COMS 4771-1 F19 Homework 2 (due October 9, 2019)

Instructions

Submit your write-up on Gradescope as a neatly typeset (not scanned nor hand-written) PDF document by 2:30 PM of the due date. (Feel free to use Word, LaTeX, etc.—whatever you like.)

On Gradescope, be sure to select the pages containing your answer for each problem. More details can be found on the Gradescope Student Workflow help page.

(If you don't select pages containing your answer to a problem, you'll receive a zero for that problem.)

Make sure **your name and your UNI** appears prominently on the first page of your write-up.

You are welcome and encouraged to discuss homework assignments with each other in small groups (two to three people). You must **list all discussants in your homework write-up**: also do this prominently on the first page of your write-up.

Remember, discussion about homework assignments may include brainstorming and verbally discussing possible solution approaches, but **must not go as far as one person telling others how to solve a problem**. In addition, **you must write-up your solutions by yourself**, and **you may not look at another student's homework write-up/solutions (whether partial or complete)**. The academic rules of conduct can be found in the course syllabus.

Source code

Please combine all requested source code files into a **single** ZIP file, along with a plain text file called `README` that contains your name and briefly describes all of the other files in the ZIP file. **Do not include the data files.** Submit this ZIP file on Courseworks.

Clarity and precision

One of the goals in this class is for you to learn to reason about machine learning problems and algorithms. To demonstrate this reasoning, you must be able to make **clear** and **precise** arguments. A clear and precise argument is not the same as a long, excessively detailed argument. Unnecessary details and irrelevant side-remarks often make an argument less clear. Non-factual statements also detract from the clarity of an argument.

Points may be deducted for answers and arguments that lack sufficient clarity or precision. We will grade your answer/argument based only on a time-economical attempt to understand it.

Problem 1 (20 points)

In this problem, you will reason about optimal predictions for squared loss risk.

Suppose $Y_1, \ldots, Y_n, Y$ are iid random variables—the distribution of Y is unknown to you. You observe $Y_1, \ldots, Y_n$ as "training data" and must make a prediction of Y.

(a) Assume Y has a probability density function given by

$$p_\theta(y) := \begin{cases} \frac{1}{\theta^2} y e^{-y/\theta} & \text{if } y > 0, \\ 0 & \text{if } y \leq 0, \end{cases}$$

for some $\theta > 0$. Suppose that θ is known to you. What is the "optimal prediction" $\hat{y}^\star$ of Y that has the smallest mean squared error $\mathbb{E}[(\hat{y}^\star - Y)^2]$? And what is this smallest mean squared error? Your answers should be given in terms of θ.

(b) (Continuing from Part (a).) In reality, θ is unknown to you. Suppose you observe $(Y_1, \ldots, Y_n) = (y_1, \ldots, y_n)$ for some positive real numbers $y_1, \ldots, y_n > 0$. Derive the following:

- the MLE $\hat{\theta}(y_1, \ldots, y_n)$ of θ given this data;
- the prediction $\hat{y}(y_1, \ldots, y_n)$ of Y based on the plug-in principle (using $\hat{\theta}(y_1, \ldots, y_n)$).

Show the steps of your derivation. The MLE and prediction should be given as simple formulas involving $y_1, \ldots, y_n$.

(c) (Continuing from Part (a).) Let $\hat{Y} := \frac{1}{n}(Y_1 + \cdots + Y_n)$, so $\hat{Y}$ is a random variable that depends on $Y_1, \ldots, Y_n$. Prove that the mean squared error of $\hat{Y}$

$$\mathbb{E}[(\hat{Y} - Y)^2]$$

is exactly $(1 + \frac{1}{n})\mathbb{E}[(\hat{y}^\star - Y)^2]$, where $\hat{y}^\star$ is the "optimal prediction" from Part (a). The expectations above are taken with respect to all of the random variables $Y_1, \ldots, Y_n, Y$.

(d) Now, instead assume $Y \sim \text{Bern}(\theta)$ for some $\theta \in [0, 1]$. Suppose that θ is known to you. What is the prediction $\hat{y}^\star$ of Y that has the smallest mean squared error $\mathbb{E}[(\hat{y}^\star - Y)^2]$? And what is this smallest mean squared error? Your answers should be given in terms of θ.

