Constant predictors

Sanjay Lall and Stephen Boyd

EE104
Stanford University

Constant predictors

- we explore the simplest possible predictor, which is *constant*
- $\hat{y} = g_\theta(x) = \theta \in \mathbf{R}^m$
- a linear regression model with $\phi(u) = 1$
- doesn't depend on u, which in fact we don't even need
- we'll use ERM to fit θ to data
- we don't need regularization since the predictor is (completely) insensitive
- different losses lead to different predictors

Losses

- we are given data $y^1, \ldots, y^n \in \mathbf{R}^m$
- we have a *loss* function $\ell : \mathbf{R} \times \mathbf{R} \to \mathbf{R}$
- $\ell(\hat{y}, y)$ quantifies how badly $\hat{y}$ approximates y
- typical losses for scalar y $(m = 1)$:
  - *quadratic loss*: $\ell(\hat{y}, y) = (\hat{y} - y)^2$
  - *absolute loss*: $\ell(\hat{y}, y) = |\hat{y} - y|$
  - *fractional loss*: for $\hat{y}, y > 0$,

$$\ell(\hat{y}, y) = \max\left\{\frac{\hat{y}}{y} - 1, \frac{y}{\hat{y}} - 1\right\} = \exp\big(|\log \hat{y} - \log y|\big) - 1$$

    (often scaled by 100 to become *percentage error*)
- typical loss for vector y $(m > 1)$: *quadratic loss*, $\ell(\hat{y}, y) = \|\hat{y} - y\|_2^2$

ERM

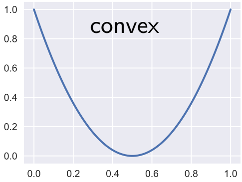
- we choose θ to minimize empirical risk, $\mathcal{L}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(\theta, y^i)$
- we'll be able to solve this minimization problem for the losses above, and others
- we'll recover some reasonable choices of a constant approximation of the data, such as mean and median

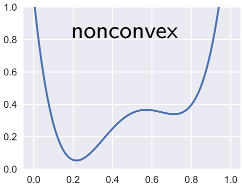
Convexity

- a function $f : \mathbf{R}^k \to \mathbf{R}$ is *convex* if it for all $w, z \in \mathbf{R}^k$ and all $\alpha \in [0, 1]$

$$f(\alpha w + (1-\alpha)z) \leq \alpha f(w) + (1-\alpha) f(z)$$

- this means the function 'curves upward' or has positive curvature
- in terms of derivatives, convexity can be expressed as
    - (if $f'(w)$ exists) $f'(w)$ is nondecreasing (as w increases)
    - (if $f''(w)$ exists) $f''(w) \geq 0$ for all w





Minimizing convex functions — optimality conditions

for a convex function f

- if f is differentiable f, w minimizes f if and only if $\nabla f(w) = 0$

for convex $f : \mathbf{R} \to \mathbf{R}$ (*i.e.*, $k = 1$)

- w minimizes f if and only if $f'_-(w) \leq 0$, $f'_+(w) \geq 0$
- $f'_+(w)$ is the *righthand derivative*, $f'_+(w) = \lim_{t \to 0, t>0} \frac{f(w+t)-f(w)}{t}$
- $f'_-(w)$ is the *lefthand derivative*, $f'_-(w) = \lim_{t \to 0, t<0} \frac{f(w+t)-f(w)}{t}$
- these both exist, even if f is not differentiable
- if $f'(w)$ exists, then $f'_-(w) = f'_+(w) = f'(w)$
- simple example: $w = 0$ minimizes $f(w) = |w|$, since $f'_-(0) = -1$, $f'_+(0) = 1$

ERM and convexity

- for the losses functions listed above (and many others), $\ell(\hat{y}, y)$ is a convex function of $\hat{y}$
- an average of convex functions is convex, so $\mathcal{L}(\theta)$ is convex
- so the optimality conditions above tell us when θ minimizes $\mathcal{L}(\theta)$
- for scalar y, θ minimizes $\mathcal{L}(\theta)$ when $\mathcal{L}'_{-}(\theta) \leq 0$, $\mathcal{L}'_{+}(\theta) \geq 0$

