BUSI97287: Advanced Machine Learning Summer 2021 Imperial College Business School
Martin Haugh

Assignment 2

Due: 11.59pm Monday 17^{th} May 2021

Rules

1. This is a group assignment. (There are approximately 3 people per group and by now you should know your assigned group.)
2. While `R` is the default package / programming language for this course you are free to use `R` or `Python` for the programming components of this assignment.
3. Within each group **I strongly encourage each person to attempt each question by his / herself first** before discussing it with other members of the group.
4. Students should **not** consult students in other groups when working on their assignments.
5. Late assignments will **not** be accepted and all assignments must be submitted through the Hub with one assignment submission per group. Your submission should include a PDF report with your answers to each question as well as any relevant code. Make sure your PDF clearly identifies each member of the group by CID and name.

---

1. **Large-Scale Ridge Regression (30 marks)**
The ridge regression problem is given by

$$\min_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\},$$

where $\boldsymbol{\beta} \in \mathbb{R}^{d+1}$,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} \quad \text{and} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \dots & x_{1d} \\ 1 & x_{21} & x_{22} & \dots & x_{2d} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{N1} & X_{N2} & \dots & x_{Nd} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{x}_1^\top \\ 1 & \mathbf{x}_2^\top \\ \vdots & \vdots \\ 1 & \mathbf{x}_N^\top \end{bmatrix}.$$

The optimal solution of this problem is given by

$$\boldsymbol{\beta}^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

(In reality we don't shrink the constant coefficient β_0 in ridge regression (or Lasso) but to simplify notation here we assume that every coefficient is being shrunk.)

(a) Suppose the dimension d of the data points is not large, but the number of data points N is *extremely* large. In fact N is so large that the matrix $\mathbf{X} \in \mathbb{R}^{N \times (d+1)}$ does not even fit in memory. All of the data, however, is available in some database that can be queried

for each data point. Explain how you could still compute the ridge estimate $\boldsymbol{\beta}^*$.
(10 marks)

Hint: $\mathbf{X}^\top\mathbf{X}$ cannot be computed directly since $\mathbf{X}$ is too large to store in memory. But perhaps $\mathbf{X}^\top\mathbf{X}$ can be written as a sum of the form $\sum_{i=1}^N \mathbf{z}_i^\top \mathbf{z}_i$ where $\mathbf{z}_i$ is a $1 \times (d+1)$ vector ...?

Solution: The issue is that we cannot load the entire matrix $\mathbf{X}$ in memory, so computing $\mathbf{X}^\top\mathbf{X}$ directly is impossible. To overcome this we can make use of the fact that $\mathbf{X}^\top\mathbf{X} = \sum_{i=1}^N \mathbf{z}_i^\top \mathbf{z}_i$, where $\mathbf{z}_i = \begin{bmatrix} 1 & \mathbf{x}_i^\top \end{bmatrix}$. Similarly, $\mathbf{X}^\top\mathbf{y} = \sum_{i=1}^N \mathbf{z}_i^\top y_i$. This leads to a simple procedure for computing $\boldsymbol{\beta}^*$: start with $\mathbf{E} := \lambda\mathbf{I}, \mathbf{f} := \mathbf{0}$ (the d-dimensional zero vector). Then for $i = 1, \ldots, N$ do $\mathbf{E} := \mathbf{E} + \mathbf{z}_i^\top \mathbf{z}_i, \mathbf{f} := \mathbf{f} + \mathbf{z}_i^\top y_i$, and finally compute $\boldsymbol{\beta}^* := \mathbf{E}^{-1}\mathbf{f}$. Note that using this procedure we only need to load one $\mathbf{z}$ vector at a time.