(e) (Continuing from Part (d).) Define the following *loss function* $\ell \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ by

$$\ell(\hat{y}, y) := \begin{cases} 2(\hat{y} - y)^2 & \text{if } \hat{y} \geq y, \\ (\hat{y} - y)^2 & \text{if } \hat{y} < y. \end{cases}$$

This loss function is a different way to measure how "bad" a prediction is. With this loss function, a prediction that is too high is more costly than one that is too low. What is the prediction $\hat{y}^\star$ of Y that has the smallest expected loss $\mathbb{E}[\ell(\hat{y}^\star, Y)]$? And what is this smallest expected loss? Your answers should be given in terms of θ.

Problem 2 (20 points)

In this problem, you will use linear regression with a subset of ProPublica's COMPAS data set, and evaluate the disparate behavior of a derived "screening tool" on different subpopulations represented in the data.

COMPAS data set

Download the COMPAS data set from Courseworks (`compas-train.csv` and `compas-test.csv`, both in CSV format). This is a subset of the data analyzed by ProPublica for their Machine Bias article. These data represent criminal defendants from Broward County, Florida. For each defendant, the data contains information available to a "screening tool" that is intended to assess the defendant's risk of recidivism (i.e., risk of committing a crime if they were to be released on parole). The data also contains, for each defendant, whether the defendant was arrested again within two years of the assessment.[1]

The data has been randomly split into two parts: one for "training" and another for "testing". There is one row per defendant. The "label" is provided in the first column (named `two_year_recid`). The remaining eight columns correspond to features that are to be used to predict the label. Each feature is either binary-valued (`sex`, `race`, `c_charge_degree`) or integer-valued (`age`, `juv_fel_count`, `juv_misd_count`, `juv_other_count`, `priors_count`). (There are several other features in the original data set that we have omitted for this assignment. We have taken a subset of the original data that includes only two possible values for the `sex` attribute and two possible values for the `race` attribute.)

Linear regression and classification

Compute the affine function $\hat{\eta}\colon \mathbb{R}^d \to \mathbb{R}$ of smallest empirical (squared loss) risk on the training data. (You can do this using affine feature expansion and ordinary least squares.) What is the (squared loss) risk of $\hat{\eta}$ on the test data?

Since the label is binary-valued, we can regard $\hat{\eta}(\boldsymbol{x})$ as an estimate of the conditional probability that the label is one given the feature vector $\boldsymbol{x}$ (i.e., an estimate of $\Pr(Y = 1 \mid \boldsymbol{X} = \boldsymbol{x})$). Therefore, it is reasonable to derive a binary classifier $\hat{f}\colon \mathbb{R}^d \to \{0, 1\}$ from $\hat{\eta}$, defined by

$$\hat{f}(\boldsymbol{x}) = \mathbb{1}_{\{\hat{\eta}(\boldsymbol{x}) > 1/2\}}.$$

What are the error rate and false positive rate of $\hat{f}$ on the test data?

(A "false positive" is a prediction that a given defendant will be arrested within two years of screening, among defendants who do not actually get arrested within two years of screening.)

Performance on subpopulations

There are many natural subpopulations represented in the data. For example, we may define two subpopulations based on the value of the `sex` feature, and we may also define two subpopulations based on the value of the `race` feature. ProPublica's article was largely concerned with the disparate behavior of a "screening tool" called "COMPAS" on two subpopulations based on the `race` feature.

[1] Note that being arrested is not the same as having committed a crime. See the recent article by Lum and Shah, 2019 for further discussion in a related context.

Choose one of the features `sex` and `race`; call this feature A. For each $a \in \{0, 1\}$, what are the error rate and false positive rate of $\hat{f}$, as evaluated on the subset of test data with $A = a$?

You should find that, like the COMPAS screening tool studied by ProPublica, the function $\hat{f}$ also has disparate behavior on the two subpopulations: the false positive rate is much higher for one subpopulation than it is for the other.

Inherent limitations?

