)\\
&= \frac{1}{\sigma^2} X^\top \mathbf{y} - \frac{1}{\sigma^2} X^\top X\boldsymbol{\beta} . && (48)
\end{aligned}$$

Setting to zero yields

$$\begin{aligned}
\frac{1}{\sigma^2} X^\top X\hat{\boldsymbol{\beta}} &= \frac{1}{\sigma^2} X^\top \mathbf{y} && (49)\\
X^\top X\hat{\boldsymbol{\beta}} &= X^\top \mathbf{y} && (50)\\
\hat{\boldsymbol{\beta}} &= (X^\top X)^{-1} X^\top \mathbf{y} . && (51)
\end{aligned}$$

Note that this is the same estimate as we would get by minimizing the squared errors of the model ('ordinary least-squares' approach, OLS).

The derivative w.r.t. σ is

$$\frac{\partial l(\sigma^2)}{\sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} ||\mathbf{y} - X\hat{\boldsymbol{\beta}}||^2 . \tag{52}$$

Setting to zero yields

$$\begin{aligned}
\frac{n}{\hat{\sigma}} &= \frac{1}{\hat{\sigma}^3} ||\mathbf{y} - X\hat{\boldsymbol{\beta}}||^2 && (53)\\
\hat{\sigma}^2 &= \frac{1}{n} ||\mathbf{y} - X\hat{\boldsymbol{\beta}}||^2 && (54)\\
\hat{\sigma}^2 &= \frac{1}{n} ||\mathbf{y} - X(X^\top X)^{-1} X^\top \mathbf{y}||^2 . && (55)
\end{aligned}$$

6 Relationship between multiple linear regression and linear discriminant analysis

In the first lecture, we derived the discriminant function for the case of two Gaussian distributed classes with equal covariance matrix Σ and different class means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

The linear discriminant had the form

$$g(\mathbf{x}) = \mathbf{x}^\top \mathbf{w} > c \quad \text{with} \tag{56}$$

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \tag{57}$$

Assume now that in our regression setting, the response variables y_i are not real valued but binary. For convenience choose

$$y_i = \begin{cases} +n/n_1 & \text{if } \mathbf{x}_i \in \omega_1 \\ -n/n_2 & \text{if } \mathbf{x}_i \in \omega_2 \ , \end{cases} \tag{58}$$

where n_1 and n_2 are the numbers of samples in each class and $n = n_1 + n_2$.

Thus we have

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (X^\top X)^{-1} X^\top \mathbf{y} && (59)\\ &= n(X^\top X)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) && (60)\\ &= (\frac{1}{n} X^\top X)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) && (61)\\ &= (\Sigma_{\text{tot}})^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) && (62)\\ & && (63) \end{aligned}$$

Σ_{tot} and Σ are in general not the same.

However, it can be shown that

$$\Sigma_{\text{tot}} = \Sigma + c(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top \ . \tag{64}$$

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (\Sigma + c(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top)^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) && (65)\\ (\Sigma + c(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top)\hat{\boldsymbol{\beta}} &= \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 && (66)\\ \Sigma\hat{\boldsymbol{\beta}} + c(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top\hat{\boldsymbol{\beta}} &= \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 && (67)\\ \Sigma\hat{\boldsymbol{\beta}} + c_2(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) &= \boldsymbol{\mu}_1 - \boldsymbol{\mu}_2 && (68)\\ \hat{\boldsymbol{\beta}} &= (1 - c_2)\Sigma^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) && (69)\\ &\sim \mathbf{w} && (70)\\ & && (71) \end{aligned}$$

Note that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^\top$ is just a rank one matrix. Any projection onto it will therefore be a multiple of $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$. We can merge this with the term on the r.h.s. . Thus, LDA and linear/OLS regression use the same projection.

7 Summary

1. This lecture introduced the maximum likelihood principle of estimating parameters of distributions.
2. We have used ML to fit Gaussian distributions to observed data.
3. Applied to class-conditional distributions, this constitutes the last step towards to deriving Bayes-optimal LDA and QDA discriminant functions under the Gaussian model.
4. We have also used ML to fit the parameters and noise level of the multiple linear regression model under the assumption of Gaussian distributed noise.
5. Therefore, we now have a way to solve two fundamental problems in machine learning, classification and regression.
6. Here, we were able to derive analytic expressions for all parameters. Depending on the noise/data distribution and structural form of the regression function, this may not be possible in general.