ASSIGNMENT 1: ESTIMATION THEORY, FISHER INFORMATION, CRLB

Institute for Machine Learning

JOHANNES KEPLER
UNIVERSITY LINZ

Institute for
Machine Learning

Contact

Heads:
Thomas Adler,
Philipp Renz,
Andreas Radler

Institute for Machine Learning
Johannes Kepler University
Altenberger Str. 69
A-4040 Linz

E-Mail: `{adler,renz,radler}@ml.jku.at`
Institute Homepage

Copyright statement:

Agenda

- Motivation and Notation
- Estimation Theory, Unbiased Estimators (Short Recap)
- Fisher Information Matrix
- Cramér-Rao lower bound
- Further Literature:
  - Lecture notes
  - Mathematics for Machine Learning (Deisenroth at. el., 2018)

What is Machine Learning?

Machine Learning:

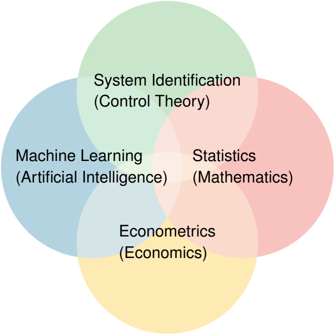
$$\text{data} + \text{model} \xrightarrow{\text{compute}} \text{prediction} \qquad (1)$$

Infos: Neil Lawrence:

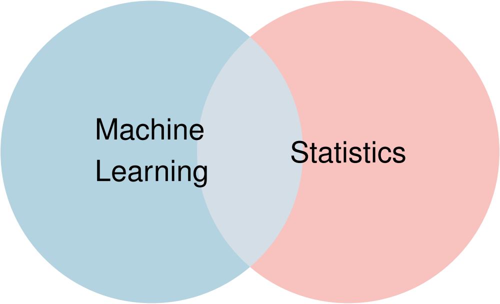
http://inverseprobability.com

What is machine learning?

Machine Learning and Related Fields:



Machine Learning vs Statistics



- Minimization of Generalization Error
- ML tries to make model predictions
- Statistical Learning Theory (Vapnik) is built on bias-variance tradeoff of model prediction

- Parameter estimation and variance analysis
- Statistics tries to estimate parameters as good as possible
- Statistics is built on bias-variance of parameter estimation

Notation

- Variable X, Y in uppercase letters are random variables
- We denote column vectors as $\boldsymbol{x} = (x_1, \ldots, x_m)^\top \in \mathbb{R}^m$
- We denote matrices as $\boldsymbol{X} \in \mathbb{R}^{n \times m}$ consisting of an n-tuple of vectors $\boldsymbol{x}_i$ such that $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$
  - If $\boldsymbol{X}$ is a data matrix, then usually n denotes the number of samples and m denotes the number of features
- An estimator $\hat{w}$ of a parameter w is indicated with a hat

Motivation

- We observe some data $\boldsymbol{X}$ and want to know which model is likely to having created $\boldsymbol{X}$.
- We assume that a *model class* defined by the *distribution* $p(x; \boldsymbol{w})$ parameterized by $\boldsymbol{w}$ created the data $\boldsymbol{X}$.
- However, we do not know the correct value of the parameters $\boldsymbol{w}$, so we have to use the observed data $\boldsymbol{X}$ to estimate it.
- It would be nice to know how likely it is for our estimated $\hat{\boldsymbol{w}}$ to actually produce the data $\boldsymbol{X}$ and how important the choice of $\hat{\boldsymbol{w}}$ w.r.t. the data $\boldsymbol{X}$ is.

Recap: Definition Expectation

- Expectation of a random variable X:
  (discrete distribution, continuous distribution)

$$\mathrm{E}(X) = \sum_j x_j \, p(x_j) \qquad \mathrm{E}(X) = \int_{-\infty}^{\infty} x \, p(x) \, \mathrm{d}x \tag{2}$$

- Expectation of a function $g(X)$ with a random variable X:
  (discrete distribution, continuous distribution)

$$\mathrm{E}(g(X)) = \sum_j g(x_j) \, p(x_j) \qquad \mathrm{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \, p(x) \, \mathrm{d}x \tag{3}$$

- Variance of a random variable X:

$$\mathrm{Var}(X) = \mathrm{E}([X - \mathrm{E}(X)]^2) = \mathrm{E}(X^2) - \mathrm{E}(X)^2 \tag{4}$$

Recap: Expectation Calculation Rules

- The expectation E of a constant c is the constant:

$$\mathrm{E}(c) = c \tag{5}$$

- Adding a constant value c to each term increases the expected value by the constant:

$$\mathrm{E}(X + c) = \mathrm{E}(X) + c \tag{6}$$

- Multiplying each term by a constant value c multiplies the expected value by that constant:

$$\mathrm{E}(c\,X) = c\,\mathrm{E}(X) \tag{7}$$

- The expected value of the sum of two random variables is the sum of the expected values (additive law of expectation):

$$\mathrm{E}(X + Y) = \mathrm{E}(X) + \mathrm{E}(Y) \tag{8}$$

Recap: Notation

- Be careful with notation
- Conditional Probability:

$$\begin{aligned} &p(x;\omega) && x \text{ is a random variable, } \omega \text{ is a parameter} \\ &p(x \mid y) && x \text{ is a random variable, } y \text{ is a random variable} \end{aligned} \tag{9}$$

