Universität zu Köln

Dr. Boqiang Huang

Mathematical Foundations of Data Analysis II — SS 2019

# Problem Set 02

Published on 23.04.2019 To be collected on 30.04.2019

**Problem 1:** (5 points)
Considering a general statistical model $y = f(x) + \epsilon$ with its estimated relationship $\hat{f}$ between $X$ and $Y$, show that the expected test mean-square-error (MSE) can always be decomposed into the sum of three fundamental quantities: the variance of $\hat{f}(x)$, the squared bias of $\hat{f}(x)$ and the variance of the error terms $\epsilon$:

$$\begin{aligned} \mathrm{E}[(y - \hat{f}(x))^2] &= \mathrm{Var}[\hat{f}(x)] + (\mathrm{Bias}[\hat{f}(x)])^2 + \mathrm{Var}[\epsilon] \\ &= \mathrm{E}[(\hat{f}(x) - \mathrm{E}[\hat{f}(x)])^2] + (\mathrm{E}[\hat{f}(x)] - f(x))^2 + E[\epsilon^2]. \end{aligned} \tag{1}$$

**Problem 2:** (5 points)
Let $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ represent $n$ observation pairs, each of which consists of a measurement of $X$ and a measurement of $Y$. Let $\hat{\beta}_0, \hat{\beta}_1$ represent the coefficient estimates of a linear model $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{y}_i$ is the prediction of $Y$ based on the $i$th value of $X$. Then $e_i = y_i - \hat{y}_i$ represents the $i$th residual, or the difference between the $i$th observed response value and the $i$th response value that is predicted by the liner model. The residual sum of squares (RSS) could be defined as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2, \tag{2}$$

or equivalently as

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2. \tag{3}$$

Please show that the least squares coefficient estimates for this simple linear regression problem will be

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned} \tag{4}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$ and $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$ are the sample means.

**Problem 3:** (5 points)
There are five students randomly selected to take a MFDA-I test before they began their FMDA-II course. We are going to investigate the relationship between the MFDA-I scores and the MFDA-II scores, which may illustrate how much knowledge that one understood from MFDA-I could support her/his study in MFDA-II.

| Student | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| FMDA-I ($x_i$) | 95 | 85 | 80 | 70 | 60 |
| FMDA-II ($y_i$) | 85 | 95 | 70 | 65 | 70 |

Applying a simple linear regression model to investigate the least squares coefficient estimates $\hat{\beta}_0, \hat{\beta}_1$, the confidence interval for $\hat{\beta}_1$, the $\mathrm{R}^2$ statistic, and the correlation $r = \mathrm{Cor}(X, Y)$.

**Problem 4:** (5 points)
A doctor is going to build up a case-based diagnosis system. Each case contains a number of features describing possible symptoms and the corresponding solution represents the classification of disease. The training data set is shown below

| Training | Fever (F) | Vomiting (V) | Diarrhea (D) | Shivering (Sh) | Classification |
|---|---|---|---|---|---|
| $c_1$ | no | no | no | no | healthy (H) |
| $c_2$ | average | no | no | no | influenza (I) |
| $c_3$ | high | no | no | yes | influenza (I) |
| $c_4$ | high | yes | yes | no | salmonella poisoning (S) |
| $c_5$ | average | no | yes | no | salmonella poisoning (S) |
| $c_6$ | no | yes | yes | no | bowel inflammation (B) |
| $c_7$ | average | yes | yes | no | bowel inflammation (B) |

Based on the experience, the doctor has an approximated similarity measure using local similarity measures and feature weights as specified below

$sim_F$

| q \ c | no | avg | high |
|---|---|---|---|
| no | 1.0 | 0.7 | 0.2 |
| avg | 0.5 | 1.0 | 0.8 |
| high | 0.0 | 0.3 | 1.0 |

$sim_V$=$sim_D$=$sim_{Sh}$

| q \ | yes | no |
|---|---|---|
| yes | 1.0 | 0.0 |
| no | 0.2 | 1.0 |

Weights

$w_F$=0.3

$w_V$=0.2

$W_D$=0.2

$w_{Sh}$=0.3

Try to calculate the similarity between all cases from the data set and the query $q = (\text{high}, \text{no}, \text{no}, \text{no})$.