EPFL, Autumn 2017
Emeric Thibaud Jonathan Koh

Statistical Machine Learning

Exercise sheet 1

Exercise 1.1 (Recap of linear models) In this exercise, consider the following setting. Let $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$ and $\mathbf{X}$ is a non-random full rank matrix of size $n \times p$. This setup contains the Gauss-Markov assumptions of a linear model.

(a) Write the residual sum of squares and derive the least squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$.

(b) Show that $\widehat{\boldsymbol{\beta}}$ is unbiased and that the variance of $\widehat{\boldsymbol{\beta}}$ is given by $\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

(c) Prove the Gauss-Markov theorem, i.e, $\widehat{\boldsymbol{\beta}}$ is the best **linear** unbiased estimator (BLUE) of $\boldsymbol{\beta}$. "Best" in the sense that for all other linear unbiased estimators $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, $\mathrm{var}(\widetilde{\boldsymbol{\beta}}) - \mathrm{var}(\widehat{\boldsymbol{\beta}})$ is a positive semidefinite matrix.

Hints: Recall that an estimator $\widetilde{\boldsymbol{\beta}}$ is linear if $\widetilde{\boldsymbol{\beta}} = \mathbf{A}\boldsymbol{y}$, for some $\mathbf{A} \in \mathbb{R}^{p\times n}$. Notice that the matrix $\mathbf{A}$ can be decomposed as $\mathbf{A} = \mathbf{D} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Exercise 1.2 (Linear regression and feature engineering) In this exercise, we will fit a linear model to data from `simreg1train.csv`. In R, use the `read.csv("...")` function to import the data.

(a) Using the results from Exercise 1.1(a), compute the least squares estimates for this dataset using your statistical software and plot the fitted values. Is the model appropriate?

(b) Calculate the training error for this dataset, given by $\mathrm{err}_f = \frac{1}{m}\sum_{i=1}^{m} L(y_i, \widehat{y}_i)$, where L is the squared error loss.

(c) For the same loss function, calculate the empirical test error for the test dataset `simreg1test.csv`.

(d) We will make the model more flexible by adding features to the design matrix $\mathbf{X}$. Add the feature $\boldsymbol{x}^2$ into your regression model, i.e., our design matrix becomes $\mathbf{X} = [\mathbf{1}, \boldsymbol{x}, \boldsymbol{x}^2]$. What happens to the training and test errors now?

(e) Add features up to $\boldsymbol{x}^k$ into your regression model, for $k = 3, 4, \ldots, 10$. Calculate the training and test errors for each $k = 1, \ldots, 10$. Make a plot of the training and test errors against k. Discuss. What happens when $k > 10$?

Exercise 1.3 (k-nearest-neighbors method) We now consider the classification training and test datasets given in `simclass1train` and `simclass1test`.

(a) Write a code that performs a kNN method for classification.

(b) Train your method on the dataset `simclass1train`, for $k = 1, 2, \ldots, 30$. Compute the training error, this time based on the zero-one loss for classification.

(c) For each k, calculate the test error of the kNN method with the dataset `simclass1test`. Plot a graph of the training and test errors over k. Discuss.

Exercise 1.4 (Curse of dimensionality, illustrated with a unit ball)

(a) Consider N data points $\boldsymbol{x}_i$ $(i = 1, \ldots, N)$ which are independent and uniformly distributed in a p-dimensional unit ball centered at the origin. Show the the median distance from the origin to the closest data point is given by

$$d(p, N) = \left\{ 1 - \left(\tfrac{1}{2}\right)^{1/N} \right\}^{1/p}. \tag{1}$$

Hint 1: Start with the definition of the median distance r, given via the probability

$$\Pr(||\boldsymbol{x}_i - \mathbf{0}|| > r, i = 1, \ldots, N) = \tfrac{1}{2}.$$

Hint 2: The volume of a p-dimensional ball of radius r is given by $C \times r^p$, where C is a constant. For example, when $p = 2$, $C = \pi$. When $p = 3$, $C = \frac{4\pi}{3}$. The value of C does not matter for this question.

(b) Assume we have $N = 500$ data points. Compare the value of $d(p, N)$ from (1) when $p = 1, 2, 3, 5, 10$. Discuss.

Exercise 1.5 (Curse of dimensionality, illustrated with a unit cube)

(a) Consider the nearest-neighbour proceduce for inputs uniformly distributed in a p-dimensional unit hypercube. Suppose we send out a hypercubical neighbourhood about a target point to capture a fraction r of the unit volume. What is the expected edge length of this hypercube?

(b) What is the expected edge length of the cube when $p = 10$ and $r = 0.1$? Discuss.