Is the disparate behavior on the two subpopulations inevitable? For each $a \in \{0, 1\}$, repeat the training and testing above using only training and test data with $A = a$. That is, compute the affine function $\hat{\eta}_a \colon \mathbb{R}^d \to \mathbb{R}$ of smallest empirical (squared loss) risk on the training data with $A = a$; form the corresponding classifier $\hat{f}_a \colon \mathbb{R}^d \to \{0, 1\}$ via $\hat{f}_a(\boldsymbol{x}) = \mathbb{1}_{\{\hat{\eta}_a(\boldsymbol{x})\}}$; then compute the error rate and false positive rate of $\hat{f}_a$ on the test data with $A = a$.

Do you still observe disparate behavior on the two subpopulations? What are the implications of this finding? (Think about the relative utility of the features for predicting the label.)

What to submit in your write-up

(a) Please indicate which feature you have selected to be A (either `sex` or `race`).
(b) Squared loss risk of $\hat{\eta}$ on test data.
(c) Error rate of $\hat{f}$ on test data.
(d) False positive rate of $\hat{f}$ on test data.
(e) The performance measures for $\hat{f}$ on the test data with $A = 0$.
(f) The performance measures for $\hat{f}$ on the test data with $A = 1$.
(g) The performance measures for $\hat{f}_0$ on the test data with $A = 0$.
(h) The performance measures for $\hat{f}_1$ on the test data with $A = 1$.
(i) (Optional.) Discuss the implications of your observations concerning the disparate behavior of $\hat{f}_0$ and $\hat{f}_1$ (or lack thereof) on their respective subpopulations.

Please also submit your source code on Courseworks. If you use a Jupyter notebook, please extract the Python source code and include a single `.py` file with this Python code. You are welcome to use any software packages you like for this problem, provided that you cite these software packages in your write-up.

Problem 3 (20 points)

In this problem, you will visualize the variability of fits with linear regression.

49 data sets

Download the synthetic data sets from Courseworks (`hw2p3_xtrain.csv`, `hw2p3_ytrain.csv`) for this problem. These two files, respectively, provide the x- and y-parts of 49 different data sets. The data sets have been randomly and independently sampled from the same probability distribution. For each $t \in \{1, \ldots, 49\}$, the t-th data set is a sample of $n = 10$ pairs $(x_{1,t}, y_{1,t}), \ldots, (x_{n,t}, y_{n,t}) \in \mathbb{R} \times \mathbb{R}$. Also get the test data (`hw2p3_xtest.csv`), a set of evenly-spaced points between -10 and 10.

Fit each data set

For each $t \in \{1, \ldots, 49\}$, compute the affine function $\hat{\eta}_{t,\text{affine}} \colon \mathbb{R} \to \mathbb{R}$ of smallest empirical (squared loss) risk on the t-th data set. (You can do this using affine feature expansion and ordinary least squares.) The functions $\hat{f}_{1,\text{affine}}, \ldots, \hat{f}_{49,\text{affine}}$ you should obtain are depicted in Figure 1.

Plot all of the curves $\hat{\eta}_{t,\text{affine}}(x)$ for $-10 \leq x \leq 10$ (using the test data) *on the same x-y plane*. So, you should have a figure that has 49 curves overlaid on top of each other. (It is fine to use the same line style and color for all 49 curves.) This figure will show the variability of the learned affine functions across different instantiations of the training data.

Now, in the same figure, plot the "average curve" $\frac{1}{49} \sum_{t=1}^{49} \hat{\eta}_{t,\text{affine}}(x)$ for $-10 \leq x \leq 10$ (again, using the test data). Change the line style and color so that it is very visible even on top of the previous 49 curves. This curve shows (an estimate of) the expected value of $\hat{\eta}_{t,\text{affine}}(x)$, where the expectation is taken over the random draw of the training data.

Set the axes of the plot to show $-10 \leq x \leq 10$ and $-10 \leq y \leq 10$. Make sure the axes and curves are clearly labeled.

Quadratic and cubic functions

Repeat the entire process above two more times.

- Instead of affine functions, use quadratic functions. This means you should use linear regression with the feature map $\phi(x) := (1, x, x^2)$.
- Instead of affine functions, use cubic functions. This means you should use linear regression with the feature map $\phi(x) := (1, x, x^2, x^3)$.

(Figure 2 and Figure 3 are analogous to Figure 1 for the quadratic and cubic cases.)

