STATS305A - Lecture 18

John Duchi
Scribed by Michael Howes

11/30/21

Contents

1 Announcements

- HW4 out today. There will be one multi-part question and one optional question.
- Etude 4 out tonight. Lots of the material for etude 4 will be covered in class on Thursday.
- Both will be due on Thursday next week at 5pm.

2 M-estimation

2.1 Recap

To do regression with M-estimators we use losses other than the squared error to measure error and fit models. For some loss $l : \mathbb{R} \to \mathbb{R}_+$ we solve the minimization problem

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n} l(y_i - x_i^T b),$$

or sometimes

$$\text{minimize } \frac{1}{n}\sum_{i=1}^{n} l(y_i - x_i^T b) + \text{Reg}(b).$$

Where $\text{Reg}(b)$ is some sort of regularizor like the norm of b.

2.2 Choosing the loss function

In ordinary least squares we use the loss function $l(t) = \frac{1}{2}t^2$. There are two guiding principles we can use to choose alternative loss functions.

(a) Suppose that we could solve $\min_f \mathbb{E}[l(Y - f(X))]$ so that $f(x) = \operatorname{argmin}_z \mathbb{E}[l(Y - z)|X = x]$. We would then like to choose the loss l so that the function f has a property that we care about. We will see this when we do quantile regression.

(b) Another approach is to choose a loss function l so that we can avoid undue influence of outlying measurements/responses y. To do this we need a loss that is Lipschitz continuous.

Definition 1. A loss function l is *Lipschitz* is there exists $C \geq 0$ such that for all $t, s \in \mathbb{R}$,

$$|l(t) - l(s)| \leq C|t - s|.$$

The smallest such C is called the *Lipschitz constant* of l and is denoted by $\|l\|_{Lip}$.

Example 1. The absolute error loss $l(t) = |t|$ is Lipschitz with Lipschitz consant 1. We claim that $\operatorname{argmin}_t \mathbb{E}[|Y - t|] = \operatorname{med}(Y)$ where $\operatorname{med}(Y)$ denotes the median of Y. To see why this is true, let $R(t) = \mathbb{E}[|Y - t|]$. We will compute the left and right derivates of R. Note that

$$l'_{\leftarrow}(t) = \text{right derivative of } l = \operatorname{sgn}_+(t) = \begin{cases} 1 & \text{if } t \geq 0, \\ -1 & \text{if } t < 0. \end{cases}$$

And

$$l'_{\rightarrow}(t) = \text{left derivative of } l = \operatorname{sgn}_-(t) = \begin{cases} 1 & \text{if } t > 0, \\ -1 & \text{if } t \leq 0. \end{cases}$$

Thus

$$R'_{\leftarrow}(t) = \mathbb{E}[\operatorname{sgn}_+(t - Y)] = \mathbb{P}(Y \leq t) - \mathbb{P}(Y > t),$$

and

$$R'_{\rightarrow}(t) = \mathbb{E}[\operatorname{sgn}_-(t - Y)] = \mathbb{P}(Y < t) - \mathbb{P}(Y \geq t).$$

Thus if $t < \operatorname{med}(Y)$, then $R'_{\rightarrow}(t) < 0$. If $t > \operatorname{med}(Y)$, then $R_{\leftarrow}(t) > 0$ and thus the minimizer of $R(t)$ is $\operatorname{med}(Y)$.

3 Outlier mitigation

We will now talk more about guiding principle (b) which was about chosing losses that are robust against outliers. Define

$$L_n(b) := \frac{1}{n}\sum_{i=1}^{n} l(y_i - x_i^T b).$$

We want to know what happens to the minimizer of $L_n(b)$ when we replace (x_k, y_k) with (x_k*, y_k^*). Let

$$L_{n,k}(b) = \frac{1}{n}\sum_{i \neq k} l(y_i - x_I^T b) + \frac{1}{n} l(y_k^* - (x_k^*)^T b).$$

Let $\widehat{\beta} = \operatorname{argmin}_b L_n(b)$ and let $\Delta_k = \operatorname{argmin}_\Delta L_{n,k}(\widehat{\beta} + \Delta)$.

