

# ORF525 - Class Notes

Bachir EL KHADIR

February 9, 2016

“I pledge my honor that this lecture summary is my own work and adheres to the guidelines in the instructions.”

## 1 Fundamental Principles of Data Analysis

### 1.1 Concentration Principle

$$\begin{array}{lcl} \text{Data} = \text{Signal} & & + \text{Noise} \\ (X_1, \dots, X_n) = \underbrace{\theta}_{\mathcal{N}(0,1)} & & + \text{uncertainty from the nature} \end{array}$$

**Important concept:** concentration phenomenon, eg. Law of large number

**Main Idea:** We need the data to have some stationary pattern to *summon* noise / uncertainty.

### 1.2 Parsimonions Principle

**Intuition:** If two explanations are equally good, we prefer the simpler one.  $\Rightarrow$  Regularization technique. **Key:** We always use the *wrong* model to control variance.

Basic concpets:

1. Sample space: All possible outs of a statistical experiment.
2. Random Sample (Data):  $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} p(x)$ ,  $p(x)$  being the density of  $X$ .
3. Realization (Observed value):  $x_1, \dots, x_n$  deterministic
4. Notation:  $\underbrace{X_1, \dots, X_n}_{X_{1:n}}, \underbrace{x_1, \dots, x_n}_{x_{1:n}}$
5. Statistic: Any measurable function of  $X_1, \dots, X_n$
6. CDF:  $F(x) := \mathbb{P}(X \leq x)$
7. PDF:  $p(x) := \frac{\partial}{\partial x} F(x)$ , could also be PMF for discrete variables.

8. We use  $p_\theta(x)$  to denote that the density is parametrized by  $\theta$
9. LLN (Estimation), CLT(Confidence Interval / p-value)
10. Statistical Model: A set of probability distributions indexed by a parameter set  $\Theta$ .

$$\mathbb{P} := \{p_\theta : \theta \in \Theta\}$$

11. Parametric Model: If there exists a finite-dimensional  $\Theta$  to index  $\mathcal{P}$ .
12. Nonparametric Model: If there doesn't exist a finite-dimensional  $\Theta$  to index  $\mathcal{P}$ . example: Sobolev space  $P := \{p(x) \text{ is continuous and } \int p'' < \infty\}$
13. Point estimation: Let  $X_1 \dots X_n \stackrel{\text{iid}}{\sim} p_\theta(x)$  we want to make a *single best guess* at  $\theta$ .  $\underbrace{X_1 \dots X_n}_{\hat{\theta}_n := g(X_1, \dots, X_n)} \sim \underbrace{p_\theta(x)}_{\theta}$  We hope that  $\hat{\theta}_n \xrightarrow{P} \theta$  as  $n \rightarrow \infty$
14. Consistent Estimation:  $\hat{\theta}_n \xrightarrow{P} \theta$  as  $n \rightarrow \infty$ . Unbiased Estimator: Define  $\text{Bias}(\hat{\theta}_n) = \mathbb{E}\hat{\theta}_n - \theta$  If  $\text{Bias}(\hat{\theta}_n) = 0 \Rightarrow \hat{\theta}_n$  is called unbiased. Question: Consistency  $\iff$  Unbiasedness. Answer: no. Example:  $X_1, \dots, X_n \sim N(\mu, 1)$

- $\hat{\theta}_n = X_1 \Rightarrow$  unbiased, not consistent
- $\hat{\theta}_n = \frac{1}{n+1} \sum X_i \Rightarrow$  biased, consistent

15. The Likelihood function of  $\theta$  related to a random sample  $X_i$  is  $\underbrace{L(X_i, \theta)}_{\text{Random quantity} := p_\theta(X_i)}$

16. Joint likelihood, The joint likelihood of  $\theta$  wrt the entire data set is defined as  $L_n(\theta) := p_\theta(X_1, \dots, X_n)$
17. Joint log-likelihood  $l_n(\theta) := \log[L_n(\theta)]$
18. Maximum likelihood estimator (MLE):  $\hat{\theta}_n$  is MLE if  $\hat{\theta}_n \in \arg \max_{\theta \in \Theta} L_n(\theta)$   
Example: Gaussian model  $\theta \in \Theta$   $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$  MLE:

- $\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i \leq n} X_i$
- $\hat{\sigma}^2 = \frac{1}{n} \sum_i (X_i - \bar{X})^2$

Question: Why MLE? Answer: Simple + systematic + optimal

**Theorem 1** (MLE). *MLE is asymptotically normal and efficient.*

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{P} \mathcal{N}(0, I^{-1}(\theta))$$

Where the Fisher information  $I(\theta) := - \int [\frac{\partial^2}{\partial \theta^2} \log p_\theta(x)] p_\theta(x) dx$  We can construct from this convergence result CI, p-values. If  $\hat{\theta}_n$  is unbiased, then  $\text{var}(\hat{\theta}_n) \geq I^{-1}(\theta)$

## 2 Regression

**Definition 1** (Regression). *The art of summarizing relationship between two variables.*

$$\underbrace{Y}_{\text{response}} \overset{??}{\longleftrightarrow} \underbrace{X}_{\text{predictor / feature / covariate}}$$

In an other word, given data  $(Y_1, X_1), \dots, (Y_n, X_n) \overset{iid}{\sim} P_{Y,X}$ , we aim to find a mapping/function  $f$ , such that  $f(X)$  is close to  $Y$ . **Loss:**

- $l(f(X), Y) = |f(X) - Y|$ :  $L_1$ -loss.
- $l(f(X), Y) = |f(X) - Y|^2$ :  $L_2$ -loss.

*Risk:*

$$R(f) = \mathbb{E}l(f(X), y) = \mathbb{E}|f(X) - y|^2$$

**Theorem 2** ( $L_2$  loss). Let  $f^* := \arg \min_f \mathbb{E}|Y - f(X)|^2$  then  $f^*(x) = E[Y|X = x]$

Question: minimize  $R(f)$ , the expectation is w.r.t  $P_{Y,X}$  which is unknown.

Stochastic optimization problem:  $R(f) = E|Y - f(X)|^2 \xrightarrow{\text{Concentration}} \hat{R}(f) = \frac{1}{n} \sum_i (Y_i - f(X_i))^2$   $\hat{f} = \arg \min_f \hat{R}(f)$

A trivial solution:

$$f(x) = \begin{cases} Y_i & \text{for } x = X_i \\ \text{anything} & \text{otherwise} \end{cases}$$

$\Rightarrow$  Overfitting.

**Definition 2** (Overfitting). *A phenomenon when a statistical mode has too much flexibility (capacity) so that the models stats to fit the noise instead of just the signal.*

Solution to overfitting:

**Regularization:** Introduce additional information onr constraints to reduce the flexibility(capacity) of the model.