(b) Suppose the data for the ridge regression problem becomes available sequentially, i.e. the k^{th} data point $\mathbf{x}_k$ arrives at time t_k. At time t_k we want to be able to compute the optimal ridge estimate $\boldsymbol{\beta}_k^*$ using all the previous data $\{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$. In the usual method for ridge regression, we will have to store all the previous data points in a database, create the data matrix, and compute the estimate. Thus, the memory requirements will grow over time. Explain how you could still compute $\boldsymbol{\beta}_k^*$ for all $k \geq 1$, while keeping only $\mathcal{O}(d^2)$ numbers in the database. **(10 marks)**

Remark: One application where the data becomes available sequentially is in *A/B testing* with so-called *contextual bandit* problems. For example, an e-commerce company may be testing various web-site designs with the goal of maximizing online sales or revenue. Rather than using just one web-site design they may allow the displayed web-site to depend on the (arriving) customer via the customer's feature vector $\mathbf{x}$. Specifically, the t^{th} customer that arrives to the company's site has a known feature vector $\mathbf{x}_t$ and is shown a different version of the company's web-site according to the A/B testing algorithm. The customer yields an outcome vector y_t which, for example, might be the dollar sales of the customer on the displayed web-site. The goal is to estimate $\mathrm{E}[y \mid \mathbf{x}, \text{design i}]$ for $i = 1, \ldots, n$ where n is the number of web-site displays under consideration. Ultimately the company wants to display $\max_i \mathrm{E}[y \mid \mathbf{x}, \text{design i}]$ to the customer with feature vector $\mathbf{x}$ but to do this it needs to learn the best one. The best bandit algorithms do this by finding a good tradeoff between *exploration* (showing customers potentially sub-optimal web-sites with a view to better learn their expected values) and *exploitation* (simply using the web-site i that currently has the highest estimated value for customer $\mathbf{x}$). A/B testing is now a standard tool for e-commerce companies of all sizes and it's not uncommon for companies to run thousands of A/B tests per year.

Solution: The solution to part (a) works here. We need to store the matrix $\mathbf{E}$ that has $(d+1)^2$ elements and the vector $\mathbf{f}$ that has $d+1$ elements. When the n^{th} data point arrives, we update $\mathbf{E} = \mathbf{E} + \mathbf{z}_n^\top \mathbf{z}_n$ and $\mathbf{f} = \mathbf{f} + \mathbf{z}_n^\top y_n$, and compute $\boldsymbol{\beta}_n^* = \mathbf{E}^{-1}\mathbf{f}$.

(c) Inverting a $(d+1) \times (d+1)$ matrix (like $(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})$) takes $\mathcal{O}(d^3)$ time so computing $\boldsymbol{\beta}_n^*$ (given a new observation $\mathbf{x}_n$) in part (b) is computationally expensive. Can you find a way of finding $\boldsymbol{\beta}_n^*$ that takes only $\mathcal{O}(d^2)$ time? **(10 marks)**

Hint: Use the Sherman-Morrison-Woodbury identity $(\mathbf{A} + \mathbf{u}\mathbf{v}^\top)^{-1} = \mathbf{A}^{-1} - \frac{\mathbf{A}^{-1}\mathbf{u}\mathbf{v}^\top \mathbf{A}^{-1}}{1+\mathbf{v}^\top \mathbf{A}^{-1}\mathbf{u}}$.

Solution: Instead of storing of $\mathbf{E}$ at each time step, we could store $\mathbf{E}^{-1}$. Then when a new data-point $\mathbf{x}_n$ arrives we perform the update

$$\mathbf{E}^{-1} := (\mathbf{E} + \mathbf{z}_n^\top \mathbf{z}_n)^{-1} = \mathbf{E}^{-1} - \frac{(\mathbf{E}^{-1}\mathbf{z}_n^\top)(\mathbf{E}^{-1}\mathbf{z}_n^\top)^\top}{1 + \mathbf{z}_n \mathbf{E}^{-1} \mathbf{z}_n^\top} \tag{1}$$

where the second equality follows from the Sherman-Morrison-Woodbury identity. Thus, the complexity of updating $\mathbf{E}^{-1}$ is only $\mathcal{O}(d^2)$ since all the matrix-vector multiplications on the r.h.s. of (1) take only $\mathcal{O}(d^2)$ time. Computing $\boldsymbol{\beta}_n^* = \mathbf{E}^{-1}\mathbf{f}$ has complexity $\mathcal{O}(d^2)$. This method is clearly superior when d is very large.