- Expectation (lots of different notations):

$$\begin{aligned} &\mathrm{E}, \quad \mathrm{E}_X, \quad \mathrm{E}_{p(x,\omega)}, \quad \mathrm{E}_{p(x)_\omega}, \quad \mathrm{E}_{X\sim p(x;\omega)}, \quad \mathrm{E}_\omega, \ \ldots \\ &\mathrm{E}_{\boldsymbol{X}}, \quad \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}, \quad \mathrm{E}_{\boldsymbol{X}\sim p(\boldsymbol{x};\boldsymbol{w})}, \quad E_{(\boldsymbol{x}_1,\boldsymbol{x}_2)}, \ \ldots \end{aligned} \tag{10}$$

- Notation you should use:

$$\begin{aligned} &\mathrm{E}, \quad \mathrm{E}_{p(x;\omega)} \\ &\mathrm{E}_{\boldsymbol{X}}, \quad \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})} \end{aligned} \tag{11}$$

Recap: Bias and Variance, Scalar Parameter

- Estimator:

$$\hat{w} = \hat{w}(\boldsymbol{X}) \tag{12}$$

- Evaluation criterion, mean squared error (MSE):

$$\mathrm{mse}(\hat{w}, w) = \mathrm{E}_{\boldsymbol{X}}[(\hat{w} - w)^2] = \mathrm{Var}(\hat{w}) + \mathrm{Bias}^2(\hat{w}, w) \tag{13}$$

- **Variance**:

$$\mathrm{Var}(\hat{w}) = \mathrm{E}_{\boldsymbol{X}}(\hat{w}^2) - \mathrm{E}_{\boldsymbol{X}}(\hat{w})^2 \tag{14}$$

- **Bias**:

$$\mathrm{Bias}(\hat{w}, w) \ = \ \mathrm{E}_{\boldsymbol{X}}(\hat{w}) \ - \ w \tag{15}$$

- An estimator is **unbiased** if

$$\mathrm{E}_{\boldsymbol{X}}(\hat{w}) \ = \ w \tag{16}$$

  i.e. on average over the training set $\boldsymbol{X}$ the estimator will yield the true parameter.

Recap: Bias and Variance, Parameter Vector

- **Bias**:

$$\boldsymbol{b}(\hat{\boldsymbol{w}}, \boldsymbol{w}) = \mathrm{E}_{\boldsymbol{X}}(\hat{\boldsymbol{w}}) - \boldsymbol{w} \tag{17}$$

$$b_*^2(\hat{\boldsymbol{w}}, \boldsymbol{w}) = \mathrm{E}_{\boldsymbol{X}}(\hat{\boldsymbol{w}} - \boldsymbol{w})^\top \mathrm{E}_{\boldsymbol{X}}(\hat{\boldsymbol{w}} - \boldsymbol{w}) \tag{18}$$

- An estimator is **unbiased** if

$$\mathrm{E}_{\boldsymbol{X}}(\hat{\boldsymbol{w}}) \; = \; \boldsymbol{w} \tag{19}$$

- **Variance** (*these are sums, not individual values):

$$\mathrm{var}_*(\hat{\boldsymbol{w}}) = \mathrm{E}_{\boldsymbol{X}}[(\hat{\boldsymbol{w}} - \mathrm{E}_{\boldsymbol{X}}(\hat{\boldsymbol{w}}))^\top (\hat{\boldsymbol{w}} - \mathrm{E}_{\boldsymbol{X}}(\hat{\boldsymbol{w}}))] \tag{20}$$

- MSE:

$$\begin{aligned} \mathrm{mse}(\hat{\boldsymbol{w}}, \boldsymbol{w}) &= E_{\boldsymbol{X}}[(\hat{\boldsymbol{w}} - \boldsymbol{w})^\top (\hat{\boldsymbol{w}} - \boldsymbol{w})] && (21) \\ &= \mathrm{var}_*(\hat{\boldsymbol{w}}) + b_*^2(\hat{\boldsymbol{w}}, \boldsymbol{w}) && (22) \\ &= \sum_{i=1}^n \mathrm{Var}(\hat{w}_i) + \sum_{i=1}^n \mathrm{Bias}^2(\hat{w}_i, w_i) && (23) \end{aligned}$$

Cramér-Rao Lower Bound and Efficiency

- How can we see if an estimator uses the data to estimate a parameter *efficiently*?
- How can we see how *efficient* we could get?
- Is there an optimal bound? If yes, how large is it?