Compare and contrast the results in the three cases (affine, quadratic, cubic). Is there something peculiar about the results in the cubic case (e.g., in the range $6 \leq x \leq 10$)? Try to explain it.

What to submit in your write-up

(a) Figure for affine case.
(b) Figure for quadratic case.
(c) Figure for cubic case.
(d) A brief paragraph comparing and contrasting the results.

No need to submit source code for this problem.

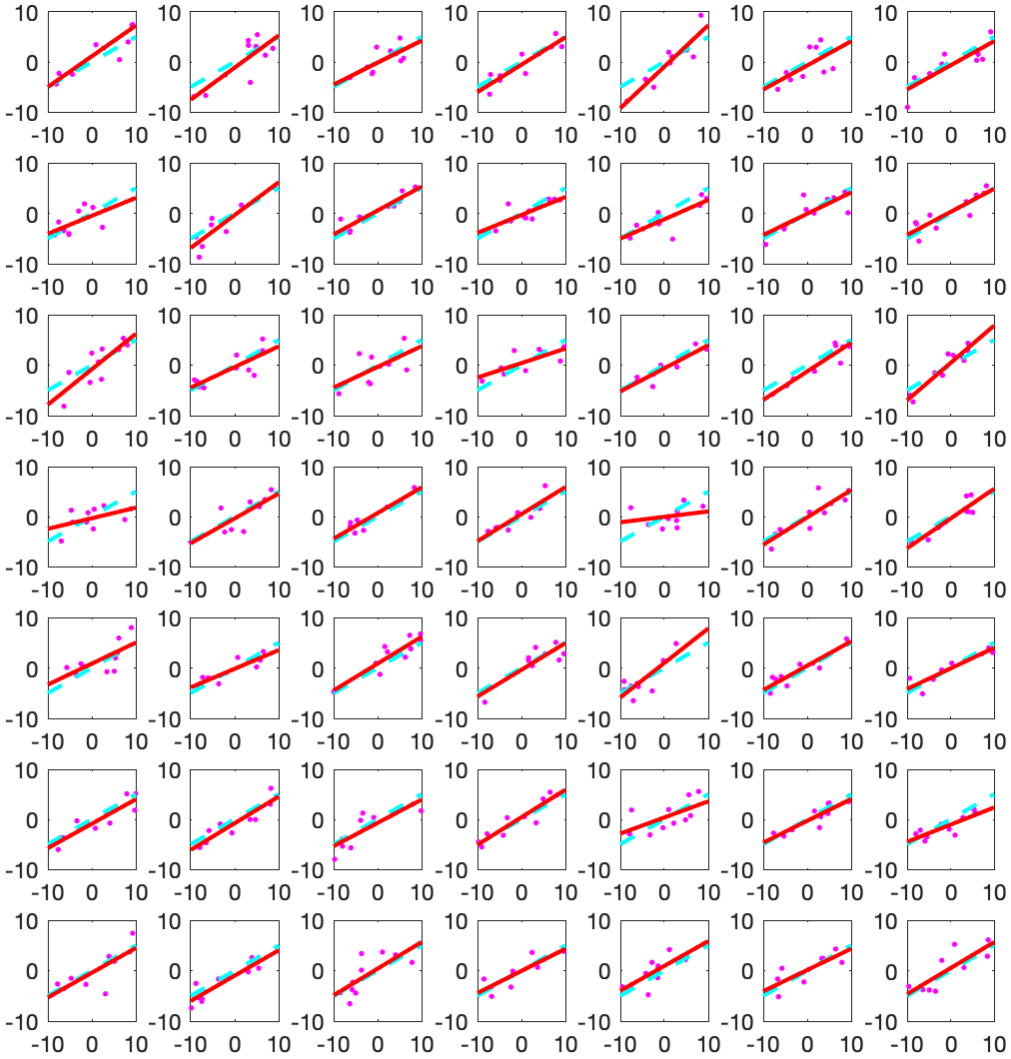

Figure 1: Learned affine functions on 49 data sets. The magenta points are the training data. The cyan dashed line shows the true regression function. The learned affine functions are shown in red.

