Statistical Data Analysis

Jana de Wiljes

wiljes@uni-potsdam.de

www.dewiljes-lab.com

24. Oktober 2022

Universität Potsdam

Learning

Problem setting

Goal: Approximate function f, that describes the link between two random variables X and Y which have the joint distribution $\pi(z) = \pi(x, y)$

Choice of parametrisation:

- choose model class $\mathcal{H}$
- and appropriate loss functional $l(y, h(x))$

Expected Risk

For $h \in \mathcal{H}$ we define the expected Risik as follows

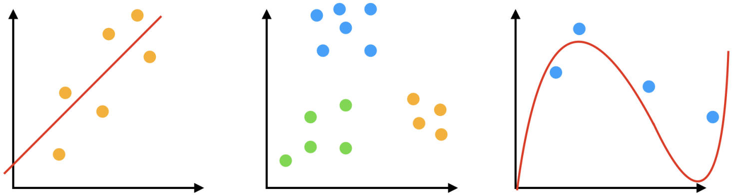
$$R(h) = \int_{\mathbf{Z}} l(y, h(x))\pi(z)dz \tag{1}$$

Approach: Want to find $h \in \mathcal{H}$ so that

$$h^* = \arg\min_{h \in \mathcal{H}} R(h) \tag{2}$$

Empirical Risk

Given in practice: independent and identical distributed Samples $S = \{(x_i, y_i)\}_{i=1}^{N}$ with $(x_i, y_i) \sim \pi(x, y)$ for $i \in \{1, \ldots, N\}$

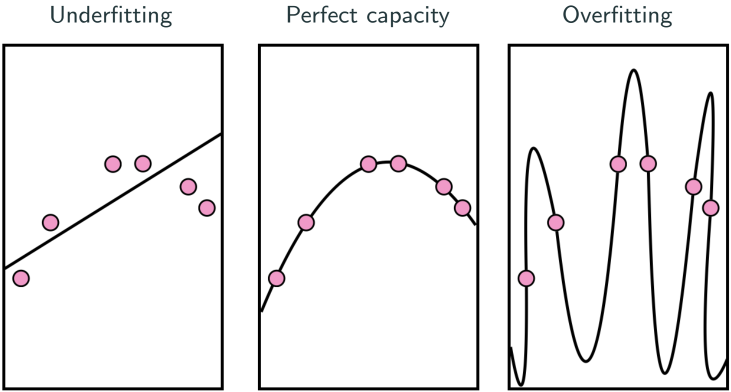


Empirical Risk

For a given sample set S we define the corresponding empirical risk as follows:

$$R_S(h) = \frac{1}{N} \sum_{i=1}^{N} l(y_i, h(x_i))$$

Empirical Risk-Minimizer



Choose: model class $\mathcal{H} = \{h(\cdot, \Theta) | \Theta \in \Omega\}$
Learning algorithm: Want to find $\Theta \in \Omega$ so that

$$\Theta^* = \arg\min_{\Theta \in \Omega} R_N(h_N(\cdot, \Theta))$$

Parameter estimation

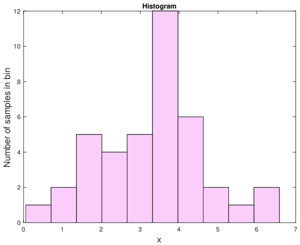
Maximum-Likelihood estimator

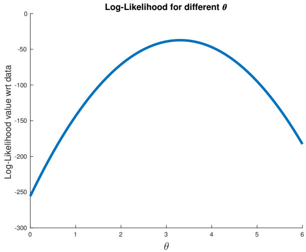
- **Choose:** a likelihood $L(x_1, \ldots, x_N; \theta)$ that describes the loss with respect to a specific parameter and the data x_i (unsupervised)
- **Goal:** find unknown θ that maximizes the fixed likelihood

Maximum-Likelihood estimator

For a fixed likelihood the so called maximum likelihood estimator is defined by

$$\widehat{\theta}^N_{\text{MLE}} = \arg\max_{\theta \in \Theta} L(x_1, \ldots, x_N; \theta) \tag{3}$$