Square loss

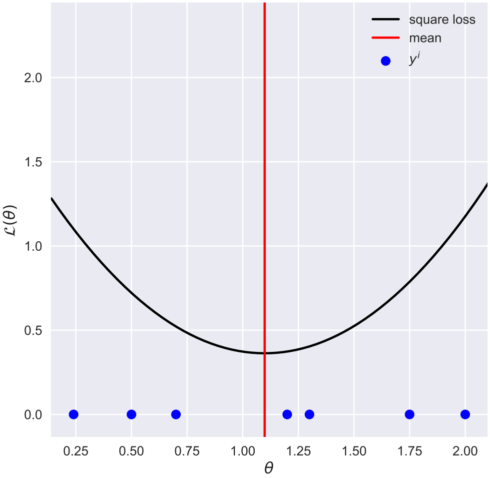
ERM with square loss

- for square loss $\ell(\hat{y}, y) = ||\hat{y} - y||_2^2$, empirical risk is *mean-square error* (MSE)

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}||\theta - y^i||_2^2$$

- a simple least squares problem, with solution $\theta = \frac{1}{n}\sum_{i=1}^{n} y^i$ (which satisfies $\nabla\mathcal{L}(\theta) = 0$)
- *i.e.*, best constant predictor with square loss is the *average* or *mean* of the data
- with this best predictor, mean square error is the *variance* of the data

ERM with square loss



Absolute loss

ERM with absolute loss

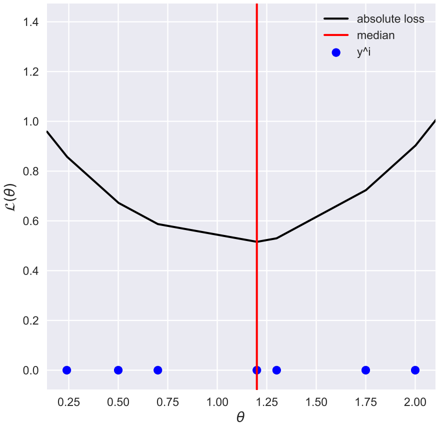
- for absolute loss $\ell(\hat{y}, y) = |\hat{y} - y|$, empirical risk is *mean-absolute error*

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}|\theta - y^i|$$

- $\mathcal{L}(\theta)$ is convex and piecewise linear, with kink points at the data values $y^1, \ldots, y^n$

- we'll see that θ is optimal if and only if it is a *median* of the data
- another reasonable constant approximation of the data

ERM with absolute loss



Median

- for $\theta \in \mathbf{R}$ define

$$n_1 = |\{y^i \mid y^i < \theta\}| \qquad \text{number of data points less than } \theta$$
$$n_2 = |\{y^i \mid y^i > \theta\}| \qquad \text{number of data points greater than } \theta$$

- we say θ is a *median* of the data if

$$\frac{n_1}{n} \leq \frac{1}{2} \quad \text{and} \quad \frac{n_2}{n} \leq \frac{1}{2}$$

- if $\theta \neq y^i$ for any i then this is the same as $\dfrac{n_1}{n} = \dfrac{1}{2}$

Median

- assume data is *sorted* so $y^1 \leq y^2 \leq \cdots \leq y^n$
- if n is odd, the median is $\theta = y^{(n+1)/2}$ (median is unique in this case)
- if n is even, θ is a median if $y^{n/2} \leq \theta \leq y^{n/2+1}$ (median is not unique in this case)

- examples:
  - the median of -3.3, -1.7, 0.4 is -1.7
  - the median of -3.3, -1.7, 0.4, 4.9 is any number in $[-1.7, 0.4]$

Medians minimize empirical risk with absolute loss

- we'll show that θ minimizes $\mathcal{L}(\theta)$ (with absolute loss) if and only if θ is a median of the data
- assume data are sorted, $y^1 \leq \cdots \leq y^n$, then

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n_1}(\theta - y^i) + \frac{1}{n}\sum_{i=1+n-n_2}^{n} -(\theta - y^i)$$

- so if θ is not equal to a data value

$$\mathcal{L}'(\theta) = \frac{d}{d\theta}\mathcal{L}(\theta) = \frac{n_1}{n} - \frac{n_2}{n}$$

- left and right derivatives are

$$\mathcal{L}'_-(\theta) = \frac{2n_1}{n} - 1 \qquad \mathcal{L}'_+(\theta) = 1 - \frac{2n_2}{n}$$