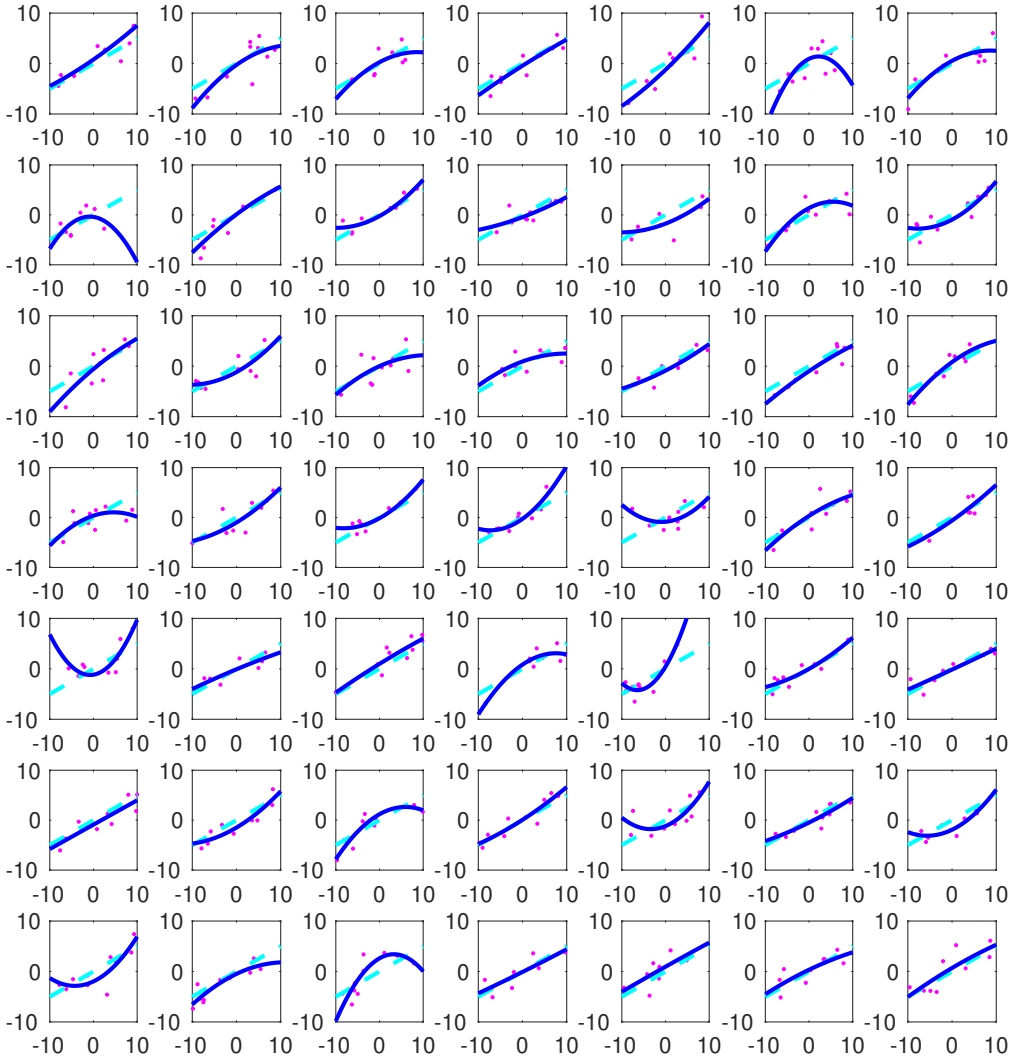

Figure 2: Learned quadratic functions on 49 data sets. The magenta points are the training data. The cyan dashed line shows the true regression function. The learned quadratic functions are shown in blue.

