STA208: Homework 1

Prof. James Sharpnack

Due 4/4 in class

In the following, show all your work. Feel free to do all the analytical questions first and then include the code and output second, but the different parts and which question that you are answering should be clearly marked. Code should be as modular as possible, points will be deducted for code that is not reusable (i.e. not broken into general purpose functions), and in the case of gratuitous hard coding.

1. **(Learning paradyms)**

   Decribe the issues involved in the following learning problems using the terminology that we learned in the first lecture. Provide a sentence or two for each problem.

   (a) A 'smart farm' has distributed sensors that detect moisture levels, and the farmers know what are the ideal moisture levels for each plant. They have many controls that adjust the irrigation system and they would like to know what settings produce the most ideal moisture levels.

   (b) Astronomers are trying to map the structure of the universe in terms of how galaxies cluster and form topological structures that they call filaments.

   (c) An online ad company wants to determine which of many ads to show each user based on their browser cookies.

   (d) NASA is mapping the strength of the gravitational field on the surface of Mars. They want you to help with determining its values in a grid of locations on the surface from remote sensing measurements.

2. **(Bayes rule)**

   Consider the classification setting with features $\mathbf{x} \in \mathbb{R}^p$ and response $y \in \{0, 1\}$. Suppose that we know the joint distribution of $\mathbb{P}(\mathbf{x}, y)$, and the conditional distributions $\mathbb{P}(y|\mathbf{x}), \mathbb{P}(\mathbf{x}|y)$ (an unlikely setting, but bear with me).

   (a) Under the Hamming loss, what is the true risk of a classifier $\hat{y} : \mathbb{R}^p \to \{0, 1\}$? Write it in terms of conditional distributions.

   (b) What is the Bayes rule, i.e. the classifier that minimizes the true risk?

   (c) Write in one sentence, what is the Bayes rule, as if you needed to describe what the Bayes risk was to someone in an elevator before you reached the lobby.

   (d) Prove that the Bayes risk is $1 - \mathbb{P}(y = y^\star(\mathbf{x})|\mathbf{x})$ where $y^\star$ is the Bayes rule.

3. **(Linear Regression)**

   Suppose that we are in the regression setting, $\mathbf{x}_i \in \mathbb{R}^p, y_i \in \mathbb{R}$ are n pairs drawn iid, let $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X}^\top = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$, and consider the linear regression estimator

   $$\hat{\boldsymbol{\beta}} := \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2. \tag{1}$$

   (a) When is the solution to this program unique? In this case, what is the unique minimizer $\hat{\boldsymbol{\beta}}$?

   (b) Give an equation that the minimizers satisfy regardless of uniqueness?

   (c) Given a solution to (1), give a reasonable prediction rule $\hat{y} : \mathbb{R}^p \to \mathbb{R}$.

(d) Suppose that $\mathbb{E}[y_i|\mathbf{x}_i] = \mathbf{x}_i^\top \boldsymbol{\beta}$ for $i = 1 \ldots, n+1$. For a new random draw $(\mathbf{x}_{n+1}, y_{n+1})$, then what is the bias of $\hat{y}(\mathbf{x}_{n+1})$, i.e. $\mathbb{E}[\hat{y}(\mathbf{x}_{n+1}) - y_{n+1}]$?

4. **(Simulation and ridge regression.)**

(a) Simulate $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^p$ with $p = 12$ and $n = 200$, iid normal with mean $\mathbf{0}$ and variance $\boldsymbol{\Sigma}$ such that

$$\Sigma_{j,k} = \rho^{|j-k|}, \quad j, k = 1, \ldots, p. \tag{2}$$

for $0 < \rho < 1$. Draw $\boldsymbol{\beta} \in \mathbb{R}^p$ such that β_j are iid normal with mean 0 and variance 1, and y_i independently normal with mean $\mathbf{x}_i^\top \boldsymbol{\beta}$ and variance 1. Print out your code (not the output), which should consist of functions for generating these objects.

(b) Derive an analytical expression for the solution to ridge regression,

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2, \tag{3}$$

as a function of $\mathbf{X}, \mathbf{y}, \lambda$.

(c) Provide code for solving ridge regression. Use any linear solver you like.

(d) Set $\rho = 0.5$. Simulate the bias of $\hat{\beta}$, $\|\mathbb{E}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$, the variance, $\mathbb{E}\|\hat{\boldsymbol{\beta}} - \mathbb{E}\hat{\boldsymbol{\beta}}\|_2^2$, and the mean square error, $\mathbb{E}\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2^2$, for many values of λ. Be sure that you see instances of overfitting and underfitting and can clearly see the point where λ is optimal. Plot these curves as functions of lambda and include your code.

5. **(Airfoil)**

Download the airfoil dataset, which is linked in the homework section of the course site. We will focus on predicting the scaled sound pressure, which is the 6th row.

(a) Set aside a test set at random.

(b) Form the coefficients for ordinary least squares with the training set. Write a function with a new $\mathbf{x}$ and $\hat{\boldsymbol{\beta}}$ as arguments and returns the prediction. Use any linear solver/Cholesky decomposition you like.

(c) Write a function that takes a new $\mathbf{x}$, k, and the training data, and outputs the k-nearest neighbor prediction with Euclidean distance.

(d) Write a function that takes a new $\mathbf{x}$, a bandwidth parameter, and the training data, and outputs the kernel prediction with boxcar kernel and Euclidean distance.

(e) Evaluate your methods on the test set, calculating the test error (empirical risk on the test set). Vary the tuning parameters and plot the test error as a function of the tuning parameters.

(f) Is the best test error a good estimate of the true risk for these methods? Why/why not? What can be done to estimate the true risk?