ASSIGNMENT 2: FISHER INFORMATION, CRLB

Johannes Kofler,
Markus Holzleitner
Institute for Machine Learning

JOHANNES KEPLER
UNIVERSITY LINZ

Institute for
Machine Learning

Copyright statement:

Agenda

- Recap: Maximum Likelihood
- Recap: Bias and Variance
- Fisher Information
- Cramer Rao Lower Bound

Recap: Maximum Likelihood

- **Given:**
  - data samples $\{\mathbf{x}\} = \{\mathbf{x}^1, \ldots, \mathbf{x}^l\}$ sampled i.i.d. from random variable X
  - a parametrized model distribution $p(\mathbf{x}; \mathbf{w})$ with parameters $\mathbf{w}$
- **Task:** find the parameter $\mathbf{w}$ that was most likely to produce this data
- **Idea:** How likely was a given $\mathbf{w}$ to produce the dataset?

$$\mathcal{L}(\{\mathbf{x}\}; \mathbf{w}) = p(\{\mathbf{x}\}; \mathbf{w}) = \prod_{i=1}^{n} p(\mathbf{x}^i; \mathbf{w})$$

- **Solution:** Find the $\hat{\mathbf{w}}$ that maximizes $\mathcal{L}(\{\mathbf{x}\}; \mathbf{w})$

$$\hat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} \mathcal{L}(\{\mathbf{x}\}; \mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \prod_{i=1}^{n} p(\mathbf{x}^i; \mathbf{w})$$

- Often, it is better to optimize a sum instead of a product: use logarithm.

$$\hat{\mathbf{w}} = \operatorname*{argmax}_{\mathbf{w}} \ln \mathcal{L}(\{\mathbf{x}\}; \mathbf{w}) = \operatorname*{argmax}_{\mathbf{w}} \sum_{i=1}^{n} \ln p(\mathbf{x}^i; \mathbf{w})$$

Recap: Bias and Variance

Bias:

$$b(\hat{\boldsymbol{w}}) = \mathbb{E}_X[\hat{\boldsymbol{w}}] - \boldsymbol{w} \tag{1}$$

(Co)Variance:

$$(\mathrm{Co})\mathrm{Var}(\hat{\boldsymbol{w}}) = \mathbb{E}_X[(\hat{\boldsymbol{w}} - \mathbb{E}_X[\hat{\boldsymbol{w}}])(\hat{\boldsymbol{w}} - \mathbb{E}_X[\hat{\boldsymbol{w}}])^T] \tag{2}$$

A estimator is **unbiased** if

$$\mathbb{E}_X[\hat{\boldsymbol{w}}] = \boldsymbol{w} \tag{3}$$

i.e. on average the estimator will yield the true parameter.

Fisher Information

- Assume we just try out some random "estimates" $\hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_k$ and take a look at the likelihoods.
- Let's say the likelihoods barely change for all our "estimates". Intuitively this seems to make the estimation harder. We can't get much information out of our trials.
- In contrast, if the likelihoods change a lot, i.e. some of our "estimates" have a high likelihood and some don't, it seems to be a lot easier to find a $\hat{\boldsymbol{w}}$ that is close to $\boldsymbol{w}$. There is more information available.

Mathematically, we can capture the above with the Fisher Information Matrix:

$$\boldsymbol{I}_F(\boldsymbol{w}) : \quad [\boldsymbol{I}_F(\boldsymbol{w})]_{ij} = \mathbb{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial w_i} \frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial w_j}\right) \tag{4}$$

- The Fisher information gives the amount of information that an observable random variable X carries about a parameter $\boldsymbol{w}$.
- This is a property of the underlying distribution.

Fisher Information

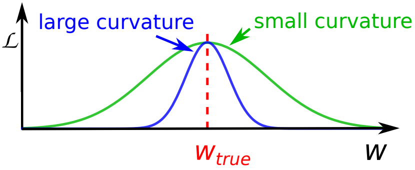
If the density function $p(\boldsymbol{x};\boldsymbol{w})$ satisfies

$$\mathbb{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}}\right) = \boldsymbol{0}\,, \tag{5}$$

then the Fisher information matrix is

$$\boldsymbol{I}_F(\boldsymbol{w}):\quad \boldsymbol{I}_F(\boldsymbol{w}) = -\mathbb{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial^2 \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}^2}\right)\,. \tag{6}$$

- In this case the interpretation is straight forward: The Fisher information is the negative expected value of the curvature of the log likelihood.
- The Fisher Information is large where the log likelihood has large (negative) curvature.



Cramer-Rao Lower Bound and Efficiency

Implications of the **Cramer-Rao Lower Bound** (CRLB):

- The **CRLB** is a lower bound for the variance of an **unbiased** estimator.
  - That means: there is no unbiased estimator with smaller variance.
  - Actually there even may not be an estimator that reaches the CRLB.
- An unbiased estimator is said to be **efficient** if it reaches the **CRLB**. It is efficient in that it efficiently makes use of the data and extracts information to estimate the parameter.

Minimal Variance Unbiased Estimator

- An efficient unbiased estimator is always the Minimal Variance Unbiased Estimator (MVUE).
- BUT: An MVUE may or may not be efficient.

Cramer-Rao Lower Bound and Efficiency

If the density function $p(\boldsymbol{x};\boldsymbol{w})$ satisfies

$$\forall_{\boldsymbol{w}} : \ \mathbb{E}_{p(\boldsymbol{x};\boldsymbol{w})}\left(\frac{\partial \ln p(\boldsymbol{x};\boldsymbol{w})}{\partial \boldsymbol{w}}\right) \ = \ \boldsymbol{0} \ ,$$

and the estimator $\boldsymbol{w}$ is unbiased, then the following holds:

$$\mathrm{Covar}(\hat{\boldsymbol{w}}) - \boldsymbol{I}_F(\boldsymbol{w})^{-1} \geq 0 \ , \tag{7}$$

where the inequality holds in the sense of positive semidefiniteness of matrices. That means the inverse of $\boldsymbol{I}_F(\boldsymbol{w})$ is a lower bound for the variance of an estimator. We call this lower bound the Cramer Rao Lower Bound, short CRLB.

Cramer-Rao Lower Bound and Efficiency

The bound is attained **iff** there exists the following decomposition of the first derivative of the log likelihood (also called the "score function"):

$$\frac{\partial \ln p(\boldsymbol{x}; \boldsymbol{w})}{\partial \boldsymbol{w}} = \boldsymbol{A}(\boldsymbol{w})(\boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{w}) \tag{8}$$

for some function $\boldsymbol{g}$ and an square matrix $\boldsymbol{A}(\boldsymbol{w})$. Then our estimator is the MVU estimator $\hat{\boldsymbol{w}} = \boldsymbol{g}(\boldsymbol{x})$ with $\mathrm{Covar}(\hat{\boldsymbol{w}}) = \boldsymbol{A}(\boldsymbol{w})^{-1} = \boldsymbol{I}_F(\boldsymbol{w})^{-1}$.

Cramer Rao Inequality / Estimator Efficiency

Scalar case:

Cramer Rao inequality:

$$\mathrm{Var}(\hat{w}) \geq \frac{1}{I_F(w)} \tag{9}$$

Efficiency of an estimator:

$$e(\hat{w}) = \frac{\frac{1}{I_F(w)}}{\mathrm{Var}(\hat{w})} \tag{10}$$