# Exercise no 5

FirstName LastName StudentNumber

Statistical Machine Learning

March 3, 2018

# 1 Mathematical Statistics

**Exercise 1.1** Derive the BIC. Suppose

$$\begin{aligned}\mathbf{y}_{n\times 1} \mid \mathbf{X}_{n\times p}, \boldsymbol{\beta}_{p\times 1} &\sim \mathcal{N}(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}) \\ \boldsymbol{\beta} \mid \mathbf{X} &\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}, \left\{\frac{1}{n}\mathbf{X}^\top\mathbf{X}\right\}^{-1}\sigma^2\right)\end{aligned}$$

- Show that

  $$-2\log f(\mathbf{y} \mid \mathbf{X}) = -2\log\left\{\int \cdots \int f(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\beta}) f(\boldsymbol{\beta} \mid \mathbf{X}) d\boldsymbol{\beta}\right\} = -2\log f(\mathbf{y} \mid \hat{\boldsymbol{\beta}}, \mathbf{X}) + p\log(n+1)$$

  Hint: first use the second order Tylor expansion of $\log f(y \mid \boldsymbol{\beta}, \mathbf{X})$ around $\hat{\boldsymbol{\beta}}$ and then take the integral. Note that this approximation is exact, because the original function and the Tylor expanded functions both are quadratic functions.

- For what $f(\boldsymbol{\beta} \mid \mathbf{X})$ the penalization term $\log(n+1)$ changes to $\log n$

  Hint: think about a constant function.

**Solution 1.1**

**Exercise 1.2** In many regression examples the error variance $\sigma$ is unknown. How do you compute AIC if $\sigma$ is unknown.

Note that the plug-in estimator of $\sigma$ cancels out the likelihood, i.e. $\hat{\sigma}^2 = \frac{1}{n-p}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

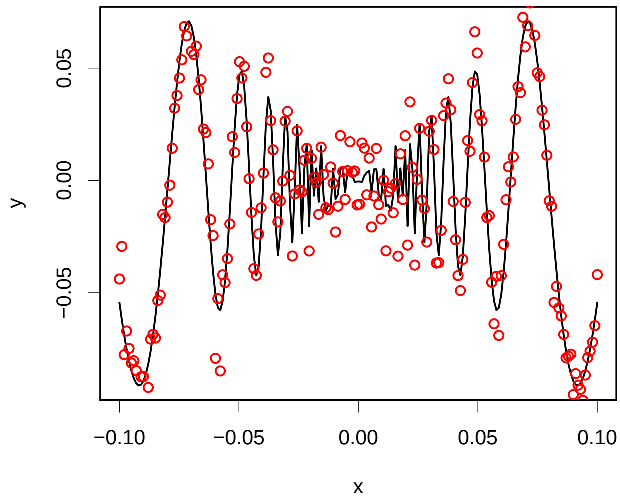
$$-2\log\text{likelihood} = \frac{1}{\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = n - p,$$

this means a naive AIC implementation reduces to AIC $= n - p + 2p = n + p$, which is only function of sample size and model dimension and is not function of data $y_i$ and predictions $\hat{y}_i$! A similar problem appears in BIC computation as well.

**Solution 1.2**

# 2 Computation

**Exercise 2.1** Suppose $x \in [-0.1, 0.1]$ and the unknown regression function $f(x) = x \sin(1/x)$. Simulate 500 data points from this model with error $\mathbb{V}(\varepsilon_i) = \sigma^2$. Set the random data generator seed to reproduce the same data and take $\sigma = 0.01$. Plot your data and the function such as the one below.




1. Step 1: use linear regression with $p = 5$ columns.

   (a) Use polynomial basis to estimate this unknown function.

   (b) Use Fourier basis to estimate this unknown function.

   (c) Choose equidistant $\xi_l$ and use the cubic spline basis $b_l(x) = \{\max(0, x - \xi_l)\}^3, l = 1, \ldots, p$.

   Plot all these expansion fits, and visually judge which one approximates the function better.

2. Step 2:

   - Choose an appropriate $p$ using BIC for your simulated data, and for all three above bases ($p$ might be different for each basis). For simplicity take $\sigma = 0.01$ to be known.
   - Choose an appropriate $p$ using leave-one-out.
   - Choose an appropriate $p$ using 10-fold cross-validation, take $B = 20$ and plot the estimated cross-validation error with its confidence bound.
   - Which basis do you prefer to use for this example? Why?

   Note: I recommend that you implement BIC, leave-one-out, and 10-fold cross-validation yourself, to make sure you understand how they work. It looks simple, but many researchers cannot implement $k$-fold cross-validation and its confidence bound properly.

**Solution 2.1**