Proposition 1 (Heuristic claim)**.** *If l is Lipschitz, smooth and symmetric and the data is "well-conditioned", then $\Delta_k = O(1/n)$ (so (x_k, y_k) has a small influence on $\widehat{\beta}$).*

Sketch of proof. Note that l being Lipschitz and differentiable implies that the derivates of l are uniformly bounded by some constant C. Observe that

$$\begin{aligned}
\nabla_\Delta L_{n,k}(\widehat{\beta}+\Delta) &= \frac{1}{n}\sum_{i\neq k} l'(x_i^T(\widehat{\beta}+\Delta)-y_i)x_i + \frac{1}{n}l'((x_k*)^T(\widehat{\beta}+\Delta)-y_k^*)x_k^* \\
&= \frac{1}{n}\sum_{i=1}^{n} l'(x_i^T(\widehat{\beta}+\Delta)-y_i)x_i \\
&\quad + \frac{1}{n}\Big(l'((x_k*)^T(\widehat{\beta}+\Delta)-y_k^*)x_k^* - l'(x_k^T(\widehat{\beta}+\Delta)-y_k)x_k\Big) \\
&= \nabla_\Delta L_n(\widehat{\beta}+\Delta) + \frac{1}{n}\left(C_k^* x_k^* - C_k x_k\right),
\end{aligned}$$

where $C_k^* = l'((x_k*)^T(\widehat{\beta}+\Delta)-y_k^*)$ and $C_k = l'(x_k^T(\widehat{\beta}+\Delta)-y_k)$. We can do a Taylor's approximation of $\nabla_\Delta L_n(\widehat{\beta}+\Delta)$ at $\Delta = 0$. We know that $\nabla_\Delta L_n(\widehat{\beta}+\Delta)\mid_{\Delta=0}= 0$ since $\widehat{\beta}$ minimizes $L_n(b)$. Thus

$$\nabla_\Delta L_{n,k}(\widehat{\beta}+\Delta) \approx \nabla^2 L_n(\widehat{\beta})\Delta + \frac{1}{n}\left(C_k^* x_k^* - C_k x_k\right).$$

If we set the right hand side equal to 0 we get

$$\Delta_k \approx \nabla^2 L_n(\widehat{\beta})^{-1}\frac{1}{n}\left(C_k^* x_k^* - C_k x_k\right).$$

This gives the approximation since l being Lipschitz implies that C_k^*, C_k are bounded and $\{x_i\}_{i=1}^n$ being well-conditioned implies that the inverse Hessian $\nabla^2 L_n(\widehat{\beta})^{-1}$ exists and isn't too wild. □

Definition 2. The *influence function* of an M-estimator is (roughly) the change in $\widehat{\beta}^*$ when we change one observation. For M-estimators with $L(b) = \mathbb{E}[l(Y - X^T b)]$ and $\beta^* = \operatorname{argmin}_b L(b)$, we define the influence function ψ to be

$$\psi(x,y) = (\Delta^2 L(\beta^*))^{-1} l'(y - x^T\beta^*)x.$$

Theorem 1. *Suppose $n\to\infty$. Let*

$$\widehat{\beta}(x,y) = \operatorname*{argmin}_b \left\{\sum_{i=1}^n l(y_i - x_i^T b) + l(y - x^T b)\right\},$$

and

$$\widehat{\beta} = \operatorname*{argmin}_b \left\{\sum_{i=1}^n l(y_i - x_i^T b)\right\}.$$