Example

- **Assume:** data $x_1, \ldots, x_i, \ldots, x_N$ is independently normally distributed ,i.e., $x_i \sim \mathcal{N}(\mu, \sigma)$
- **Aim:** trying to determine the unknown parameter θ that corresponds to the mean μ of the normal distribution (here σ is assumed to be known) by maximizing

$$L(x_1, \ldots, x_N; \theta) := \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x_i - \theta)^2}{2\sigma^2} \right) \tag{4}$$

- **Method:** apply the log function

$$\begin{aligned} \log L(x_1, \ldots, x_N; \theta) &= \log \left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(- \frac{(x_i - \theta)^2}{2\sigma^2} \right) \right) \\ &= - \sum_{i=1}^{N} \frac{(x_i - \theta)^2}{2\sigma^2} + \frac{1}{\sqrt{2\pi}\sigma} \end{aligned}$$

Bayes estimator

Def: The a-posteriori-distribution of θ is the conditional distribution given the information $X_1 = x_1, \ldots, X_n = x_n$, i.e.,

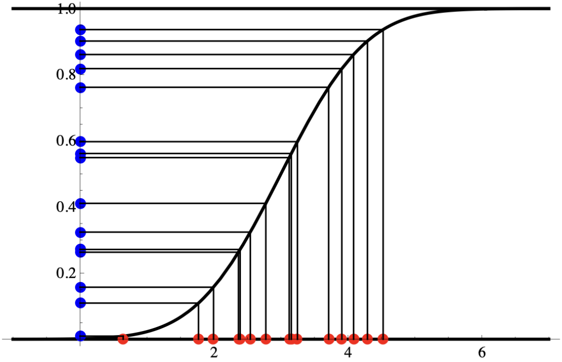
$$q(\theta_i | x_1, \ldots, x_n) := \mathbb{P}[\theta = \theta_i | X_1 = x_1, \ldots, X_n = x_n], \quad i = 1, 2, \ldots$$

Def: The Bayes estimator is defined as the expectation of the a-posteriori-distribution

$$\widehat{\theta}_{\text{Bayes}} = \sum_i \theta_i q(\theta_i | x_1, \ldots, x_n)$$

Maximum-spacing method

Lemma: Let the cumulative distribution function F_θ be continuous and strictly monoton increasing. Under $\mathbb{P}_\theta$ the random variables $F_\theta(X_1), \ldots, F_\theta(X_n)$ are independent and uniformly distributed on the $(0, 1)$ interval.



Maximum-spacing method

Lemma: The maximum-spacing method is defined via

$$\widehat{\theta}_{MS} = \arg\max_{\theta \in \Theta} \prod_{i=1}^{n+1} (F_\theta(x_{(i)}) - F_\theta(x_{(i-1)})) \tag{5}$$

Linear regression

Model for simple linear regression

Model:

$$Y_i = f(X_i, \beta) + \epsilon_i, \quad i = 1, \ldots, n \tag{6}$$

where ϵ_i are iid with $\mathbb{E}[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$

Data: it is possible to observe realisations

$$(y_i, x_i) \quad i = 1, \ldots, n \tag{7}$$

Goal: estimate parameters β of the function to obtain approximative $f(x, \hat{\beta})$

Note: note that f approximates $\mathbb{E}[Y_i|X_i]$

Model for simple linear regression

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1, \dots, n \tag{8}$$

where ϵ_i are iid with $\mathbb{E}[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$

Data:

$$(y_i, x_i) \quad i = 1, \dots, n \tag{9}$$

Goal: estimate $f(x, \hat{\beta}) = \hat{\beta}_0 + \hat{\beta}_1 x$

The Ordinary Multiple Linear Regression Model

Model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \cdots + \beta_p X_{i,p} + \epsilon_i, \quad i = 1, \ldots, n \tag{10}$$

where ϵ_i are iid with $\mathbb{E}[\epsilon_i] = 0$ and $Var(\epsilon_i) = \sigma^2$

