Exercise no 6

FirstName LastName StudentNumber

Statistical Machine Learning

March 10, 2018

1 Mathematical Statistics

Exercise 1.1 Show that the kernel smoothing (weighted average) is the solution of the following optimization if $f_\theta(x) = \theta_0$

$$\hat{\theta}(x_0) = \text{argmin}_\theta \sum_{i=1}^{N} K(x_0, x_i)\{y_i - f_\theta(x_i)\}^2,$$

$$\hat{f}(x_0) = f_{\hat{\theta}}(x_0)$$

- Find the link between this optimization problem and the weighted linear regression.
- Find the solution of $\hat{f}$ for $f_\theta(x) = \theta_0 + \theta_1 x$?
- Find the solution of $\hat{f}$ for $f_\theta(x) = \theta_0 + \sum_{j=1}^{M} \theta_j x^j$?

Solution 1.1

2 Computation

Exercise 2.1 Take *lpsa* as the response variable y and *lcp* as the dependent variable x for the prostate cancer data set

`https://web.stanford.edu/~hastie/ElemStatLearn/datasets/prostate.data`

Use polynomial order $M \in \{0, 1, 2, 3\}$ local regression of Exercise 1.1 for different values of M, different kernels, different bandwidth λ on prostate cancer data. Which M, which kernel, which bandwidth λ do you prefer?

Use Rbf kernel, and use 10-fold cross-validation to tune λ for this data. Compare different fits as order $M \in \{0, 1, 2, 3\}$ changes in a graph.

Solution 2.1 Put your graph here.

3 Application

Exercise 3.1

- Use k-nearest neighbours classifier and tune k using 10-fold cross-validation on the zip data to recognize digit 3 from digit 8. All digits $\{0, \ldots, 9\}$ are available here.

  `https://web.stanford.edu/~hastie/ElemStatLearn/datasets/zip.train.gz`

- predict accuracy of the test set

  `https://web.stanford.edu/~hastie/ElemStatLearn/datasets/zip.test.gz`

Solution 3.1 Put the *confusion matrix* of the test set here.