Brendan Neal 1001160236|nealbre1 2018-10-13

CSC411 Homework 3

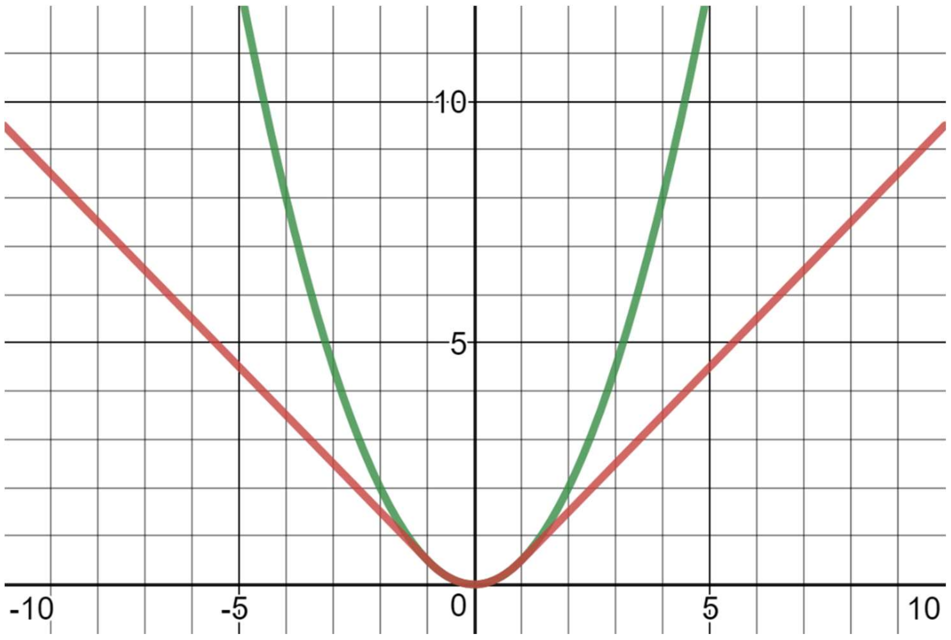
Question 1: Robust Regression

1a: Sketching Huber vs. Squared Error Loss

Error functions to plot

Full Equation	$\boldsymbol{t = 0}$
$L_\delta(y,t) = \begin{cases} \frac{1}{2}(y-t)^2, & \lvert y-t \rvert \le \delta \\ \delta\left(\lvert y-t \rvert - \frac{1}{2}\delta\right), & \lvert y-t \rvert > \delta \end{cases}$	$L_\delta(y,0) = \begin{cases} \frac{1}{2}y^2, & \lvert y \rvert \le \delta \\ \delta\left(\lvert y \rvert - \frac{1}{2}\delta\right), & \lvert y \rvert > \delta \end{cases}$
$L_{SE}(y,t) = \frac{1}{2}(y-t)^2$	$L_{SE}(y,0) = \frac{1}{2}y^2$

Plot of Error Functions



LEGEND	
RED	$L_\delta(y,0), \delta = 1$
GREEN	$L_{SE}(y,0)$
HORIZONTAL AXIS	y

Huber loss should be more robust to outliers because it is less affected by errors that occur outside the range $[-\delta, \delta]$.

Outside this range, Huber loss behaves linearly and therefore has a constant rate of change. Conversely, squared error has an increasing rate of change. Thus, the error of a prediction affects the squared loss far more than Huber loss. Since outliers have (comparably) large errors, Huber loss should be far less affected by them.