Then, we have shown heuristically,

$$\widehat{\beta}(x,y) = \widehat{\beta} + \frac{1}{n}\psi(x,y) + o(1/n).$$

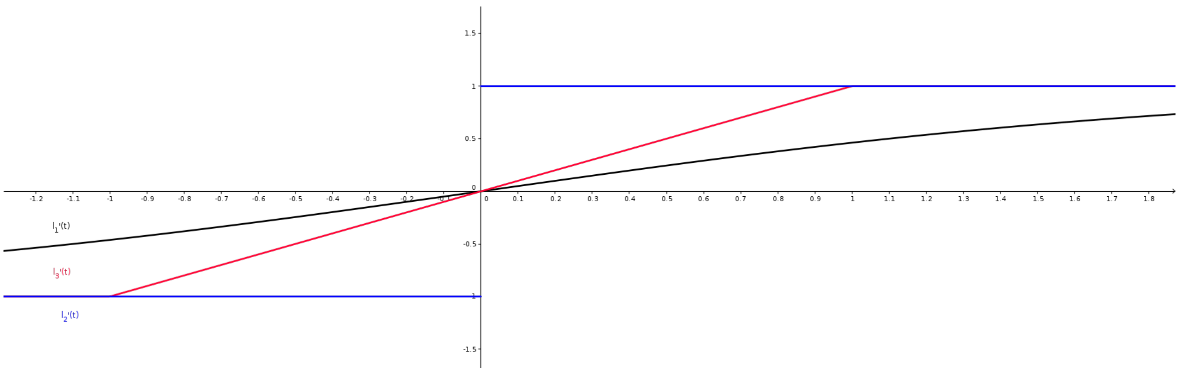
The upshot is that if the derivative $l'(t)$ is bounded, then the estimator is robust. If $l'(t)$ is unbounded, then we the estimator is not robust. For example the following losses

$$\begin{aligned}
l_1(t) &= \log(1+e^t) + \log(1+e^{-t}), \\
l_2(t) &= |t| \\
l_3(t) &= \begin{cases} \frac{1}{2}t^2 & \text{if } |t|\leq 1, \\ |t| - \frac{1}{2} & \text{if } |t| > 1. \end{cases}
\end{aligned}$$

have corresponding derivatives which are all bounded on $\mathbb{R}$

$$\begin{aligned}
l_1'(t) &= \frac{1}{1+e^{-t}} - \frac{1}{1+e^{t}}, \\
l_2'(t) &= \operatorname{sgn}(t) \\
l_3'(t) &= \begin{cases} t & \text{if } |t| \leq 1, \\ \operatorname{sgn}(t) & \text{if } |t| > 1. \end{cases}
\end{aligned}$$

The derivatives look like this with l_1' in black, l_2' in blue and l_3' in red. Note that l_2' and l_3' overlap for $|t| \geq 1$.



4 Qunatile regression

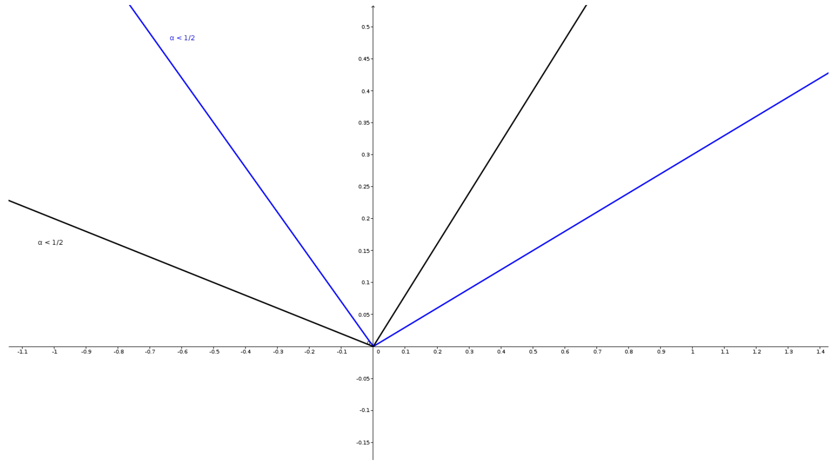
Often, instead of predicting Y, we want an interval where Y is likely to be. That is, we would like a confidence set $\widehat{c}(x) = [q_\alpha(x), q_{1-\alpha}(x)]$ such that $\mathbb{P}(Y \in \widehat{c}(x) | X = x) = 1 - 2\alpha$ (this is impossible but let's try anyway). There are many applications where we are more interested in $\widehat{c}(x)$ than $\widehat{y}$. For example we might be

- Predicting election results,
- Predicting patient survival times,
- Predicting air quality.

What loss should we choose to get something like this? We could fit two (or more) models that predict the α, $1-\alpha$ quantiles of $Y|X = x$. To do this we use the *pinball loss* (also called the quantile loss). For $\alpha \in (0, 1)$, let l_α be the loss function given by

$$l_\alpha(t) = \alpha(t)_+ + (1-\alpha)(-t)_+,$$

where $(t)_+ = \max 0, t$ so that $(t)_+ = 0$ for $t \leq 0$ and $(t)_+ = t$ for $t \geq 0$. The function l_α looks something like this. The blue loss is typical for $\alpha > 1/2$ and the black loss is typical for $\alpha < 1/2$. These losses are chosen with guiding principle (a) in mind. We will now show that the minimizers of $\mathbb{E}[l_\alpha(Y - t)]$ are α-quanties of Y.



