COMS 4771-2 F18 Homework 2 (due October 10, 2018)

Instructions

Submit your write-up on Gradescope as a neatly typeset (not scanned nor handwritten) PDF document by 11:59 PM of the due date.

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:

- https://gradescope.com/help#help-center-section-student-workflow

(If you don't select the 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.

Source code

Please combine all requested source code files into a *single* ZIP file[1], 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 reason about these things, you must be able to make *clear* and *precise* claims and arguments about them.

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. And 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. Moreover, a time-economical attempt will be made to interpret such answers/arguments, and the grade you receive will be based on this interpretation.

[1] See https://en.wikipedia.org/wiki/Zip_(file_format).

Problem 1 (30 points)

In this problem, you will practice using linear regression on a simple data set.

Download the wine data set `wine.mat` from the course website. The first feature (given in `data`) is a constant feature, always equal to one; I have already applied "affine expansion" to the data. The next $d = 11$ features are `fixed acidity`, `volatile acidity`, ..., `alcohol` as described in http://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality.names. I have applied a "standardization" transformation to these features so that, on the training data, the empirical mean of each feature is zero, and the empirical standard deviation of each feature is one; the same transformation was applied to the test data. The output (given in `labels`) is `quality`, the quality grade of the wine (an integer between 0 and 10).

Ordinary least squares

Compute the ordinary least squares estimator based on the training data (i.e., the squared loss ERM). You can use any software package you like to do this provided that you include proper citations.[2] **Make sure the solution satisfies the appropriate normal equations! This is a way to check if you've done things correctly.** Compute the test squared loss risk on the test data (`testdata` and `testlabels`).

Sparse linear predictor

Write a program to find a coefficient vector $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_d)$ where at most three of the $\hat{\beta}_i$ for $1 \leq i \leq d$ are non-zero, with smallest empirical risk on the training data. (The coefficient $\hat{\beta}_0$ corresponds to the always-one feature and is also allowed to be non-zero.) This can be done by enumerating over all $\binom{d}{3}$ triplets of the d features. For this chosen coefficient vector, compute the test squared loss risk on the test data.

For each of the (at most) three variables $v \in \{\texttt{fixed acidity}, \texttt{volatile acidity}, \ldots, \texttt{alcohol}\}$ with non-zero coefficients in $\hat{\beta}$, determine the two variables (different from v) with highest and second-highest (in absolute value) sample Pearson correlation with variable v, as based on the test data. Record both the variable names and the corresponding correlation values.

What to submit in your write-up:

(a) Test risks of the ordinary least squares estimator and of the sparse linear predictor.
(b) Names (e.g., `fixed acidity`, `volatile acidity`) of the variables with non-zero coefficients in the sparse linear predictor, along with the coefficient values. Also give the value of the coefficient $\hat{\beta}_0$.
(c) For each variable v with non-zero coefficient, the names of the two other variables most correlated (in absolute value) with variable v, along with the corresponding correlation values.

Please submit your source code on Courseworks.

[2] See https://libguides.mit.edu/c.php?g=551454&p=3900280.

Problem 2 (35 points)

In this problem, you will study an "approximation error-vs-variance" trade-off in the *fixed design* setting (which was covered in the reading assignment).

Suppose you are given n distinct real numbers $z_1, \ldots, z_n \in [-\pi, \pi]$, and you observe the realizations of n uncorrelated real-valued random variables $Y_1, \ldots, Y_n$, where $\mathbb{E}(Y_i) = h(z_i)$ for some unknown continuous function $h\colon \mathbb{R} \to \mathbb{R}$, and $\mathrm{var}(Y_i) = 1$ for all $i = 1, \ldots, n$. Taking inspiration from the Weierstrass approximation theorem, you construct a feature vector for each z_i using a degree-k polynomial expansion (for some $k \in \mathbb{N}$),

$$x_i := (1, z_i, z_i^2, \ldots, z_i^k) \quad \text{for each } i = 1, \ldots, n,$$

and use ordinary least squares to construct $\hat{\beta} \in \mathbb{R}^{k+1}$ to approximate each $h(z_i)$ with $x_i^\top \hat{\beta}$.

(a) Suppose it turns out that

$$h(z_i) = \cos(z_i) \quad \text{for all } i = 1, \ldots, n.$$

Derive an upper-bound on the expected (fixed design) risk

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(x_i^\top \hat{\beta} - \mathbb{E}(Y_i))^2\right].$$

The bound should be expressed as a simple function of k and n. Give detailed justifications for each step in your derivation. (It is fine to use asymptotic notations, like $O(1/n)$.)

Hint: Use Taylor's (remainder) theorem.