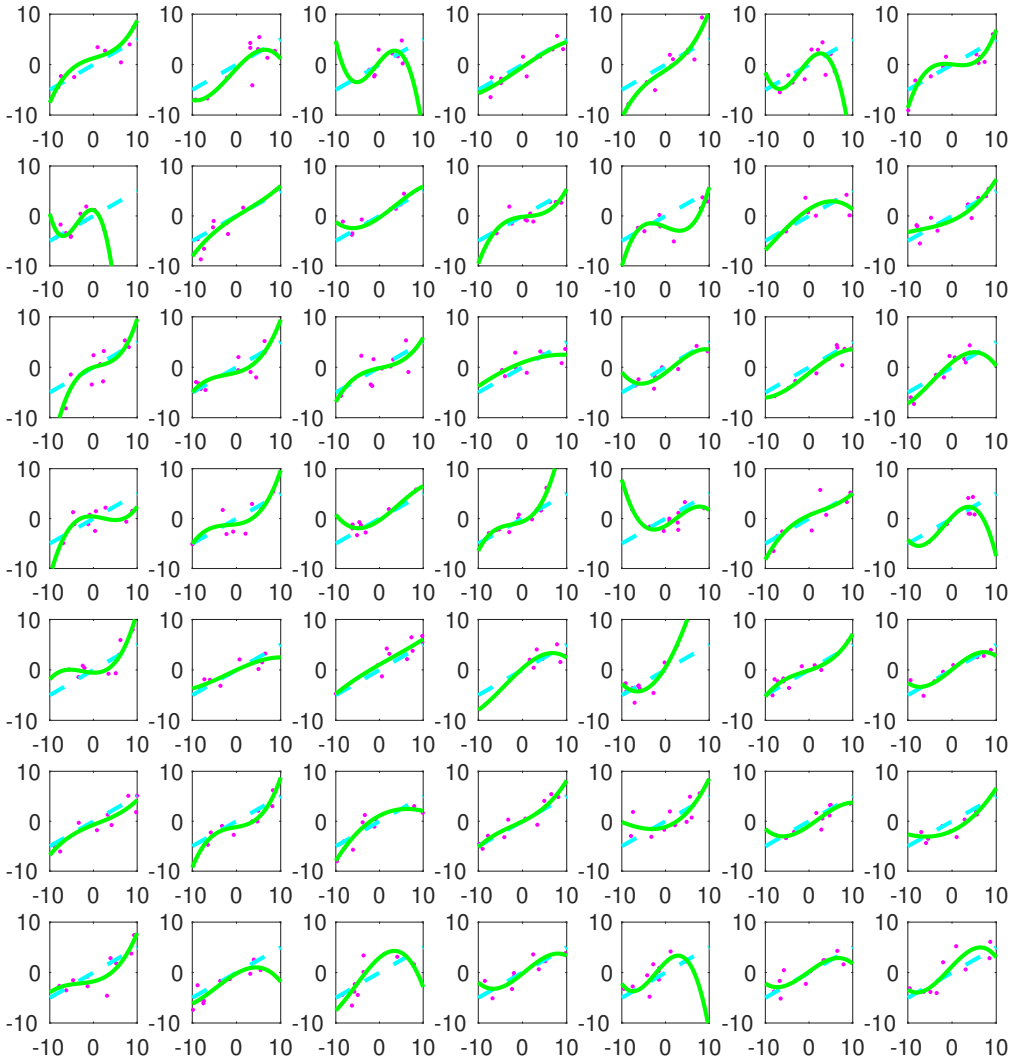

Figure 3: Learned cubic functions on 49 data sets. The magenta points are the training data. The cyan dashed line shows the true regression function. The learned cubic functions are shown in green.

Problem 4 (20 points)

In this problem, you will prove a basic relationship between the empirical risk and true risk of an empirical risk minimizer.

Let $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n), (\boldsymbol{X}, Y)$ be iid random examples drawn from a probability distribution P over $\mathbb{R}^d \times \mathbb{R}$. Let $\mathcal{R}$ denote the true (squared loss) risk with respect to P (i.e., $\mathcal{R}(\boldsymbol{w}) = \mathbb{E}[(\boldsymbol{X}^\intercal \boldsymbol{w} - Y)^2]$), and let $\widehat{\mathcal{R}}$ denote the empirical (squared loss) risk based on $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_n, Y_n)$ (i.e., $\widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{X}_i^\intercal \boldsymbol{w} - Y_i)^2$).

Let $\hat{\boldsymbol{w}}$ be a linear function of smallest empirical risk $\widehat{\mathcal{R}}$. Prove that

$$\mathbb{E}\left[\widehat{\mathcal{R}}(\hat{\boldsymbol{w}})\right] \leq \mathbb{E}\left[\mathcal{R}(\hat{\boldsymbol{w}})\right].$$

Problem 5 (20 points)

In this problem, you will modify the basic ERM approach to handle non-representative training data.

