Problem Set

Machine Learning Concepts

CEU, Winter 2024

You should upload your work to Moodle by 5pm on Friday, March 15. Please work on your own, you can ask me or Janos Divenyi for help if you are stuck.

1. Consider the simplest possible predicitive model

$$Y_i = \beta_0 + \epsilon_i$$

where ϵ_i, $i = 1, \dots, n$ are independent and identically distributed random variables with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$. (Thus, in this model we try to predict Y using only a constant, i.e., $f(x) = \beta_0$ regardless of any X variables we might have at our disposal.) The ridge estimator of β_0 solves

$$\min_b \left[\sum_{i=1}^{n} (Y_i - b)^2 + \lambda b^2 \right]$$

for some $\lambda \geq 0$. In the special case $\lambda = 0$, the solution is of course the OLS estimator. It is easy to show that the general solution to this problem is given by $\hat{\beta}_0^{ridge} = \sum_{i=1}^{n} Y_i/(n + \lambda)$.

a) How does the regularized estimator (predictor) $\hat{\beta}_0^{ridge}$ compare with the OLS estimator?

b) Suppose that $\beta_0 = 1$ and $\epsilon \sim N(0, \sigma^2)$ with $\sigma^2 = 4$. Generate a sample of size $n = 20$ from the model and compute the predicted value $\hat{Y} = \hat{f}(x) = \hat{\beta}_0^{ridge}$ for a grid of λ values over the interval $[0, 20]$.

c) Repeat part b), say, 1000 times so that you end up with 1000 estimates of β_0 for all the λ values that you have picked. For each value of λ, compute $bias^2[\hat{\beta}_0^{ridge}]$, $Var[\hat{\beta}_0^{ridge}]$ and $MSE[\hat{\beta}_0^{ridge}] = bias^2[\hat{\beta}_0^{ridge}] + Var[\hat{\beta}_0^{ridge}]$.

d) Plot $bias^2[\hat{\beta}_0^{ridge}]$, $Var[\hat{\beta}_0^{ridge}]$ and $MSE[\hat{\beta}_0^{ridge}]$ as a function of λ and interpret the results. Can a ridge regression give a better prediction than OLS?

2. ISLR Exercise 3 in Section 6.8 (p. 260). Please use the version of the textbook posted online (7th printing).

3. Consider the 'dense' regression model discussed in the last lecture:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_{50} X_{50,i} + \epsilon_i,$$

where $X_1, \ldots, X_{50}$ are correlated jointly normal random variables, $\epsilon \sim N(0, 2^2)$, and the regression coefficients are arbitrarily chosen numbers between 0 and 1 (not shown).

The file PCA_data.csv contains a training sample of size $n = 500$ and a test sample of size $m = 500$ generated from this model. The exercise below asks you to do the exercise that produces the last column of the 'Dense DGP' table on slide 22, Lecture 3. (I posted it on Moodle.)

a) Load the data set called `PCA_data.csv` (posted). Designate the first $N_{tr} = 500$ observations as the training sample and the last $N_{te} = 500$ as the test sample.

b) Compute the first 10 principal component vectors and the corresponding scores $Z_1^*, \ldots, Z_{10}^*$ for $(X_1, X_2, \ldots, X_{50})$. For simplicity, you can use the whole data set for this (both the training sample as well as the test sample).

c) Estimate an OLS regression of Y on a constant and $X_1, \ldots, X_{50}$ over the training sample. Estimate OLS regressions of Y on a constant and $Z_1^*, \ldots, Z_k^*$ over the training sample for $k = 1, 5, 10$.

d) Use the four models estimated under part c) the obtain predictions for the outcomes Y_i in the test sample. Compute the mean squared prediction error for the four different predictions and report these numbers. You should get results similar to those on slide 22, but there will be some differences because the whole experiment is performed only once. (The slide averages over many experiments.)

e) Consider again the original 'Dense DGP' table on slide 22, Lecture 3. Discuss and explain the MSPE patterns you see in the first column ($N_{tr} = 75$) and the last column ($N_{tr} = 500$).