Machine learning in empirical economic research

Problem set

Andreas Dzemski

June 11, 2019

Problem 1

The Ridge regressor is given by

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}^{p_n}} \sum_{i=1}^{n} \left(y - \beta_0 - x_i' \beta_1\right)^2 + \lambda \|\beta_1\|_2^2$$

1. Show that Ridge regression minimizes a strictly convex function and conclude that $\hat{\beta}^{\text{ridge}}$ is always uniquely defined.

2. Show that

$$\hat{\beta}^{\text{ridge}} = \left(\mathbf{X}'\mathbf{X} + \lambda \operatorname{diag}\left((0, 1, \dots 1)'\right)\right)^{-1} \mathbf{X}'\mathbf{y}$$

3. Suppose that $p_n < n$ and that the off-diagonal elements of $\mathbf{X}'\mathbf{X}/n$ are zero. In that case $\hat{\beta}^{\text{ols}}$ is defined. Show that

$$\left(\frac{(\mathbf{X}'\mathbf{X}/n)_{jj}}{(\mathbf{X}'\mathbf{X}/n)_{jj} + \lambda}\right) \hat{\beta}^{\text{ols}}.$$

   Relate this result to the "shrinkage" property of Ridge regression.

Problem 2

The L_q-penalized least squares estimator is given by

$$\hat{\beta}^{L_q} = \arg\min_{\beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}^{p_n}} \sum_{i=1}^{n} \left(y - \beta_0 - x_i' \beta_1\right)^2 + \lambda \|\beta_1\|_q^q.$$

1. Show that this definition is equivalent to

$$\hat{\beta}^{L_q} = \arg\min_{\beta_0 \in \mathbb{R}, \beta_1 \in \mathbb{R}^{p_n}} \sum_{i=1}^{n} \left(y - \beta_0 - x_i' \beta_1\right)^2$$
$$\text{subject to:} \quad \|\beta_1\|_q \leq s_\lambda$$

   for some s_λ.

2. Verify for L_q-penalized linear regression that $\mathcal{F}_\lambda \subset \mathcal{F}_{\lambda'}$ for $\lambda < \lambda'$.

3. Let $\mathbf{X}_{+1}$ denote the design matrix including intercept ($n \times (p_n + 1)$ matrix). Suppose that $\mathbf{X}_{+1}'\mathbf{X}_{+1}/n$ has full rank. Let $\hat{\beta}^{\text{ols}}$ denote the OLS estimator

$$\hat{\beta}^{\text{ols}} = \left(\mathbf{X}_{+1}'\mathbf{X}_{+1}\right)^{-1} \mathbf{X}_{+1}'\mathbf{y}.$$

   Show that the contour sets

$$\left\{\beta \in \mathbb{R}^{p_n+1} : \|\mathbf{y} - \mathbf{X}_{+1}\beta\|_2^2 = c\right\}$$

   indexed by $c \in \mathbb{R}$ are empty or ellipsoids centered at $\hat{\beta}^{\text{ols}}$.

Problem 3

We consider the the simulation excercise from the slides for the "sparse uncorrelated" design. The repository with the simulation code used to generate the slides can be found at `https://github.com/adzemski/ML_notes`.

1. In the Gaussian case $\hat{\beta}_1^{\text{ols}}$ is the MLE estimator. Explain the efficiency result for MLE estimators (Cramér-Rao lower bound). Theoretically, can $\hat{\beta}_1^{\text{post}}$ beat $\hat{\beta}_1^{\text{ols}}$ in terms of efficiency?

2. Extend the simulation excercise to simulate also the variance and the MSE of $\hat{\beta}_1^{\text{post}}$ and $\hat{\beta}_1^{\text{ols}}$. Which estimator is more efficient?

Problem 4

Implement and simulate the double selection procedure from Belloni, Chernozhukov, and Hansen (2014) for the "sparse correlated" desgin from the slides. The repository with the simulation code used to generate the slides can be found at `https://github.com/adzemski/ML_notes`.

Problem 5

This problem reviews estimating treatment effects by reweighting methods.