---

2. **Estimating the Scofflaw Rate in Chicago (30 marks)**

A city[1] that has parking meters on many of its streets wanted to estimate how much money it was losing because people were not paying the meters. Ideally, it would select a random sample of cars parked at meters at random times, and determine for each whether the meter read "violation" at the moment the car was approached. Then it would estimate p, the proportion of parked cars in violation of the meter, by $\hat{p}$, the fraction of such cars within the random sample.

However, such a sampling scheme was not practical. Even if the city knew exactly which meters had cars parked at them, a random selection among those cars might first pick a vehicle nine miles North of the city center, and then pick a second vehicle three miles South. Moving among the vehicles thus sampled would be hopelessly inefficient and expensive. Another scheme was necessary.

What the city decided to do instead was to pick metered blocks at random (e.g., 33rd Street between 3rd and 4th Avenues). Then, at a random time during a given interval (e.g., noon - 1 PM), a surveyor would arrive at each such block and count both the number of cars parked there and the number of cars in violation of the meters. If a total of n blocks were sampled, and z_i was the number of cars parked on block i and and x_i the number of "violation" cars there, then the overall "scofflaw rate" p would be estimated by $\tilde{p}$ that follows:

$$\tilde{p} = \frac{\sum_{i=1}^n x_i}{\sum_{i=1}^n z_i}. \tag{2}$$

(a) In words, what approximation is being made in (2)? **(5 marks)**

[1] This question (and the next one) is taken from an excellent new textbook *Applied Statistics: Models and Intuition* (2015, Dynamic Ideas) by Arnold I Barnett. It provides an entertaining yet lucid introduction to statistics with real-world applications. If you're interested in brushing up on your (classical) statistics and want to gain some real insight into how to think about real-world problems then this is a perfect book for you.

Solution: $\sum_{i=1}^n x_i$ is the total number of cars that are in violation on the n blocks sampled, while $\sum_{i=1}^n z_i$ is the total number of sampled cars. Therefore $\tilde{p}$ is the fraction of observed cars that were in violation.

(b) Why does that approximation seem a reasonable one? **(5 marks)**

Solution: The approximation seems reasonable for a few reasons. First, we note that if we had been able to sample all blocks in the city then we would have $\tilde{p} = p$, as desired. Second, while the scheme did not sample cars randomly (for the practical reasons given above) it adjusts for this fact in the construction of $\tilde{p}$ where more attention is paid to blocks with many parked cars than to blocks with few. We can make this clear with a simple example. Suppose that two blocks are sampled, block A with two parked cars and one scofflaw (i.e., in violation), while block B has 16 parked cars and two scofflaws. Then the scofflaw rate was 50% on block A and 12.5% on block B, but

$$\tilde{p} = \frac{1+2}{2+16} = \frac{3}{18} = 16.7\%$$

which is considerably less than the average of 50% and 12.5% which is 31.25%. The busier street therefore got more weight in the calculation which is seems to be a strength of the calculation in (2).

(c) In working out the margin of error, would it be reasonable to apply the usual formulas as if $\tilde{p}$ were estimated by the fraction of violations within a purely random sample of $\sum_{i=1}^n z_i$ different cars? (In this case the "usual formulas" would say that an approximate 95% confidence interval (CI) for p is given by

$$\tilde{p} \pm 1.96 \times \sqrt{\frac{\tilde{p}(1-\tilde{p})}{\sum_{i=1}^n z_i}} \tag{3}$$

for example.) Justify your answer. **(10 marks)**

Solution: Using a formula like (3) would be incorrect since it assumes that the scofflaw "status" of different cars are independent. But that need not be the case. For example, in some parts of the city scofflaw rates might be high whereas they might be low in other parts. That would imply the scofflaw "status" of different cars that were sampled from the same block would be positively correlated and not independent. As with the previous question, it is easy to construct an extreme yet simple example of this. Suppose that there are only 2 categories of blocks: those where everyone pays the parking meter and those where no-one pays. Then if a block is sampled and the first car is a scofflaw we know for sure the second car on the block (and the others) will also be a scofflaw. So the independence assumption required for (3) does not hold under the sampling scheme here and so (3) would be problematic.

