# Machine Learning for Economic Analysis
# Problem Set 2

Jonas Lieber*

Due: 11:59pm Wed, Jan 31, 2023

**Problem 1.** *We reconsider the model of problem 1 of the first problem set*

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

*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* $\mu$ *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* $\mu$. *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.,*

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

*Recall the decomposition*

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

1. ***Graduate students only, all other questions are for all students.*** *Show the following formula for the fitted values of a linear regression with p parameters*

$$\frac{1}{n}\sum_{i=1}^{n} Cov(\hat{y}_i, y_i) = \frac{p}{n}\sigma^2.$$

2. *Where have you used knowledge of the true function* $f$ *in the last problem set in part d?*

3. *In this problem, we consider the more realistic case that* $f$ *is not known. Instead, you are given a dataset with i.i.d. draws* $(X_1, Y_1), \ldots (Y_n, X_n)$. *You can find the dataset on canvas as ps2data.csv.*

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

(a) *How would you do part d again without knowing the function $f$? Use the prediction point $x^* = 1.7$.*

(b) *Implement your procedure and try to reproduce the plot from part d without knowing $f$.*

4. *Implement a function for cross-validation. It should take as input*

   (a) *a dataset (of any dimension, including numbers of observations that are not a multiple of $K$)*

   (b) *$K$, a hyperparameter for cross-validation,*

   (c) *a function of the dataset and a set of hyperparameters (the function should be able to accept multiple sets of hyperparameters, possibly of different length).*

   *Do not use a function from a package, library etc. Code this function from scratch. Use this function to choose the best degree of a polynomial (between 0 and 30) that you can find using cross-validation.*

5. *How could you evaluate the conditional MSE for $X^* = 1.7$ for a model that you have selected with cross-validation? There should be two methods. Please describe which method you prefer and why.*

6. *Implement your preferred method to evaluate the conditional MSE for the chosen model for $X^* = 1.7$.*

7. ***Bonus**: Guess the function $f$.*

**Problem 2.** *Show that the formula for expected optimism shown in class also applies to 0-1 loss when $Y$ is binary.*

**Problem 3.** 1. *Generate a dataset of 1000 individuals, with potential outcomes $Y_i(d)$, a treatment $D_i$ that is not independent of the potential outcomes, and outcomes so that*

$$\frac{1}{n}\sum_{\substack{i=1\\ i:D_i=1}}^{n} Y_i - \frac{1}{n}\sum_{\substack{i=1\\ i:D_i=0}}^{n} Y_i \tag{2}$$

*does not approximate the ATE well.*

2. *Now consider the same potential outcomes and a treatment that is independent of the potential outcomes. Generate the observable outcomes. Does (2) estimate the ATE?*

**Problem 4.** ***graduate students only*** *Does the formula for expected optimism shown in class also apply when $Y$ can take on three values, say 0, 1 and 2?*