CAP 6610 Machine Learning, Spring 2025

Homework 1 Solution

1. *Regression to the mean.* Consider a data set in which the (scalar) x_i is the parent's height (average of mother's and father's height), and y_i is their child's height. Assume that over the data set the parent and child heights have the same mean value μ, and the same standard deviation σ. We will also assume that the correlation coefficient ρ is strictly between 0 and 1. (These assumptions hold, at least approximately, in real data sets that are large enough). Consider the least squares regression model that predicts the child's height from the parent's height. Show that this prediction of the child's height lies (strictly) between the parent's height and the mean height μ (unless the parent's height happens to be exactly the mean μ). For example, if the parents are tall, i.e., have height above the mean, we predict that the child will be shorter, but still taller than the mean. This phenomenon, called *regression to the mean*, was first observed by the early statistician Sir Francis Galton (who indeed, studied a data set of parent's and child's heights).

   Hint. Applying the formula derived on page 14 of `lec1.pdf` slides, in the simpler case that x is a scalar, the prediction model is

   $$\hat{y} = \mu_y + \rho\frac{\sigma_y}{\sigma_x}(x - \mu_x),$$

   where we used the fact that the correlation coefficient is defined as $\rho = s_{xy}/\sigma_x\sigma_y$. Plug in the assumption that $\mu_x = \mu_y = \mu$ and $\sigma_x = \sigma_y = \sigma$ to show the above.

   Solution. From the prediction model, we have

   $$\hat{y} = \mu + \rho(x - \mu) = (1 - \rho)\mu + \rho x.$$

   Since ρ satisfies $0 < \rho < 1$, we see that the predicted child's height is a convex combination of the mean μ and parent's height x. But a convex combination of any two numbers lies between the two numbers. When the parents are tall, *i.e.*, $x > \mu$, we have

   $$\mu < \hat{y} < x,$$

   i.e., the predicted child's height is above average but also less than the parent's height.

2. *More bases always decreases training error.* Consider two least squares regression models, one with m basis functions

   $$f(\boldsymbol{x}) = \theta_1\varphi_1(\boldsymbol{x}) + \cdots + \theta_m\varphi_m(\boldsymbol{x}),$$

   and the other with the same m bases plus another one

   $$\tilde{f}(\boldsymbol{x}) = \tilde{\theta}_1\varphi_1(\boldsymbol{x}) + \cdots + \tilde{\theta}_m\varphi_m(\boldsymbol{x}) + \tilde{\theta}_{m+1}\varphi_{m+1}(\boldsymbol{x}).$$

   Show that on a training data set $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, we always have

   $$\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(\boldsymbol{x}_i)\right)^2 \geq \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \tilde{f}(\boldsymbol{x}_i)\right)^2.$$

   (Of course, it doesn't mean $\tilde{f}(\boldsymbol{x})$ is a better prediction model. It is their performance on the test set that matters.)

Solution. Suppose $\tilde{\boldsymbol{\theta}}^*$ is the optimal solution for $\tilde{f}(\boldsymbol{x})$. For any vector $\tilde{\boldsymbol{\theta}}$, we have

$$\frac{1}{n}\|\boldsymbol{y} - \tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\theta}}^*\|^2 \leq \frac{1}{n}\|\boldsymbol{y} - \tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\theta}}\|^2,$$

where $\tilde{\boldsymbol{\Phi}}_{n\times(m+1)} = [\boldsymbol{\Phi}, \boldsymbol{\varphi}_{m+1}]$ and $\boldsymbol{\Phi}_{n\times m} = [\boldsymbol{\varphi}_1, \cdots, \boldsymbol{\varphi}_m]$.

Suppose $\boldsymbol{\theta}^*$ is the optimal solution for $f(\boldsymbol{x})$. Given $\tilde{\boldsymbol{\theta}} = \begin{bmatrix}\boldsymbol{\theta}^* \\ 0\end{bmatrix}$, inequality still satisfies, *i.e.*,

$$\frac{1}{n}\|\boldsymbol{y} - \tilde{\boldsymbol{\Phi}}\tilde{\boldsymbol{\theta}}^*\|^2 \leq \frac{1}{n}\left\|\boldsymbol{y} - [\boldsymbol{\Phi}, \boldsymbol{\varphi}_{m+1}]\begin{bmatrix}\boldsymbol{\theta}^* \\ 0\end{bmatrix}\right\|^2 = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\theta}^*\|^2.$$

Q.E.D.

