Assignment 3

CS780/880: Introduction to Machine Learning

Due: By 12:40PM Thu Mar 9th, 2017
Submission: Turn in as a PDF on myCourses, or printed and turned in at class
Discussion forum: https://piazza.com/unh/spring2017/cs780cs880

Problem 1 [15%] Suppose we estimate the regression coefficients in a linear regression model by choosing β to minimize:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right) + \lambda\sum_{j=1}^{p}\beta_j^2$$

for a particular value of λ. For parts (a) through (e), indicate which of (i.) through (v.) is correct. *Briefly* justify your answer.

(a) As we increase λ from 0, the *training RSS* will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.
ii. Decrease initially, and then eventually start increasing in a U shape.
iii. Steadily increase.
iv. Steadily decrease.
v. Remain constant.

(b) Repeat (a) for *test RSS*.

(c) Repeat (a) for variance.

(d) Repeat (a) for (squared) bias.

Problem 2 [15%] You will derive the bias-variance decomposition of MSE as described in Eq. (2.7) in ISL. The bias-variance decomposition is defined as follows. For simplicity, assume that $\text{Var}[\epsilon] = 0$, in which case the decomposition becomes:

$$\underbrace{\mathbb{E}[(y_0 - \hat{f}(x_0))^2]}_{\text{test MSE}} = \underbrace{\text{Var}[\hat{f}(y_0)]}_{\text{Variance}} + \left(\underbrace{\mathbb{E}[f(y_0) - \hat{f}(y_0)]}_{\text{Bias}}\right)^2.$$

In deriving the decomposition, take the following steps:

(a) Using $y_0 = f(x_0)$ rewrite the test MSE as a function of the estimated model $\hat{f}$ and the *true* model f

(b) Rewrite the test MSE using the following property of the variance operator: $\mathbb{E}[W^2] = \text{Var}[W] + \mathbb{E}[W]^2$. Substitute the appropriate value for W.

(c) Derive the decomposition using another property of variance: $\text{Var}[A \pm B] = \text{Var}[A] + \text{Var}[B]$ if A and B *are independent* (note $\pm$ on the left and $+$ on the right). Also note that $f(x_0)$ is a *constant* since it does not depend on data, and thus $\text{Var}[f(x_0)] = 0$.

CS880 Graduate: Problem 3 [30%] You will now derive the Bayesian connection to the lasso and ridge regression discussed in Section 6.2.2. of ISL.

(a) Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from a normal distribution $N(0, \sigma^2)$ distribution. Write out the likelihood for the data as a function of values β.

(b) Assume the following prior for $\beta : \beta_1, \ldots, \beta_p$ are independent and identically distributed according to a double-exponential distribution with mean 0 and common scale parameter b: i.e. $p(\beta) = \frac{1}{2b}\exp(-|\beta|/b)$. Write out the posterior for β in this setting using Bayes theorem.

(c) Argue that the lasso estimate is the value of β with maximal probability under this posterior distribution. Compute log of the probability in order to make this point. *Hint: The denominator (= the probability of data) can be ignored in computing the maximum probability.*

(d) Now assume the following prior for $\beta : \beta_1, \ldots, \beta_p$ are independent and identically distributed according to a normal distribution with mean zero and variance c. Write out the posterior for β in this setting using Bayes theorem.

(e) Argue that the ridge regression estimate is the value of β with maximal probability under this posterior distribution. Compute log of the probability in order to make this point. *Hint: The denominator (= the probability of data) can be ignored in computing the maximum probability.*

CS780 Undergraduate: Problem 3 [30%] You will now derive the Bayesian connection to the ridge regression discussed in Section 6.2.2. of ISL.

(a) Suppose that $y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \epsilon_i$ where $\epsilon_1, \ldots, \epsilon_n$ are independent and identically distributed from a normal distribution $N(0, 1)$ distribution. Write out the likelihood for the data as a function of values β.

(b) Assume the following prior for $\beta : \beta_1, \ldots, \beta_p$ are independent and identically distributed according to a normal distribution with mean zero and variance c. Write out the posterior for β in this setting using Bayes theorem.

(c) Argue that the ridge regression estimate is the value of β with maximal probability under this posterior distribution. Compute log of the probability in order to make this point. *Hint: The denominator (= the probability of data) can be ignored in computing the maximum probability.*

Problem 4 [20%] In this problem, you will perform cross-validation on a simulated data set.

(a) Generate a simulated data set as follows:

```
set .seed(1)
y = rnorm(100)
x = rnorm(100)
y = x - 2*x^2 + rnorm(100)
```

In this data set, what is n and what is p? Write out the model used to generate the data in equation form.

(b) Create a scatterplot of X against Y. Comment on what you find.

(c) Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:

1. $Y = \beta_0 + \beta_1 X + \epsilon$
2. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
3. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$
4. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

Note you may find it helpful to use the `data.frame()` function to create a single data set containing both X and Y.

(d) Repeat (c) using another random seed, and report your results. Are your results the same as what you got in (c)? Why?

(e) Which of the models in (c) had the smallest LOOCV error? Is this what you expected? Explain your answer.

Problem 5 [20%] We will now try to predict per capita crime rate in the Boston data set.

(a) Try out some of the regression methods explored in this chapter, such as best subset selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

(b) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using validation set error, cross-validation, or some other reasonable alternative, as opposed to using training error.

(c) Does your chosen model involve all of the features in the data set? Why or why not?