(d) Is $\tilde{p}$ an unbiased estimator for p? Justify your answer by either proving that it is unbiased or come up with a counter-example which shows that $\tilde{p}$ is biased. (*Hint:* Imagine a city that has three blocks with parking meters, two of which will be sampled at random. Can you find a set of (x_i, z_i)'s for which $\mathbb{E}[\tilde{p}] \neq p$? The example can be simple.) **(10 marks)**

Solution: $\tilde{p}$ is a biased estimator. (But of course as you should know by now, bias isn't necessarily a bad thing!) To see this, consider the following example. Imagine a city that has three blocks with parking meters, two of which will be sampled at random. The first block has 4 cars parked on it and all of them are scofflaws. The second and third blocks each have 2 cars parked on them and all of these cares have to paid to park. It then follows that the true value of p is

$$p = \frac{\text{total \# of scofflaws}}{\text{total \# of cars}} = \frac{4}{8} = \frac{1}{2}.$$

But the scheme that samples two blocks at random will result in three possible values of $\tilde{p}$. Specially, we will have

$$\tilde{p} = \begin{cases} \frac{0}{4}, & \text{if the sample omits block 1} \\ \frac{4}{6}, & \text{if the sample omits block 2} \\ \frac{4}{6}, & \text{if the sample omits block 3.} \end{cases}$$

Each of the values of $\tilde{p}$ has probability 1/3 since the blocks are sampled randomly and so

$$\begin{aligned} \mathbb{E}\left[\tilde{p}\right] &= \frac{1}{3} \times \frac{0}{4} + \frac{1}{3} \times \frac{4}{6} + \frac{1}{3} \times \frac{4}{6} \\ &= \frac{4}{9}. \end{aligned}$$

We therefore see that $\tilde{p}$ is biased.

---

3. **Bootstrapping to Estimating the Scofflaw Rate in Chicago (35 marks)**
Continuing on from the previous question, suppose that the city samples 12 metered blocks at random, and reaches the following data of the form (x_i, z_i):

(4,12)	(3,8)	(3,9)	(3,16)	(2,7)	(4,15)
(4,10)	(3,15)	(2,6)	(1,12)	(3,8)	(2,14)

(a) What is the city's estimate $\tilde{p}$ of the citywide "scofflaw" rate p? **(5 marks)**

Solution: Applying (2) to the data we obtain an estimate of $\tilde{p} = .258$.

(b) Why is this a situation where bootstrapping seems desirable? **(5 marks)**

Solution: Bootstrapping is desirable here because (as with any parameter estimator) we would like to assess the uncertainty / sampling variability associated with it. This cannot be done with $\tilde{p}$ using classical sampling theory (based on the Central Limit Theorem) because $\tilde{p}$ is not a sum (or average) of IID random variables. In fact it is the ratio of two (likely correlated) random variables, $X = \sum_{i=1}^{n} X_i$ and $Z = \sum_{i=1}^{n} Z_i$. Bootstrapping is perfect for this situation as the bootstrap samples allow us to approximate the sampling distribution of the test statistic, i.e. $\tilde{p}$.

(c) In a bootstrap analysis, what would be the approximate distribution of (X, Z) across metered blocks? **(5 marks)**

Solution: The distribution of (X, Z) across metered blocks would be approximated by the empirical distribution of the data. Since there are 12 data-points this empirical distribution is given by

$$P\left(X_i = j,\ Z_i = k\right) = \begin{cases} \frac{1}{12}, & \text{for each } (j, k) \text{ pair that arose in the data-set} \\ 0, & \text{otherwise.} \end{cases}$$

(d) Perform a bootstrap analysis with 1,000 bootstrap samples, and assess whether the estimation procedure suffers an appreciable bias. **(10 marks)**

Solution: See the `RStudio Notebook` *Bootstrap_Chicago_Scofflaws.rmd* for the code for this question and the ones below. The bias at 0.0009473981 is very small indeed.