- θ is optimal means $\mathcal{L}'_-(\theta) \leq 0$ and $\mathcal{L}'_+(\theta) \geq 0$, which is

$$\frac{n_1}{n} \leq \frac{1}{2} \qquad \frac{n_2}{n} \leq \frac{1}{2}$$

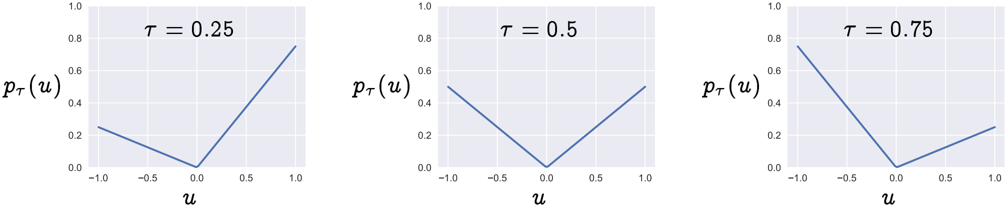
Tilted absolute loss

Tilted absolute value function

- for $\tau \in [0, 1]$ the *tilted absolute value function* is

$$p_\tau(u) = \begin{cases} -\tau u & u < 0 \\ (1-\tau)u & u \geq 0 \end{cases}$$

- can be expressed as $p_\tau(u) = (1/2 - \tau)u + (1/2)|u|$

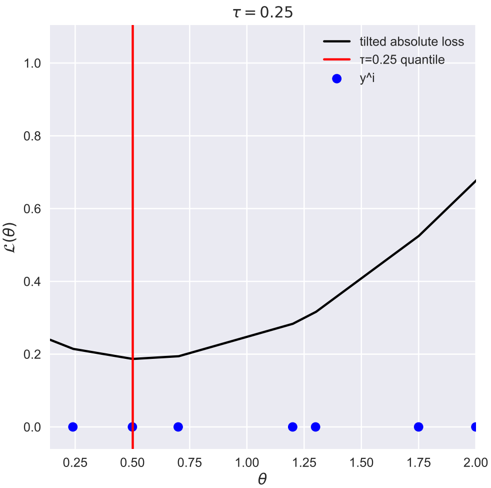


ERM with tilted absolute value loss

- empirical risk with *tilted absolute loss* $\ell(\hat{y}, y) = p_\tau(\hat{y} - y)$ is $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n p_\tau(\hat{y} - y)$
- $\mathcal{L}(\theta)$ is convex and piecewise linear, with kink points at the data values $y^1, \ldots, y^n$
- for $\tau < 1/2$, it's worse (more loss) to over-estimate y ($\hat{y} > y$) than to under-estimate
- for $\tau > 1/2$, it's worse (more loss) to under-estimate y than to overestimate

- we'll see that θ is optimal if it is a *τ-quantile* of the data
- roughly, the fraction of y^i's less than θ is around τ

ERM with tilted absolute loss



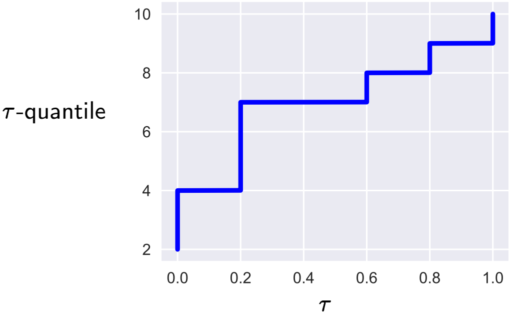
Quantiles

- for $\tau \in [0,1]$, we call θ a *τ-quantile* of the data if

$$\boxed{\frac{n_1}{n} \leq \tau \leq 1 - \frac{n_2}{n}}$$

- if $\theta \neq y^i$ for all i then this is the same as $\tau = n_1/n$
- some common quantiles have names like
  - median ($\tau = 0.5$)
  - quartiles ($\tau = 0.25, 0.5, 0.75$)
  - deciles ($\tau = 0.1, 0.2, \ldots, 0.9$)
  - percentiles ($\tau = 0.01, 0.02, \ldots, 0.99$)

Quantiles



- if the data is (4,7,7,8,9) then
  - the 0.1 quantile is 4
  - the 0.2 quantile is any number in [4,7]
  - the 0.5 quantile is 7