Define $R_\alpha(t) = \mathbb{E}[l_\alpha(Y-t)]$, then

$$R'_\alpha(t) = \mathbb{E}[-\alpha \mathbf{1}(Y > t) + (1-\alpha)\mathbf{1}(Y \le t)] = (1-\alpha)\mathbb{P}(Y \le t) - \alpha \mathbb{P}(Y > t).$$

If $\mathbb{P}(Y \le t) > \alpha$, then $\mathbb{P}(Y > t) < 1-\alpha$ and $R'_\alpha(t) > 0$ so t is too large. If $\mathbb{P}(Y \le t) < \alpha$, then $\mathbb{P}(Y > t) > 1-\alpha$ and $R'_\alpha(t) < 0$ so t is too small. Thus

$$\operatorname*{argmin}_t R_\alpha(t) = \inf\{t : \mathbb{P}(Y \le t) \ge \alpha\} = \alpha\text{-quantile of } Y.$$

Quantile regression is the following procedure. For $\alpha \in (0, 1/2)$, fit two models

$$\widehat{\beta}_\alpha = \operatorname*{argmin}_b \sum_{i=1}^n l_\alpha(y_i - x_i^T b) \quad \text{and} \quad \widehat{\beta}_{1-\alpha} = \operatorname*{argmin}_b \sum_{i=1}^n l_{1-\alpha}(y_i - x_i^T b).$$

Then our prediction intervals are

$$\widehat{c}(x) = [\widehat{\beta}_\alpha^T x, \widehat{\beta}_{1-\alpha}^T x].$$

More generally, instead of linear functions we could consider fitting $f \in \mathcal{F}$ where $\mathcal{F}$ is some function class. We would then have

$$\widehat{f}_\alpha = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n l_\alpha(y_i - f(x_i)) \quad \text{and} \quad \widehat{f}_{1-\alpha} = \operatorname*{argmin}_{f \in \mathcal{F}} \sum_{i=1}^n l_{1-\alpha}(y_i - f(x_i)).$$

We would then again define

$$\widehat{c}(x) = [\widehat{f}_\alpha(x), \widehat{f}_{1-\alpha}(x)].$$

Some issues:

- There is no guarantee that $\widehat{f}_\alpha(x) \le \widehat{f}_{1-\alpha}(x)$. We can fix this by defining

$$\widehat{c}(x) = [\min\{\widehat{f}_\alpha(x), \widehat{f}_{1-\alpha}(x)\}, \max\{\widehat{f}_\alpha(x), \widehat{f}_{1-\alpha}(x)\}].$$

- We do not have any gaurantee that $\widehat{c}(x)$ is a valid confidence interval for Y given $X = x$. That is we have no reason to believe that

$$\mathbb{P}(Y \in \widehat{c}(x) | X = x) = 1 - \alpha.$$

The second issue cannot be resolved. It is a fact that without strong assumptions, you cannot have *conditional converage*. That is we cannot find a procedure $\widehat{c}$ based on a i.i.d. sample $\{x_i, y_i\}_{i=1}^n$ such that

$$\mathbb{P}(Y_{n+1} \in \widehat{c}(x_{n+1}) | X_{n+1} = x_{n+1}) = 1 - \alpha + o(1),$$

where Y_{n+1}, x_{n+1} is a new independent date point from the same distribution as our sample.

We can achieve *marginal coverage*. That is the exists a procedure $\widehat{c}$ such that

$$\mathbb{P}(Y_{n+1} \in \widehat{c}(X_{n+1})) \geq 1 - \alpha.$$

Furthermore if Y is continuous, then we can find $\widehat{c}$ such that

$$\mathbb{P}(Y_{n+1} \in \widehat{c}(X_{n+1})) = 1 - \alpha \pm \frac{1}{n}.$$

These ideas will be discussed on Thursday when we look at conformal confidence intervals.