(b) If k is chosen (as a function of n) to make the expected risk from part (a) as small as possible, what is this best possible expected risk? Your answer should be expressed only in terms of n. Briefly justify your answer. (It is fine to use asymptotic notations, like $O(1/n)$.)

(c) Let $n = 1001$, and let $z_1, \ldots, z_n$ be uniformly spaced in the interval $[-\pi, \pi]$ (with $z_1 = -\pi$ and $z_n = \pi$). Simulate the random variables $Y_1, \ldots, Y_n$, where $Y_i = \sin(3z_i/2) + \varepsilon_i$ for each i, and $\varepsilon_1, \ldots, \varepsilon_n$ are iid $\mathrm{N}(0, 1)$-random variables. For each $k = 1, \ldots, 20$, compute the predictions of $\mathbb{E}(Y_i)$ obtained using ordinary least squares on the degree-k polynomial expansion, along with the fixed design risk

$$\frac{1}{n}\sum_{i=1}^{n}(x_i^\top \hat{\beta} - \mathbb{E}(Y_i))^2.$$

Repeat this process 1000 times, and average the 1000 resulting fixed design risks (to estimate the expected fixed design risk) for each k. Report these average fixed design risks for each k, both in a table and in a plot. (Make sure the table and plot are clearly labeled.)

Please submit your source code on Courseworks.

Problem 3 (35 points)

In this problem, you will carry out a simulation that illustrates a phenomenon called *Freedman's paradox*.

Download the data set `freedman.mat` from the course website. The $n \times d$ matrix in `data` contains n feature vectors $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^d$ (as rows), and the $n \times 1$ vector in `labels` contains the corresponding labels $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}$. Note that $n < d$, so we don't expect ordinary least squares to work well.

Note that the features and labels are approximately *standardized* (i.e., mean zero and variance one).

Consider the following three-step procedure:

1. Compute the following estimates of the correlations between the features and the label:

$$\hat{\rho}_j := \frac{1}{n} \sum_{i=1}^{n} x_j^{(i)} y^{(i)} \quad \text{for } j = 1, \ldots, d.$$

2. Let $\widehat{J} := \{j \in \{1, \ldots, d\} : |\hat{\rho}_j| > 1.75/\sqrt{n}\}$ be the features for which the estimated correlation is larger than $1.75/\sqrt{n}$, and discard the features not in $\widehat{J}$.
3. Construct the ordinary least squares estimate $\hat{\beta}(\widehat{J}) \in \mathbb{R}^d$ using only features in $\widehat{J}$. In other words,

$$\hat{\beta}(\widehat{J}) \in \underset{\substack{\beta \in \mathbb{R}^d: \\ \beta_j = 0 \text{ for all } j \notin \widehat{J}}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} (\beta^\top x^{(i)} - y^{(i)})^2.$$

Steps 1 & 2 comprise a *screening procedure* whose purpose is to remove irrelevant features. If the number of features $k := |\widehat{J}|$ remaining after the screening is much smaller than n, then one might think that ordinary least squares (using just the features in $\widehat{J}$) will work well.

(a) How many features remain (in $\widehat{J}$) after the screening? (It should be much smaller than n.)

(b) What is the empirical risk of $\hat{\beta}(\widehat{J})$, i.e.,

$$\frac{1}{n} \sum_{i=1}^{n} (\hat{\beta}(\widehat{J})^\top x^{(i)} - y^{(i)})^2?$$

(It should be much smaller than the empirical variance of the labels.)

(c) Construct $\hat{\beta}(I)$ in the same manner as $\hat{\beta}(\widehat{J})$ in Step 3 above, except using $I := \{1, \ldots, 75\}$ in place of $\widehat{J}$. What is the empirical risk of $\hat{\beta}(I)$? (It should be higher than what you got in part (b).)

(d) It turns out that the features and labels are drawn independently from $\mathrm{N}(0, 1)$—i.e., there is no real correlation between the features and labels. Yet based on the relatively small empirical risk of $\hat{\beta}(\widehat{J})$, together with the small number of features used in $\hat{\beta}(\widehat{J})$ relative to the sample size n, you might have thought that there is a relationship between the features and labels! If this seems unbelievable to you, write some code to generate the data yourself, and vary the numerical parameters of the simulation (e.g., n and d) to see how robust the phenomenon is to changes to these parameters. Try to explain why this happens, *in your own words*.

(e) Suppose you were able to carry out Step 3 using a new, independent but identically distributed set of data. What would you expect $\hat{\beta}(\widehat{J})$ to be? And what would you expect its empirical risk (on this new data set) to be? Approximate answers are fine, but briefly justify your answers.

Please submit your source code on Courseworks.