Download the synthetic data sets from Courseworks (`hw2p5_train.csv`, `hw2p5_test.csv`) for this problem. The first column represents the label, and the second column represents the (scalar) input feature. Both data sets have $n = 1000$ data points.

Empirical risk minimization and outliers

Compute the affine function $\hat{f}\colon \mathbb{R} \to \mathbb{R}$ of smallest empirical (squared loss) risk on the training data. What is its test risk (i.e., empirical risk on the test data)?

Also, plot $|\hat{f}(\boldsymbol{x}_i) - y_i|$ as a function of i, where $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ are the training examples. The training examples are ordered in a particular way (see below), so indeed we want to see this plot with the example index i as the horizontal axis. This plot gives an indication about which training examples might be considered "outliers" based on their fit. A method like "iterative trimming" would throw out the examples that are least well-fit.

(Don't actually throw away any training examples!)

Subpopulations

Suppose you learn that the training data was obtained from two subpopulations, but one subpopulation is underrepresented. In the true population, the subpopulations are equally represented. But in the training data, the first 900 data correspond to subpopulation 1, and the last 100 data correspond to subpopulation 2. (Hopefully you see the relevance to the "outlier-ness" plot.)

Recall that ERM is based on the plug-in principle, where the empirical distribution P_n is plugged in for the true (unknown) data distribution. However, we know now that the empirical distribution P_n based on the training data is not a good substitute for the true data distribution.

Let Q_n be the distribution that puts probability mass $\frac{1}{2 \cdot 900}$ on the first 900 training examples, and puts probability mass $\frac{1}{2 \cdot 100}$ on the last 100 training examples. Under the distribution Q_n, the first 900 training examples have the same overall probability mass as the last 100 training examples. We apply plug-in principle with Q_n and thus seek to minimize the risk under Q_n:

$$\widehat{\mathcal{R}}_{Q_n}(g) := \frac{1}{2 \cdot 900} \sum_{i=1}^{900} (g(\boldsymbol{x}_i) - y_i)^2 + \frac{1}{2 \cdot 100} \sum_{i=901}^{1000} (g(\boldsymbol{x}_i) - y_i)^2.$$

(Above, $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_{900}, y_{900})$ are the first 900 training examples, and $(\boldsymbol{x}_{901}, y_{901}), \ldots, (\boldsymbol{x}_{1000}, y_{1000})$ are the last 100 training examples.)

Explain how to find the affine function of smallest $\widehat{\mathcal{R}}_{Q_n}$ value by solving a system of linear equations that are just a slight variation of the normal equations.

Compute the affine function $\hat{g}\colon \mathbb{R} \to \mathbb{R}$ of smallest $\widehat{\mathcal{R}}_{Q_n}$ value. What is its test risk (i.e., empirical risk on the test data)?

In the test data, the first 500 test examples are from subpopulation 1, and the last 500 test examples are from subpopulation 2. What are the test risks of $\hat{f}$ and $\hat{g}$ on each subpopulation?

Separating subpopulations

Another approach to dealing with subpopulations in data is to separately fit a model to each subpopulation. This, of course, requires one to know which subpopulation an input point belongs to. We are fortunate that this is the case for this problem.

Compute the affine function $\hat{h}_1 \colon \mathbb{R} \to \mathbb{R}$ of smallest empirical risk on the training data for subpopulation 1 (i.e., just the first 900 training examples). Then compute the affine function $\hat{h}_2 \colon \mathbb{R} \to \mathbb{R}$ of smallest empirical risk on the training data for subpopulation 2 (i.e., just the last 100 training examples). What is the test risk of $\hat{h}_1$ on subpopulation 1, and what is the test risk of $\hat{h}_2$ on subpopulation 2?

What to submit in your write-up

(a) Test risk of $\hat{f}$ and plot of "outlier-ness".
(b) Description of approach for finding $\hat{g}$.
(c) Test risk of $\hat{g}$.
(d) Test risks of $\hat{f}$ and $\hat{g}$ on the two subpopulations. (Four test risks to compute.)
(e) Test risks of $\hat{h}_1$ and $\hat{h}_2$ on their respective subpopulations.

No need to submit source code for this problem.