Data:

$$(y_i, x_i) \quad i = 1, \ldots, n \tag{11}$$

Goal: estimate $\hat{f}(x_1, \ldots, x_p, \hat{\beta}_1, \ldots, \hat{\beta}_p) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots, \hat{\beta}_p x_p$

Multivariate Random Variables

Def: Let $\mathbf{X}$ be a vector of (univariate) random variables, i.e., $\mathbf{X} = (X_1, \ldots, X_p)^\top$ with $\mathbb{E}[X_i] = \mu_i$. $\mathbf{X}$ is called a multivariate random variable and we denote $\mathbb{E}[\mathbf{X}] = \mu$

Note:

- Variance $\text{Var}(X_i) = \mathbb{E}[(X_i - \mathbb{E}(X_i))^2] = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_i - \mathbb{E}(X_i))]$
- Covariance $\text{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mathbb{E}(X_i))(X_j - \mathbb{E}(X_j))]$

Least squares Estimation: minimize the sum of squared errors

Least squares estimation: minimize the sum of squared errors

$$L(\beta) = \sum_{i=1}^{N}(y_i - x_i^\top \beta_i)^2 = \sum_{i=1}^{N} \epsilon_i^2 = \epsilon^\top \epsilon \tag{12}$$

with respect to $\beta \in \mathbb{R}^{p+1}$

$$\hat{\beta} = \arg\min_{\beta} \sum_{i=1}^{N} \epsilon_i^2 = \epsilon^\top \epsilon = (Y - X\beta)^\top (Y - X\beta) \tag{13}$$

$$Y = \underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix}}_{\in \mathbb{R}^{N \times 1}}, \quad X = \underbrace{\begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ \vdots & & & \\ 1 & x_{N1} & \dots & x_{Np} \end{bmatrix}}_{\in \mathbb{R}^{N \times p+1}}, \beta = \underbrace{\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_N \end{bmatrix}}_{\in \mathbb{R}^{p+1 \times 1}}, \epsilon = \underbrace{\begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix}}_{\in \mathbb{R}^{N \times 1}},$$

Covariance

Def: The covariance of the multivariate random variable $\mathbf{X}$ is defined by

$$\Sigma := \mathrm{Cov}(\mathbf{X}) = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^\top] \tag{14}$$

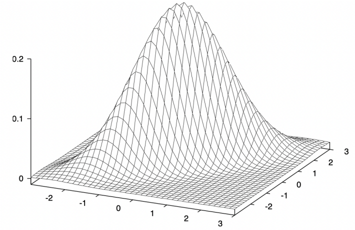
Example:

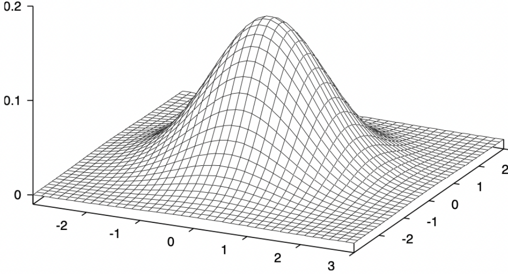
$$\Sigma = \begin{pmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) \end{pmatrix} \tag{15}$$

Properties of Σ:

- quadractic
- symmetric
- positive-semidefinite

Multivariate Normal Distribution





$$\mathbf{X} \sim \mathcal{N}_p(\mu, \Sigma) \tag{16}$$

Positive semi-definite

Lemma: Let $\mathbf{B}$ be an $n \times (p+1)$ matrix. Then the matrix $\mathbf{B}^\top \mathbf{B}$ is symmetric and postive semi-definite. It is positive definite, if $\mathbf{B}$ has full column rank. Then, besides $\mathbf{B}^\top \mathbf{B}$ also $\mathbf{B}\mathbf{B}^\top$ is postive semi-definite.

LS-estimator

Theorem: The LS-estimator of the unknown parameters β is

$$\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \tag{17}$$

if $\mathbf{X}$ has full column rank p+1.