Fisher Information

- Assumption: A model/distribution $p(x; \boldsymbol{w})$, parameterized by $\boldsymbol{w}$, created some data $\boldsymbol{X}$
- We can observe the data $\boldsymbol{X}$ but the parameter-vector $\boldsymbol{w}$ is unknown
- **Fisher Information**: Shows how much “information” the created observable data $\boldsymbol{X}$ holds about the unknown model parameters $\boldsymbol{w}$ that were used to produce them

Fisher Information: Likelihood

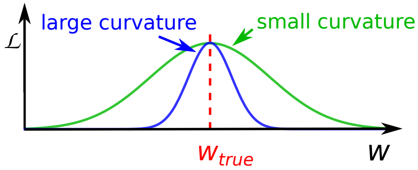
- The likelihood $\mathcal{L}(w)$ is a measure of how likely a parameter w, which parameterizes a model $p(x; w)$, is to produce some observed i.i.d. data set $\{x_1, \ldots, x_n\}$.
- **Likelihood:**

$$\mathcal{L}(w) = \prod_{i=1}^{n} p(x_i; w) \tag{24}$$

- **Log-Likelihood:**

$$\ln \mathcal{L}(w) = \sum_{i=1}^{n} \ln p(x_i; w) \tag{25}$$

Fisher Information: Idea



- Large curvature leads to small variance of the estimator $\rightarrow$ only a few w would produce $\boldsymbol{X}$ well and those w will be close to w_{true}
- The smaller the variance the more efficient the parameter estimation.
- Fisher Information Matrix $\boldsymbol{I}_F(w)$ is the variance of the derivative of the log-likelihood, which (under regularity conditions) is the negative curvature.

Fisher Information: Formula (1)

- The Fisher Information for a scalar parameter w is defined as

$$I_F(w) \;:=\; \mathrm{E}_{p(x;w)}\left[\left(\frac{\partial}{\partial w}\ln\mathcal{L}(w)\right)^2\right]. \tag{26}$$

- Under the regularity condition that differentiation and integration can be exchanged, we have

$$\forall_w : \quad \mathrm{E}_{p(x;w)}\left(\frac{\partial \ln\mathcal{L}(w)}{\partial w}\right) \;=\; 0\,. \tag{27}$$

- Then, the Fisher Information is the variance of the derivative of the log-likelihood $\ln\mathcal{L}(w)$:

$$I_F(w) \;=\; \mathrm{Var}_{p(x;w)}\left(\frac{\partial}{\partial w}\ln\mathcal{L}(w)\right) \tag{28}$$

Fisher Information: Formula (2)

- With $\boldsymbol{w} = (w_1, \ldots, w_N)^\top$ a parameter vector, the Fisher information will become an $N \times N$ matrix:

$$\left[\boldsymbol{I}_F(\boldsymbol{w})\right]_{ij} \;=\; \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial \ln \mathcal{L}(\boldsymbol{w})}{\partial w_i}\;\frac{\partial \ln \mathcal{L}(\boldsymbol{w})}{\partial w_j}\right) \tag{29}$$

$$\left[\boldsymbol{I}_F(\boldsymbol{w})\right] \;=\; \mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left[\left(\frac{\partial \ln \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}}\right)^T\left(\frac{\partial \ln \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}}\right)\right] \tag{30}$$

- A regularity condition also for second derivatives implies that the Fisher information represents the (negative) curvature:

$$\left[\boldsymbol{I}_F(\boldsymbol{w})\right]_{ij} \;=\; -\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial^2 \ln \mathcal{L}(\boldsymbol{w})}{\partial w_i \partial w_j}\right) \tag{31}$$

$$\boldsymbol{I}_F(\boldsymbol{w}) \;=\; -\mathrm{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial^2 \ln \mathcal{L}(\boldsymbol{w})}{\partial \boldsymbol{w}^\top\,\partial \boldsymbol{w}}\right) \tag{32}$$

Cramér-Rao Lower Bound and Efficiency

- The **Cramér-Rao Lower Bound** (CRLB) is
  - a **lower bound** for the variance of an **unbiased estimator**
  - the inverse of the **Fisher information (matrix)**
- For scalar parameter:

$$\mathrm{Var}(\hat{\omega}) \geq \frac{1}{I_F(\omega)} \tag{33}$$

- For vector parameter:

$$\mathrm{Covar}(\hat{\boldsymbol{w}}) \geq \boldsymbol{I}_F^{-1}(\boldsymbol{w}) \tag{34}$$

- An unbiased estimator is said to be **efficient** if its variance reaches the **CRLB.** It is efficient in the sense that it efficiently makes use of the data and extracts maximal information to estimate the parameter.

Minimal Variance Unbiased Estimator

- In statistics, a minimum-variance unbiased estimator (MVUE) is an unbiased estimator that has lower variance than any other unbiased estimator for all possible values of the parameter.
- Hence, an efficient unbiased estimator (reaching the CRLB) is always the MVUE.
- However: An MVUE may or may not be efficient. An MVUE may be “optimal” but still not reach the CRLB.