Let x denote a covariate vector and let w denote a binary indicator of treatment. Let $y(1)$ and $y(0)$ denote the latent outcomes for the treated and untreated state, respectively. The observed outcome is given by $y = y(1)w + y(0)(1-w)$. We assume

$$y(0), y(1) \perp w \mid x \qquad \text{(unconfoundedness)}.$$

To keep things simple[1] we assume that x is continous with respect to Lebesgue measure. Let f_x denote the unconditional density of x and let $f_{x|w=w'}$ denote the conditional density of x when the treatment takes the value $w = w' \in \{0, 1\}$. You may assume

$$\operatorname{supp}(f_{x|w=0}) = \operatorname{supp}(f_{x|w=1}) \qquad \text{(overlapping support)}.$$

Let p^w denote the unconditional probability of treatment and let $e_1(x')$ denote the conditional probability of treatment if $x = x'$. Let $f(x, w)$ denote the joint density of x and w While this is not a density with respect to Lebesgue measure (what is it?) it still satisfies the equations

$$f_{x|w=1} = \frac{f(x,1)}{p^w} = \frac{e(x)f_x(x)}{p^w} \qquad \text{and} \qquad f_{x|w=0} = \frac{f(x,0)}{1-p^w} = \frac{(1-e(x))f_x(x)}{1-p^w}.$$

Make sure you understand why these equations hold. Our goal is recovering the *average treatment effect on the treated*

$$ATT = \mathbb{E}\left[y(1) - y(0) \mid w = 1\right].$$

1. Argue that $E[y(1) \mid w = 1]$ is trivially identified.

2. Show that there is a constant a depending only of p^w such that

$$\mathbb{E}\left[y(0) \mid w = 1\right] = a\mathbb{E}\left[y\frac{e(x)}{1-e(x)} \mid w = 0\right].$$

   Argue that the expression on the right-hand side is identified up to the value of a.

3. Can you relax the overlapping support assumption without changing your argument? *Hint: Your argument relies on existence of a certain Radon-Nikodym derivative.*

4. Show that

$$a = \left(\mathbb{E}\left[\frac{e(x)}{1-e(x)} \mid w = 0\right]\right)^{-1}.$$

   Hint: $1 = \int f_{x|w=1}(x)\,dx$.

[1] We want to be able to use the notion of a density from elementary calculus, rather than dealing with serious measure theory and Lebesgue integration. Our results will not depend on this simplification.

5. Give conditions under which

$$\frac{1}{n}\sum_{i:w_i=0}\frac{e(x_i)}{1-e(x_i)} = \mathbb{E}\left[\frac{e(x)}{1-e(x)} \mid w=0\right] + o_p\left(n^{-1/2}\right).$$

Relate these conditions to restrictions on how much mass is assigned to $e(x)$ close to 1. Assume that there is an estimator $\hat{e}$ of the propensity score e

$$\frac{1}{n}\sum_{i:w_i=0}\frac{\hat{e}(x_i)}{1-\hat{e}(x_i)} = \frac{1}{n}\sum_{i:w_i=0}\frac{e(x_i)}{1-e(x_i)} + o_p\left(n^{-1/2}\right).$$

Use your results above to argue that an estimator of the treatment effect on the treated is given by

$$\widehat{ATT} = \frac{1}{n}\sum_{i:w_i=1} y_i - \sum_{i:w_i=0}\hat{\gamma}_i y_i,$$

where

$$\hat{\gamma}_i = \frac{\frac{\hat{e}(x_i)}{1-\hat{e}(x_i)}}{\sum_{j:w_j=0}\frac{\hat{e}(x_j)}{1-\hat{e}(x_j)}}.$$

6. We now assume a linear model for the untreated outcome, i.e., we assume[2]

$$Y(0) = x'\beta^c.$$

We consider a general weight function $\gamma(x)$ and identifying the untreated outcome of the treated by

$$\mathbb{E}[y\gamma(x) \mid w=0].$$

For a given weight function γ, define the estimand

$$a(\gamma) = \mathbb{E}[y \mid w=1] - \mathbb{E}[x\gamma(x) \mid w=0].$$

Show that the linearity assumption implies

$$ATT - a(\gamma) = \left(\mathbb{E}[x\gamma(x) \mid w=0] - \mathbb{E}[x \mid w=1]\right)'\beta^c.$$

Relate this to the claim in Athey, Imbens, and Wager (2018), that the bias of a balancing approach to treatment effect estimation is zero if

$$\frac{1}{n}\sum_{i:w_i=1} x_i - \sum_{i:w_i=0} x_i\hat{\gamma}_i$$

is zero.

[2]Here we don't allow an error term. However, assuming $Y(0) = x'\beta^c + \epsilon$ with $\mathbb{E}[\epsilon \mid x, w] = 0$ gives similar results.

7. Now assume that for a smooth (i.e., several times differentiable but not necessarily linear) function m^c

$$y(0) = m^c(x).$$

Can you extend the insights from the linear case to derive conditions for an approximately unbiased balancing estimator for the smooth, non-linear case? *Hint: Taylor expansion.*

Problem 6

1. What is a *quadratic programming problem*?
2. Verify the claim in Athey, Imbens, and Wager (2018) that Step 1 of their procedure is a quadratic programming problem.