Brendan Neal 1001160236|nealbre1 2018-10-13

1b: Partial Derivatives of Huber Loss

Final Answer

$$\frac{dL_\delta}{dw} = x \cdot H'_\delta(y-t) \qquad\qquad \frac{dL_\delta}{db} = H'_\delta(y-t)$$

Where

$$H'_\delta(a) = \frac{dH_\delta}{da}(a) = \begin{cases} a, & |a| \le \delta \\ -\delta, & a < -\delta \\ \delta, & a > \delta \end{cases}$$

Derivation

Expanding delta multiplication in last case

$$H_\delta(a) = \begin{cases} \frac{1}{2}a^2, & |a| \le \delta \\ \delta\left(|a| - \frac{1}{2}\delta\right), & |a| > \delta \end{cases}$$

$$\Rightarrow H_\delta(a) = \begin{cases} \frac{1}{2}a^2, & |a| \le \delta \\ \delta|a| - \frac{1}{2}\delta^2, & |a| > \delta \end{cases}$$

Eliminating absolute values in last case

$$\Rightarrow H_\delta(a) = \begin{cases} \frac{1}{2}a^2, & |a| \le \delta \\ -\delta a - \frac{1}{2}\delta^2, & a < -\delta \\ \delta a - \frac{1}{2}\delta^2, & a > \delta \end{cases}$$

Derivative of piecewise is piecewise of component derivatives

$$\Rightarrow \frac{dH_\delta}{da}(a) = \begin{cases} \left(\frac{1}{2}a^2\right)\left(\frac{d}{da}\right), & |a| \le \delta \\ \left(-\delta a - \frac{1}{2}\delta^2\right)\left(\frac{d}{da}\right), & a < -\delta \\ \left(\delta a - \frac{1}{2}\delta^2\right)\left(\frac{d}{da}\right), & a > \delta \end{cases}$$

Evaluating the derivatives

$$\Rightarrow \frac{dH_\delta}{da}(a) = \begin{cases} a, & |a| \le \delta \\ -\delta, & a < -\delta \\ \delta, & a > \delta \end{cases}$$

Desired partial derivatives can obtained using implicit differentiation

$$\frac{dH_\delta}{dw}(a) = \frac{dH_\delta}{da}(a) \times \frac{da}{dw} \qquad\Bigg|\qquad \frac{dH_\delta}{db} = \frac{dH_\delta}{da}(a) \times \frac{da}{db}$$

Determining a's partial derivatives:

$$a = y - t = w^T x + b - t$$

$$\frac{da}{db} = \left(\frac{d}{db}\right)(w^T x + b - t) = 1$$

$$\frac{da}{dw} = \left(\frac{d}{dw}\right)(w^T x + b - t) = x$$

Evaluating multiplication of derivatives

$$\frac{dH_\delta}{dw}(a) = \frac{dH_\delta}{da}(a) \times \frac{da}{dw}$$

$$\Rightarrow \frac{dH_\delta}{dw}(a) = x\frac{dH_\delta}{da}(a)$$

$$\frac{dH_\delta}{db}(a) = \frac{dH_\delta}{da}(a) \times \frac{da}{db}$$

$$\Rightarrow \frac{dH_\delta}{db}(a) = \frac{dH_\delta}{da} \times 1$$

$$\Rightarrow \frac{dH_\delta}{db}(a) = \frac{dH_\delta}{da}(a)$$

For brevity, define

$$H'_\delta(a) = \frac{dH_\delta}{da}(a) = \begin{cases} a, & |a| \le \delta \\ -\delta, & a < -\delta \\ \delta, & a > \delta \end{cases}$$

$$\Rightarrow \frac{dH_\delta}{dw}(a) = xH'_\delta(a)$$

$$\Rightarrow \frac{dH_\delta}{db}(a) = H'_\delta(a)$$

Recall $L_\delta(y,t) = H_\delta(y-t)$

$$\Rightarrow \frac{dL_\delta}{dw}(y,t) = \frac{dH_\delta}{dw}(y-t)$$

$$\Rightarrow \frac{dL_\delta}{dw}(y,t) = xH'_\delta(y-t)$$

$$\Rightarrow \frac{dL_\delta}{db}(y,t) = \frac{dH_\delta}{db}(y-t)$$

$$\Rightarrow \frac{dL_\delta}{db}(y,t) = H'_\delta(y-t)$$

Question 2: Locally Weighted Regression

2a: Matrix Expression of Weighted Least Squares

$$w^* = \underset{w}{\operatorname{argmin}}\left(\frac{1}{2}\sum_{i=1}^{N} a^{(i)}\left(y^{(i)} - w^T x^{(i)}\right)^2 + \frac{\lambda}{2}\left|\left|w\right|\right|^2\right)$$