τ-quantile minimizes empirical risk with tilted absolute loss

θ minimizes $\mathcal{L}(\theta)$ if and only if it is a τ-quantile

- assume data are sorted, $y^1 \le \cdots \le y^n$, then

$$\mathcal{L}(\theta) = p_\tau(\theta - y^1) + \cdots + p_\tau(\theta - y^n) = \frac{1}{n}\sum_{i=1}^{n_1}(1-\tau)(\theta - y^i) + \frac{1}{n}\sum_{i=1+n-n_2}^{n} -\tau(\theta - y^i)$$

- if θ is not equal to a data value, then $\mathcal{L}'(\theta) = (n_1(1-\tau) - \tau n_2)/n$
- left and right derivatives are

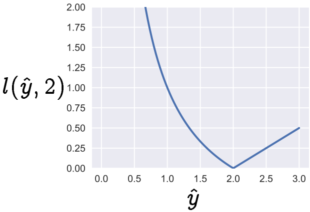
$$\mathcal{L}'_-(\theta) = (n_1(1-\tau) - \tau(n-n_1))/n = \frac{n_1}{n} - \tau$$

$$\mathcal{L}'_+(\theta) = ((n-n_2)(1-\tau) - \tau n_2)/n = 1 - \tau - \frac{n_2}{n}$$

- θ is optimal means $\mathcal{L}'_-(\theta) \le 0$ and $\mathcal{L}'_+(\theta) \ge 0$, which means $\dfrac{n_1}{n} \le \tau \le 1 - \dfrac{n_2}{n}$

Fractional loss

ERM with fractional loss

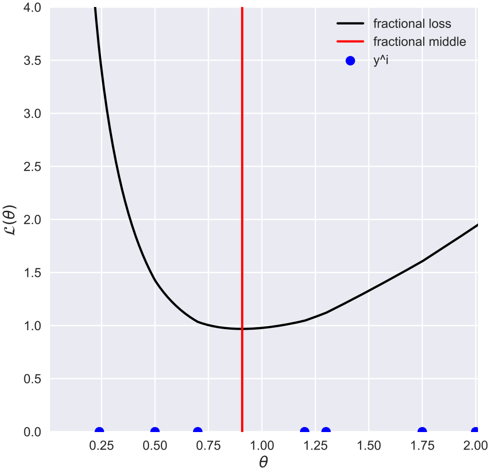


- fractional loss $\ell(\hat{y}, y) = \max\left\{\frac{\hat{y}}{y} - 1, \frac{y}{\hat{y}} - 1\right\} = \exp\left(|\log \hat{y} - \log y|\right) - 1$
- empirical risk is

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n} \max\left\{\frac{\theta}{y^i} - 1, \frac{y^i}{\theta} - 1\right\}$$

- a convex function, with kink points at $y^1, \ldots, y^n$
- we call θ that minimizes $\mathcal{L}(\theta)$ the *fractional middle* of $y^1, \ldots, y^n$ (not a standard term)

ERM with fractional loss



ERM with fractional loss

- with $y^1 \leq \cdots \leq y^n$ and $y^k \leq \theta \leq y^{k+1}$, we have

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{k}\left(\frac{y^i}{\theta} - 1\right) + \frac{1}{n}\sum_{i=k+1}^{n}\left(\frac{\theta}{y^i} - 1\right) = -1 + \frac{1}{n}\sum_{i=1}^{k}\frac{y^i}{\theta} + \frac{1}{n}\sum_{i=k+1}^{n}\frac{\theta}{y^i}$$

- so for $y^k < \theta < y^{k+1}$ we have

$$\mathcal{L}'(\theta) = -\frac{1}{\theta^2}\left(\frac{1}{n}\sum_{i=1}^{k} y^i\right) + \frac{1}{n}\sum_{i=k+1}^{n}\frac{1}{y^i}$$

- $\mathcal{L}'(\theta)$ is an increasing function of θ (since it is convex)
- first find k so that $\mathcal{L}'_+(y^k) \leq 0$ and $\mathcal{L}'_-(y^{k+1}) \geq 0$ (using above expression evaluated at y^k and y^{k+1})
- setting $\mathcal{L}'(\theta)$ to zero we get

$$\theta = \left(\frac{\sum_{i=1}^{k} y^i}{\sum_{i=k+1}^{n} 1/y^i}\right)^{1/2}$$

Summary

Summary

- the simplest predictor is a constant, $\hat{y} = g_\theta(u) = \theta$
- for different losses, ERM gives different θs
- for some common losses, we recover well known predictors of a set of data
  - square loss given mean
  - absolute loss gives median
  - tilted absolute loss gives quantile