Homework 7

36-465/665, Spring 2021

Due at 6 pm on Thursday, 23 March 2021

Agenda: Practice working with penalized and constrained model fitting; filling in the details of the algorithmic-stability bound from lecture.

1. *Ridge regression: penalty view* This question will guide us through finding an explicit solution to the problem of minimizing the mean squared error with a penalty on the squared length of the coefficient vector. Specifically, we fix $\lambda > 0$, and ask for

$$\hat{\beta}_\lambda = \operatorname*{argmin}_{b \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} (y_i - (x_i \cdot b))^2 + \lambda \|b\|^2 \tag{1}$$

To simplify the book-keeping, we'll assume that the variables y and x are both centered, so $n^{-1} \sum_i y_i = 0$, $n^{-1} \sum_i x_i = 0$.

   a. (4) As a warm-up, assume that $p = 1$, so x is one dimensional (a scalar). Write out the optimization problem, and show that its solution is

$$\hat{\beta}_\lambda = \frac{n^{-1} \sum_{i=1}^{n} x_i y_i}{\lambda + n^{-1} \sum_{i=1}^{n} x_i^2} \tag{2}$$

   How does this differ from what we'd get from ordinary least squares (OLS)? Does it coincide with OLS in some limit for λ and/or other variables? *Hint*: As usual, take derivative and set equal to 0.

   b. (5) The general, $p > 1$ version of the problem can be written in matrix form as

$$\hat{\beta}_\lambda = \operatorname*{argmin}_{b \in \mathbb{R}^p} \frac{1}{n} (\mathbf{y} - \mathbf{x}\mathbf{b})^T (\mathbf{y} - \mathbf{x}\mathbf{b}) + \lambda \mathbf{b}^T \mathbf{b} \tag{3}$$

   (What are the dimensions of $\mathbf{b}$ as a matrix?) Show that the solution is

$$\hat{\beta}_\lambda = (\mathbf{x}^T \mathbf{x} + n\lambda \mathbf{I})^{-1} \mathbf{x}^T \mathbf{y} \tag{4}$$

   c. (2) Does the formula in Q1b reduce to the formula in Q1a when $p = 1$? Should it?

2. *Ridge regression: constraint form.* In class, I talked about how penalties and constraints are equivalent to each other via Lagrange multipliers ("a fine is a price"). We'll explore that in this problem, taking $p = 1$ for simplicity. Let's abbreviate the unconstrained OLS estimate as $\hat{\beta}_0$. The constraint we'll impose is that $b^2 \leq c$.

   a. (5) Explain why, if $\hat{\beta}_0^2 \leq c$, the constraint is not binding, and the Lagrange multiplier should be 0. Does this mean the Lagrangian should be 0?

   b. (5) Explain why the Lagrangian for the constrained problem is

$$\mathcal{L}(b, \lambda) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i b)^2 + \lambda (b^2 - c) \tag{5}$$

c. (5) Show that the two first-order conditions for minimizing the Lagrangian, assuming the constraint is binding, are

$$(b^*)^2 = c \tag{6}$$

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - b^* x_i)x_i = \lambda^* b^* \tag{7}$$

d. (5) Explain why, if the constraint is binding, $b^* = \sqrt{c}\operatorname{sgn}\hat{\beta}_0$. *Hint*: why should the constrained $\hat{\beta}$ should have the same sign as $\hat{\beta}_0$?

e. (5) Use the first-order conditions, and Q2c, to show that, when the constraint is binding,

$$\lambda^* = \frac{1}{\sqrt{c}\operatorname{sgn}\hat{\beta}_0}\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n} x_i^2 \tag{8}$$

3. *Classifiers with false-positive control.* Suppose we want to predict a binary variable Y from a covariate X. Any way of doing prediction like this is a classifier, and it's describable as a region C: if $x \in C$ then the classifier predicts $Y = 1$, and if $x \notin C$ then the classifier predicts $Y = 0$. (See HW1, Q1–3.) We can make two kinds of errors: false positives, where $Y = 0$ but we guess $Y = 1$, and false negatives, where $Y = 1$ but we guess $Y = 0$. The 0-1 loss doesn't distinguish between these errors. Using different losses (as in HW1Q2) is one way to do so, but another approach is to constrain the false positive rate to be $\leq \alpha$ for some acceptably small α, and minimize the false negative rate. (This can be especially useful when true positives are rare but important to find.) Say that $p(x)$ is the PDF of X conditional on $Y = 1$ and $q(x)$ is the PDF of X conditional on $Y = 0$.