Using calculus to find the minimizing w (since formula is convex, occurs when derivative WRT w is 0

$$0 = \frac{d}{dw}\left(\frac{1}{2}\sum_{i=1}^{N} a^{(i)}\left(y^{(i)} - w^T x^{(i)}\right)^2 + \frac{\lambda}{2}\left|\left|w\right|\right|^2\right)$$

Breaking up the addition to compute the derivative

$$\frac{d}{dw}\left(\frac{1}{2}\sum_{i=1}^{N} a^{(i)}\left(y^{(i)} - w^T x^{(i)}\right)^2\right)$$

Use chain rule on bracketed group

$$= \frac{1}{2}\sum_{i=1}^{N} a^{(i)}(2)\left(-x^{(i)}\right)\left(y^{(i)} - w^T x^{(i)}\right)$$

$$= -\sum_{i=1}^{N} a^{(i)}\left(x^{(i)}\right)\left(y^{(i)} - w^T x^{(i)}\right)$$

Expanding multiplication

$$= -\sum_{i=1}^{N} x^{(i)} a^{(i)} y^{(i)} + \sum_{i=1}^{N} x^{(i)} a^{(i)} w^T x^{(i)}$$

Since $w, x^{(i)}$ are vectors, $w^T x^{(i)} = \left(x^{(i)}\right)^T w$

$$= -\sum_{i=1}^{N} x^{(i)} a^{(i)} y^{(i)} + \sum_{i=1}^{N} x^{(i)} a^{(i)} \left(x^{(i)}\right)^T w$$

$$\frac{d}{dw}\left(\frac{\lambda}{2}\left|\left|w\right|\right|^2\right)$$

Derivative of vector's squared norm WRT the vector is twice the vector

$$= \frac{\lambda}{2} \times 2w^T$$

$$= \lambda w$$

$$\Rightarrow 0 = -\sum_{i=1}^{N} x^{(i)}\, a^{(i)} y^{(i)} + \sum_{i=1}^{N} x^{(i)} a^{(i)} \left(x^{(i)}\right)^T w + \lambda w$$

Isolating terms with w to other side

$$\Rightarrow \sum_{i=1}^{N} x^{(i)} a^{(i)} \left(x^{(i)}\right)^T w + \lambda w = \sum_{i=1}^{N} x^{(i)} a^{(i)} y^{(i)}$$

Factoring out w

$$\left(\sum_{i=1}^{N} x^{(i)} a^{(i)} \left(x^{(i)}\right)^T + \lambda\right) w = \sum_{i=1}^{N} x^{(i)} a^{(i)} y^{(i)}$$

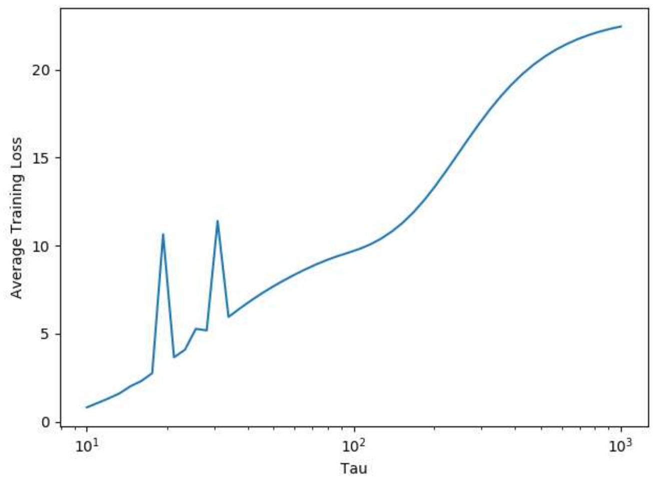
Expressing as matrices:

$$(X^T A X + \lambda I) w = X^T A y$$

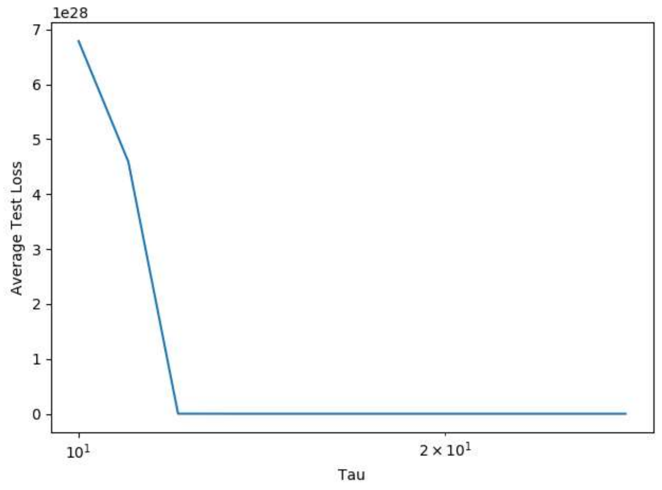
$$\Rightarrow w = (X^T A X + \lambda I)^{-1} X^T A y$$

2c: Ro vs Training and Validation Loss

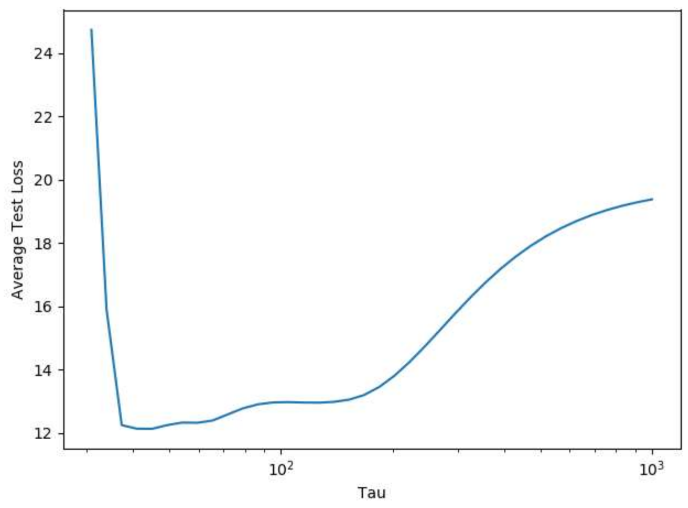
Training Loss (Mean Squared Error)



Test / Validation Loss Part 1 (Mean Squared Error)



Test / Validation Loss Part 2 (Mean Squared Error)



2d: Algorithm Performance Under Extreme Ro Values

Intuitively, locally weighted regression makes points closer to the datum play a larger role in the datum's prediction. Tau is used to regulate the relative importance ("weight") of each point.

Smaller tau values mean that the algorithm pays more attention to points near the datum when making a prediction. Therefore as tau approaches 0, the algorithm would pay more and more attention to the datum's closest point. Eventually, the closest point to the datum would effectively decide the entirety of the datum's prediction.

Conversely, larger tau values would cause the algorithm to consider points more fairly. As tau approaches infinity, the exponents would approach zero. This would effectively make all the weights equal to (1 / number of data points). This effectively eliminates the weights for the regression, causing it to degenerate into simple linear regression.

In reality, the algorithm seemed to obey these predictions. For small values of tau, the algorithm had low training loss but very high validation loss. This is because the algorithm was overfitting the training data by only paying attention to the closest points to the datum. Conversely, the algorithm performed poorly for the validation data because its prediction was too biased to the training data.

Validation loss hit a minimum around $\tau = 20$, but rose after this point before starting to plateau. This is because the algorithm stopped overfitting by paying more attention to points further away from the datum, increasing its generalization. However, after this point, the validation loss rose because the algorithm started to fit the training data more and became less generalized.

Training loss continually increased with tau, but started to plateau. This signifies the algorithm starting to degenerate into simple linear regression when every point is weighed roughly evenly. When this occurs the same prediction will be output regardless of tau.