Proof:

Proof

Positive semi-definite

Proposition: The hat-matrix $\mathbf{H} = (h_{ij})_{1 \leq i,j, \leq b}$ has the followi-ng properties:

1. $\mathbf{H}$ is symmetric
2. $\mathbf{H}$ is idempotent, i.e., $\mathbf{HH} = \mathbf{H}$
3. $rk(\mathbf{H}) = tr(\mathbf{H}) = p + 1$
4. $0 \leq h_{ii} \leq 1, \quad \forall i = 1, \ldots, n$
5. the matrix $\mathbf{I}_n - \mathbf{H}$ is also symmetric and idempotent with $rk(\mathbf{I}_n - \mathbf{H}) = n - p - 1$

ML-estimator

Theorem: The ML-estimator of the unknown parameters σ^2 is $\hat{\sigma}^2_{ML} = \frac{\hat{\epsilon}\hat{\epsilon}}{n}$ with $\hat{\epsilon} = \mathbf{y} - \mathbf{X}\hat{\beta}$.

Proof

ML-estimator

Proposition: For the ML-estimator $\hat{\sigma}^2_{ML}$ of σ^2 the following property holds:

$$\mathbb{E}[\sigma^2_{ML}] = \frac{n - p - 1}{n}\sigma^2 \tag{18}$$

Proof

Adjusted estimator

Proposition: The adjusted estimator

$$\hat{\sigma}^2_{ad} = \frac{\hat{\epsilon}\hat{\epsilon}}{n-p-1} \tag{19}$$

of the unknown parameter σ^2 can be written as

$$\hat{\sigma}^2_{ad} = \frac{\mathbf{y}^\top \mathbf{y} - \hat{\beta}^\top \mathbf{X}^\top \mathbf{y}}{n-p-1} \tag{20}$$

Proof

ML estimator

Proposition: The LS-estimator $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$ is equivalent to the ML-estimator based on maximization of the log-likelihood

$$l(\beta, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)^\top(\mathbf{y} - \mathbf{X}\beta) \tag{21}$$

Proof

LS estimator

Proposition: The LS-estimator $\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$ and the REML-estimator $\hat{\sigma}^2 = \frac{1}{n-p-1}\hat{\epsilon}^\top\hat{\epsilon}$ the following properties hold:

1. $\mathbb{E}[\hat{\beta}] = \beta$, $\mathrm{Cov}(\hat{\beta}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$
2. $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$

Sequential Learning Algorithm

New challenge: Need to learn Θ sequentially or online, i.e.,

$$\Theta^j = \arg\min_{\Theta \in \Omega} R_j(h_j(\cdot, \Theta^j | \Theta^{j-1}))$$

But why?

- sequential decision involved, i.e., $h(\cdot, \Theta, a_t)$
- data can only be collected individually in time and we already want to start predicting
- nonstationary Θ_t

Sequential update of Linear Regression Parameter

Choose: $h(x, \Theta) = x\Theta$ and we assume $y = h(x, \Theta) + \epsilon$ where $\epsilon \sim \mathcal{N}(0, 1)$

Linear regression with batch data:

$$R_N(h) = \frac{1}{N}\sum_{i=1}^{N}(y_i - x_i\Theta)^2 \quad \text{where } \Theta_N^* = (\sum_{i=1}^{N} x_i^2)^{-1}(\sum_{i=1}^{N} x_i y_i)$$

Sherman-Morrison formula

$$\left(A + uv^{\mathsf{T}}\right)^{-1} = A^{-1} - \frac{A^{-1}uv^{\mathsf{T}}A^{-1}}{1 + v^{\mathsf{T}}A^{-1}u}. \tag{22}$$

New data: (x_{N+1}, y_{N+1}) yield a new parameter

$$\Theta^* = (\sum_{i=1}^{N} x_i^2 + x_{N+1}x_{N+1})^{-1}(\sum_{i=1}^{N} x_i y_i + x_{N+1}y_{N+1}) \tag{23}$$

Using the Sherman-Morrison formula we can update recursively.

New challenge

Setting:

- h is known and links noisy and partial observations y to x
- x is not given directly as a sample but we have a model f that we can use as a surrogate to generate samples

Goal: estimating the associate density conditioned on the data and using the prior information given via f