a. (5) Explain what the optimization problem

$$C^* = \operatorname*{argmax}_{C} \int_C p(x)dx \tag{9}$$

$$\text{subject to} \tag{10}$$

$$\int_C q(x)dx \leq \alpha \tag{11}$$

has to do with finding classifiers with controlled false-positive rates.

b. (5) Explain what the optimization problem

$$\max_{C,\lambda} \lambda\alpha + \int_C (p(x) - \lambda q(x))dx \tag{12}$$

has to do with finding classifiers with controlled false-positive rates.

c. (5) Explain why, if $p(x) > \lambda q(x)$, the point x should be in C^*, but if $p(x) < \lambda q(x)$, x should not be in C^*. Also explain why we don't care whether x is in C^* when $p(x) = \lambda q(x)$.

d. (5) Explain why the optimal classifier $s^*(x)$ will always have the form

$$s^*(x) = \begin{cases} 1 & p(x)/q(x) \geq \lambda \\ 0 & p(x)/q(x) < \lambda \end{cases} \tag{13}$$

4. *Stability bounds.* In class, we said that an algorithm A is β-**error stable** when it has the following property: given any two data sets $Z_{1:n}$ and $Z'_{1:n}$, which differ in only one data point, and any new data point z,

$$|\ell(z, A(Z_{1:n})) - \ell(z, A(Z'_{1:n}))| \leq \beta \tag{14}$$

We will use this, and the assumption that $0 \leq \ell \leq m$, to bound $r(A(Z_{1:n})) - \hat{r}(A(Z_{1:n}))$, i.e., how much the algorithm over-fits. It will be convenient to introduce the abbreviation $\Phi(Z_{1:n}) \equiv r(A(Z_{1:n})) - \hat{r}(A(Z_{1:n}))$. We'll first show that $\Phi(Z_{1:n})$ has the bounded difference property, so it will concentrate around its expectation value, and then relate that expectation to β.

a. (5) Show that

$$|\Phi(Z_{1:n}) - \Phi(Z'_{1:n})| \leq |r(A(Z_{1:n})) - r(A(Z'_{1:n}))| + |\hat{r}(A(Z_{1:n})) - \hat{r}(A(Z'_{1:n}))| \tag{15}$$

b. (5) Use the assumption that A is β-stable to show that

$$|r(A(Z_{1:n})) - r(A(Z'_{1:n}))| \leq \beta \tag{16}$$

c. (5) Use the assumption that A is β-stable and that $0 \leq \ell \leq m$ to show that

$$|\hat{r}(A(Z_{1:n})) - \hat{r}(A(Z'_{1:n}))| \leq \frac{(n-1)\beta + m}{n} < \frac{n\beta + m}{n} \tag{17}$$

Hint: Remember that Z and Z' differ in only one data point, which you can take to be the n^{th} one for convenience.

d. (5) Show that $\Phi(Z_{1:n})$ has the bounded difference property with bound $2\beta + m/n$.

e. (5) Show that

$$\mathbb{P}\left(\Phi(Z_{1:n}) \leq \mathbb{E}\left[\Phi(Z_{1:n})\right] + (2n\beta + m)\sqrt{\frac{\log 1/\alpha}{2n}}\right) \geq 1 - \alpha \tag{18}$$

Hint: Use the bounded difference inequality to limit the probability that Φ is ϵ bigger than its expectation, set the probability to α and solve for ϵ (as we've done in earlier homeworks).

f. (3) Use error stability to show that $\left|\mathbb{E}\left[r(A(Z_{1:n})) - \hat{r}(A(Z_{1:n}))\right]\right| \leq \beta_n$.

g. (5) Put the earlier pieces of this question together to show that

$$\mathbb{P}\left(r(A(Z_{1:n})) \leq \hat{r}(A(Z_{1:n})) + \beta_n + (2n\beta_n + m)\sqrt{\frac{\log 1/\alpha}{2n}}\right) \geq 1 - \alpha \tag{19}$$

5. (1) How much time did you spend on this problem set?

Presentation rubric (10): The text is laid out cleanly, with clear divisions between problems and sub-problems. The writing itself is well-organized, free of grammatical and other mechanical errors, and easy to follow. Plots are carefully labeled, with informative and legible titles, axis labels, and (if called for) sub-titles and legends; they are placed near the text of the corresponding problem. All quantitative and mathematical claims are supported by appropriate derivations, included in the text, or calculations in code. Numerical results are reported to appropriate precision.