(e) Construct a 95% confidence interval for p based on percentiles in the bootstrap distribution of $\tilde{p}$? **(5 marks)**

Solution: From the slides we know that

$$(2\tilde{p} - q_u,\ 2\tilde{p} - q_l)$$

yields an approximate $(1 - \alpha)\%$ CI for p where q_l and q_u are the $\alpha/2$ lower- and upper-sample quantiles, respectively, of the bootstrap samples of the test statistic. This yields an approximate 95% CI of $[.199, .311]$.

(f) Create a histogram of the bootstrap statistic under resampling, and in the context of this histogram discuss why adjusting the bootstrap estimate of the "raw" confidence interval is not important here. (The raw confidence interval is the confidence interval $[q_l, q_u]$ where q_l and q_u are the $\alpha/2$ lower- and upper-sample quantiles, respectively, of the bootstrap samples of the test statistic.) **(5 marks)**

Solution: The raw CI, $[q_l, q_u]$, and the adjusted CI, $[2\tilde{p} - q_u,\ 2\tilde{p} - q_l]$, are very similar because there is very little bias and the bootstrap sampling distribution of p is approximately symmetric. This means $\tilde{p} \approx (q_l + q_u)/2$ which, when substituted into the adjusted CI yields the raw CI.

Remark: Note that the bootstrap samples of p can also be easily used to estimate, for example, the probability that the true "scofflaw" rate is 20% or smaller.

---

4. **Quantile Regression (20 marks)**

The usual regression that you're familiar with estimates $\mathrm{E}[y \mid \mathbf{x}]$ where $\mathbf{x}$ is some feature vector. You've also seen logistic regression (and other classification algorithms!) where the goal is to estimate $p(\mathbf{x}) := \mathrm{P}(Y = 1 \mid \mathbf{x})$ where $Y \in \{0, 1\}$ is binary. In this question we're going to consider *quantile regression* where the goal is to estimate $y_q(\mathbf{x})$, the q-quantile of a distribution given some feature vector $\mathbf{x}$. Quantile regression arises in many applications, e.g. economics, finance, epidemiology etc., where we care about estimating the tail, e.g. the 1% tail, of a distribution rather than it's mean as a function of some independent variables $\mathbf{X} \in \mathbb{R}^d$.

Note that part (a) of this question is challenging and entirely optional. Only those of you with a strong mathematical background and who are familiar with the fundamental theorem of calculus (FTC) can tackle it. It is worth zero(!) marks so you should feel free to just use the result of part (a) and move directly to parts (b) and (c).

Remark: Quantile regression is available in `R` via (for example) the `quantreg` package. See the `R Notebook` *Quantile_Regression.Rmd* for an example. The `quantreg` package can also accommodate Lasso penalty terms to aid model sparsity.

(a) Let $Y \in \mathbb{R}$ be a random variable with CDF $F(y)$ and PDF $f(y)$. Consider the (convex) optimization problem

$$\min_{\beta} \ \mathrm{E}\left[q(Y-\beta)^+ + (1-q)(Y-\beta)^-\right] \tag{4}$$

where $(x)^+ := \max\{x, 0\}$ and $(x)^- := \max\{-x, 0\}$. Define the q-th quantile y_q of the random variable Y as any point that satisfies the equation

$$F(y_q) = q.$$

Show that any y_q is an optimal solution of (4). **(0 marks)**

Hint: Write out the expression $\mathrm{E}\left[q(Y-\beta)^+ + (1-q)(Y-\beta)^-\right]$ in terms of the density $f(y)$ of Y and take derivatives with respect to β. (This is where the FTC is required.)

