Machine Learning for Economic Analysis
Problem Set 1

Jonas Lieber*

Due: 11:59pm Wednesday, January 24, 2023

Problem 1. *Consider the model*

$$Y = f(X) + U,$$

where $Y, U \in \mathbb{R}$ *and* $X \in \mathbb{R}^k$ *and* $f : \mathbb{R}^k \to \mathbb{R}$*. Denote the joint distribution of* (X, Y, U) *by* μ *and assume that*

1. $E_\mu[U|X] = 0$,
2. $E_\mu[U^2|X] = \sigma^2$.

Consider an i.i.d. dataset $(X_1, Y_1, U_1), \ldots, (X_n, Y_n, U_n)$ *drawn from* μ*. Assume that the researcher observes only* $(X_1, Y_1), \ldots, (X_n, Y_n)$*. Suppose the researcher runs some algorithm to produce an estimate of* f *using this dataset. Denote this estimate by* $\hat{f}$*. We are interested in the performance of* $\hat{f}$ *on an unseen independent datapoint* (X^*, Y^*)*, identically distributed to the data, in the conditional mean-square sense, i.e.,*

$$MSE(\hat{f}|X^*) := \mathbb{E}[(\hat{f}(X^*) - Y^*)^2|X^*].$$

1. ***For graduate students only, all following parts are for all students*** *Show the following decomposition*

$$MSE(\hat{f}|X^*) = \mathbb{E}\left[\left(\hat{f}(X^*) - \mathbb{E}[\hat{f}(X^*)|X^*]\right)^2 \middle| X^*\right] + \left(\mathbb{E}[\hat{f}(X^*)|X^*] - f(X^*)\right)^2 + \sigma^2. \tag{1}$$

2. *Consider the function* $f(x) = sin(x)$ *from* $x = 0$ *to* $x = 2\pi$*. We consider linear regression of a polynomial expansion.*

*Department of Economics, Yale University. jonas.lieber@yale.edu

(a) *Write a function that creates a sample for reproducible pseudo-random* $x \sim U[0, 2\pi]$ *and*

$$Y|X = x \sim \mathcal{N}(sin(x), 0.2)$$

of length $n = 30$.

(b) *Plot* $x, f(x)$ *and* y.

(c) *Regress* y *on a polynomial expansion of* x *of degree up to* d, *i.e., a regression on* $x^0, x^1, \ldots, x^d$. *For* $d = 0, 1, 2, 3, 29$, *plot*

i. *the true function* $f(x)$ *and the estimated functions* $\hat{f}_d(x)$, *for the different values of* d *together with your random sample (as "fat" dots),*

ii. *the error* $f(x) - \hat{f}_d(x)$, *for the different values of* d,

iii. *the residuals* $Y_i - \hat{f}_d(X_i)$ *for the different values of* d.

You should have 3 figures with several functions in each figure. Use colors and a legend in each figure.

(d) *Now run the polynomial regression for* $d = 0, 1, 2, \ldots, 12$. *Evaluate the conditional MSE and the three terms from the decomposition* (1) *for each of the models at the point* $X^* = 1.5\pi$. *This involves repeatedly sampling* X *and* $Y|X$, *say 1000 times. Plot each term as a function of* d.

(e) *Interpret the figure and decomposition* (1) *in light of the "underfitting"/"overfitting" trade-off that we discussed in class. .*

3. *What would Breiman think of this exercise?*

Problem 2. *Write a function that draws 2 pseudo-random vectors* X_1, X_2 *of dimension* p. *They are to be i.i.d. with i.i.d. components that are distributed as standard normals. The function is to return the Euclidian distance between the two vectors. Execute this function 1000 times for* $p = 1, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ *and create a plot of the average distances as a function of* p. *What is this procedure estimating? Interpret the results in light of Tuesday's lecture.*

Problem 3. ***For graduate students only*** *Prove that the Bayes classifier solves the Bayes problem for binary labels* Y. *Bonus points if you have proved it for any discrete* Y.