MATH 829 - Spring 2016
Introduction to data mining and analysis
Homework #1

Instructions:

(1) You are allowed to work in groups of 1-4.
(2) Show (and briefly explain) all of your work to receive full credit.
(3) Submit your work before **Friday, February 26th, 2016**.

Theoretical part

Problem 1.

a) Consider the linear regression problem $y = X\beta$ with only one predictor, i.e., $y, X \in \mathbb{R}^{n\times 1}$. Prove that the least squares solution can be written as

$$\hat{\beta}_{\text{LS}} = \frac{\langle y, X\rangle}{\langle X, X\rangle},$$

where $\langle a, b\rangle := \sum_{i=1}^n a_i b_i$ denotes the usual dot product on $\mathbb{R}^n$.

b) More generally, suppose $X \in \mathbb{R}^{n\times p}$ has orthogonal columns $x_1, \dots, x_p \in \mathbb{R}^n$. Show that the least squares solution $\hat{\beta}_{\text{LS}} = (\beta_1, \dots, \beta_p)$ satisfies:

$$\beta_i = \frac{\langle y, x_i\rangle}{\langle x_i, x_i\rangle} \qquad (i = 1, \dots, p).$$

c) Explain why Algorithm 3.1 in ESL provides the least squares solution in the general case where the columns are not orthogonal and an intercept is included into the model.

Problem 2. Let $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n\times p}$. Suppose the first column of X is $\mathbf{1}_{n\times 1} = (1, 1, \dots, 1)^T \in \mathbb{R}^n$ so that the model includes an intercept. Let $\hat{y} = X\hat{\beta}$ denote the least squares estimate of y. Denote the residuals by $\hat{\varepsilon} := y - \hat{y}$.

a) Prove that the mean of the residuals is equal to zero, i.e., show that $\frac{1}{n}\sum_{i=1}^n \hat{\varepsilon} = 0$. (Hint: According to the normal equations, $X^T(X\hat{\beta} - y) = -X^T\hat{\varepsilon} = 0$.)

b) Provide a simple example to show that the residuals may not have mean zero if an intercept is not included in the model.

Problem 3. Let $y, \hat{y}, X, \hat{\varepsilon}$ be as in Problem 2. The following steps will guide you to prove that in a least squares model with an intercept, the R^2 coefficient is equal to the square of the sample correlation coefficient between the output and the predicted values.

Let $\overline{y} = \frac{1}{n}\sum_{i=1}^n y_i = (\mathbf{1}_{n\times 1}^T y)/n$ and $\overline{\hat{y}} = \frac{1}{n}\sum_{i=1}^n \hat{y}_i = (\mathbf{1}_{n\times 1}^T \hat{y})/n$. Define

$$\text{SSE} := \sum_{i=1}^n (y_i - \hat{y}_i)^2 \qquad \text{(error sum of squares)},$$

$$\text{SST} := \sum_{i=1}^n (y_i - \overline{y})^2 \qquad \text{(total sum of squares)}.$$

Recall that we defined in class

$$R^2 = 1 - \frac{SSE}{SST}.$$

Define the sample variance of y and $\hat{y}$ by

$$\hat{V}(y) := \frac{1}{n} \cdot \mathrm{SST} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2, \qquad \hat{V}(\hat{y}) = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{\hat{y}})^2.$$

Also, denote the sample covariance and sample correlation coefficients between $z, w \in \mathbb{R}^n$ respectively by

$$\hat{C}(z,w) := \frac{1}{n}\sum_{i=1}^{n}(z_i - \overline{z})(w_i - \overline{w}), \qquad \hat{\rho}(z,w) := \frac{\hat{C}(z,w)}{\sqrt{\hat{V}(z)}\sqrt{\hat{V}(w)}}.$$

a) Prove that the prediction $\hat{y}$ is uncorrelated with the residuals, i.e., show that

$$\hat{C}(\hat{y}, \hat{\varepsilon}) = 0.$$

(Hint: By proceeding as in Problem 2, we obtain $X^T\hat{\varepsilon} = 0$. Use $\hat{y} = X\hat{\beta}$, to conclude that $\hat{y}^T\varepsilon = 0$.)

b) Define $A := I_n - \frac{1}{n}\mathbf{1}_{n\times 1}\mathbf{1}_{n\times 1}^T \in \mathbb{R}^{n\times n}$. Verify the relations

$$\begin{aligned} &A^TA = A^2 = A, \\ &Ay = y - \overline{y}\mathbf{1}_{n\times 1}, \\ &\hat{C}(y,\hat{y}) = \frac{1}{n}(Ay)^TA\hat{y} = \frac{1}{n}y^TA\hat{y}. \end{aligned}$$

c) Show that $\hat{C}(y,\hat{y}) = \hat{V}(\hat{y})$. (Hint: Use $\hat{C}(y,\hat{y}) = \frac{1}{n}y^TA\hat{y} = \frac{1}{n}(\hat{y}+\hat{\varepsilon})^TA\hat{y}$).

d) Show that $\hat{V}(y) = \hat{V}(\hat{y}) + \hat{V}(\hat{\varepsilon})$. (Hint: Write $\hat{V}(y) = \frac{1}{n}y^TAy$, substitute $y = \hat{y} + \hat{\varepsilon}$, and simplify).

e) Use the previous calculations to conclude that $R^2 = \hat{\rho}(y,\hat{y})^2$.

Python part

Problem 4.

a) Implement Algorithm 3.1 of ESL in Python.

b) Use scikit-learn to verify that your implementation is correct.

Problem 5.

a) Verify the relation $R^2 = \hat{\rho}(y,\hat{y})^2$ proved in Problem 3 e) on randomly generated data.

b) Show that the relation is generally false if an intercept is not included into the model.

Problem 6. The file `water.csv` contains data from a study that relates water fluoridation and cavity rates for 7,257 children in 21 cities. The variable `FLUORIDE` contains the level of fluoride in public water supplies for each city (in parts per million), and the variable `CARIES` the number of cavities per 100 children.

a) Display the data using a scatter plot. Does the relationship between the variables look linear? Can you see outliers? Should you keep outliers into the dataset?

b) Fit a linear regression model to the data and compute the average training error of the model.

c) Fit a model of the form $\log(\texttt{CARIES}) = \log(c + \texttt{FLUORIDE}) \cdot \beta$ where $c > 0$ is a constant. Optimize on c to pick the model with the smallest training error.

d) Draw a histogram of the residuals for the model chosen in c).

Problem 7.

a) Compute the prediction error for all subsets of variables for the `cars` dataset.

b) Build the best model you can to predict the price of a car by selecting a good subset of predictors and using transformations of the variables.