Solution: First note that the function $h(\beta) = \mathrm{E}\left[q(Y-\beta)^+ + (1-q)(Y-\beta)^-\right]$ is a convex function of β. Thus, the first order conditions, i.e. $h'(\beta) = 0$, completely characterize the minima. Expanding $h(\beta)$ we obtain

$$h(\beta) = q\int_{\beta}^{\infty}(y-\beta)f(y)dy + (1-q)\int_{-\infty}^{\beta}(\beta-y)f(y)dy$$

from which it follows (using the FTC) that

$$\begin{aligned} h'(\beta) &= \left(-q(\beta-\beta)f(\beta) - q\int_{\beta}^{\infty} f(y)dy\right) \\ &\quad + \left((1-q)(\beta-\beta)f(\beta) + (1-q)\int_{-\infty}^{\beta} f(y)dy\right) \\ &= -q + F(\beta) \end{aligned}$$

The set $\{\beta : F(\beta) = q\}$ is therefore optimal.

(b) Now suppose $(Y, \mathbf{X}) \in \mathbb{R} \times \mathbb{R}^d$ is a random vector with a joint density. Use the result in part (a) to show that an optimal solution of the optimization problem

$$\min_{\{\beta(\mathbf{x}):\mathbb{R}^d \mapsto \mathbb{R}\}} \mathrm{E}\left[q(Y-\beta(\mathbf{X}))^+ + (1-q)(Y-\beta(\mathbf{X}))^-\right] \tag{5}$$

is given by

$$\beta^*(\mathbf{x}) = \text{Conditional } q\text{-quantile of } Y \text{ given } \mathbf{X} = \mathbf{x}.$$

Note that the minimization is over all functions $\beta(\mathbf{x})$ that map $\mathbb{R}^d$ to $\mathbb{R}$. **(10 marks)**

Solution: Using conditional expectations, it follows that

$$\begin{aligned} &\mathrm{E}\left[q(Y-\beta(\mathbf{X}))^+ + (1-q)(Y-\beta(\mathbf{X}))^-\right] \\ &\quad = \mathrm{E}\left[\mathrm{E}\left[q(Y-\beta(\mathbf{x})^+ + (1-q)(Y-\beta(\mathbf{x}))^- \mid \mathbf{X}=\mathbf{x}\right]\right]. \end{aligned}$$

Since all possible functions are allowed, we can write the minimization problem as

$$\mathrm{E}\left[\min_{\beta} \mathrm{E}\left[q(Y-\beta)^+ + (1-q)(Y-\beta)^- \mid \mathbf{X}=\mathbf{x}\right]\right]$$

so that we solve a separate minimization problem for each value of $\mathbf{x}$. But each of these minimization problems is identical to the problem solved in part (a). Hence it follows that $\beta^*(x)$ = conditional q-quantile of Y given $\mathbf{X} = \mathbf{x}$, is an optimal solution.

(c) Suppose we restrict $\beta(\mathbf{X})$ to be of the form $\beta(\mathbf{X}) = \left[X^\top 1\right] w$ where w is $(d+1)\times 1$ and that we have N IID samples $\{(Y_i, \mathbf{X}_i) : i = 1, \ldots, N\}$. Show that the empirical approximation for the optimization problem (5) is given by

$$\min_{w \in \mathbb{R}^{d+1}} \|y - Mw\|_1 + (2q-1)\, \mathbf{1}^\top (y - Mw) \tag{6}$$

where

$$y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} \qquad M = \begin{bmatrix} \mathbf{X}_1^\top & 1 \\ \mathbf{X}_2^\top & 1 \\ \vdots & \vdots \\ \mathbf{X}_N^\top & 1 \end{bmatrix}.$$

Note that (6) is a convex optimization problem and therefore is straightforward to solve. (Indeed one can also add a Lasso penalty term to the objective.) **(10 marks)**

Solution: The direct empirical approximation is given by

$$\min_{w} \sum_{i=1}^{N} q(Y_i - M_i w)^+ + (1-q)(Y_i - M_i w)^-$$

where M_i denotes the i-th row of M. Next, we use the fact that $z = z^+ - z^-$ and $|z| = z^+ + z^-$ for any z to write this as

$$\min_{w} \sum_{i=1}^{N} \frac{q}{2}\Big(|Y_i - M_i w| + (Y_i - M_i w)\Big) + \frac{1-q}{2}\Big(|Y_i - M_i w| - (Y_i - M_i w)\Big)$$

The result follows by collecting terms and multiplying across by 2.