3. *Numerical check of the least squares solution.* Use your favorite language to generate a random 40×10 matrix $\boldsymbol{\Phi}$ and a random 40-vector $\boldsymbol{\psi}$. Compute the least squares solution $\boldsymbol{\theta}^\star = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\psi}$ and the associated loss $\|\boldsymbol{\Phi}\boldsymbol{\theta}^\star - \boldsymbol{\psi}\|^2$. (There may be several ways to do this, depending on the software package you use. In MATLAB or Julia, the command is simply `Phi\psi`.) Generate a random 10-vectors $\boldsymbol{\delta}$ and verify that $\|\boldsymbol{\Phi}(\boldsymbol{\theta}^\star + \boldsymbol{\delta}) - \boldsymbol{\psi}\|^2 > \|\boldsymbol{\Phi}\boldsymbol{\theta}^\star - \boldsymbol{\psi}\|^2$ holds. Repeat several times with different values of $\boldsymbol{\delta}$.

   Submit your code, including the code that checks whether the expected inequality that involves $\boldsymbol{\delta}$ holds.

   Solution. Complete code and unambiguous results will be accepted.

4. We test the performance of ridge regression methods on the wine data set `http://archive.ics.uci.edu/ml/datasets/Wine+Quality`. We will only consider the red wine data set, with 1599 samples. We use the first 1400 samples for training, and the last 199 samples for testing. The goal is to build a ridge regression model of the first 11 features (together with a constant term) to predict the quality of the wine, by solving the following least squares problem

$$\underset{\boldsymbol{w},\beta}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top\boldsymbol{w} + \beta - y_i)^2 + \lambda\|\boldsymbol{w}\|^2,$$

   with $\lambda = 0, 10^{-3}, 10^{-2}, 10^{-1}$. Report their prediction performance on the test set in terms of mean squared error (MSE).

   Solution. The least squares problem is equivalent to

$$\underset{\boldsymbol{w},\beta}{\text{minimize}} \quad \|[\boldsymbol{X} \quad \mathbf{1}]\begin{bmatrix}\boldsymbol{w} \\ \beta\end{bmatrix} - \boldsymbol{y}\|^2 + \lambda\|\boldsymbol{w}\|^2,$$

   which is also equivalent to

$$\underset{\boldsymbol{w},\beta}{\text{minimize}} \quad \| \begin{bmatrix}\boldsymbol{X} & \mathbf{1} \\ \sqrt{\lambda}\boldsymbol{I} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{w} \\ \beta\end{bmatrix} - \begin{bmatrix}\boldsymbol{y} \\ \mathbf{0}\end{bmatrix} \|^2.$$

   Taking $\Phi = \begin{bmatrix}\boldsymbol{X} & \mathbf{1} \\ \sqrt{\lambda}\boldsymbol{I} & \mathbf{0}\end{bmatrix}$, $\hat{\boldsymbol{w}} = \begin{bmatrix}\boldsymbol{w} \\ \beta\end{bmatrix}$ and $\boldsymbol{\psi} = \begin{bmatrix}\boldsymbol{y} \\ \mathbf{0}\end{bmatrix}$, the solution of this problem is $\hat{\boldsymbol{w}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\psi}$.

   The MSEs on the test set are 0.4869, 0.4873, 0.4876, and 0.4877, respectively.

5. We test the performance of regularized classification methods on the ionosphere data set `https://archive.ics.uci.edu/ml/datasets/ionosphere`. There are 351 samples. We use the first 300 samples for training, and the last 51 samples for testing. The goal is to build a classification model of the 34 features (together with a constant term) to predict the binary (± 1) outcome. All models are trained by solving the following optimization problem

$$\underset{\boldsymbol{w},\beta}{\text{minimize}} \quad \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^\top \boldsymbol{w} + \beta - y_i)^2 + \lambda\|\boldsymbol{w}\|^2,$$

with $\lambda = 0, 10^{-3}, 10^{-2}, 10^{-1}$. Report their prediction accuracy (ratio of correct predictions over the total number of samples) on the test set.

Solution. Same as the last question, the solution of this problem is $\hat{\boldsymbol{w}} = (\boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{\psi}$, given $\Phi = \begin{bmatrix} \boldsymbol{X} & \mathbf{1} \\ \sqrt{\lambda}\boldsymbol{I} & \mathbf{0} \end{bmatrix}$, $\hat{\boldsymbol{w}} = \begin{bmatrix} \boldsymbol{w} \\ \beta \end{bmatrix}$ and $\boldsymbol{\psi} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix}$.

The prediction accuracies on the test set are all 100%.