HW3 Solutions

October 2018

1 Robust Regression

a)

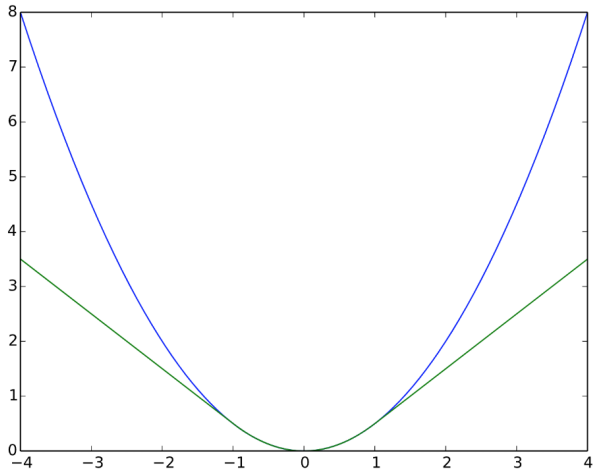


Figure 1: Green: Huber Loss. Blue: Squared Error Loss

We can expect the Huber loss to be more robust to outliers because, for large errors, the loss scales linearly compared to quadratically for the squared error loss.

b)

We begin with the derivative of the Huber loss:

$$H_{\delta}^{'}(a) \begin{cases} a & |a| \leq \delta \\ \delta & a > \delta \\ -\delta & a < -\delta \end{cases}$$

Now we take partial derivatives wrt to the weights and biases:

$$\begin{aligned}\frac{\partial L_\delta(y,t)}{\partial \mathbf{w}} &= \frac{\partial H_\delta(y-t)}{\partial \mathbf{w}} \\ &= \frac{\partial H_\delta(y-t)}{\partial (y-t)}\frac{\partial (y-t)}{\partial \mathbf{w}} \\ &= H_\delta^{'}(y-t)\frac{\partial(\mathbf{w}^T\mathbf{x}+b)}{\partial \mathbf{w}} \\ &= H_\delta^{'}(y-t)\cdot \mathbf{x}\end{aligned}$$

$$\begin{aligned}\frac{\partial L_\delta(y,t)}{\partial b} &= \frac{\partial H_\delta(y-t)}{\partial b} \\ &= \frac{\partial H_\delta(y-t)}{\partial (y-t)}\frac{\partial (y-t)}{\partial b} \\ &= H_\delta^{'}(y-t)\frac{\partial(\mathbf{w}^T\mathbf{x}+b)}{\partial b} \\ &= H_\delta^{'}(y-t)\end{aligned}$$

2 Locally Weighted Regression

a)

The loss can be written as:

$$L(w) = \frac{1}{2}(\boldsymbol{y}-\boldsymbol{w}\boldsymbol{X})^T\boldsymbol{A}(\boldsymbol{y}-\boldsymbol{w}\boldsymbol{X}) + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}$$

Take the derivative of the loss wrt w:

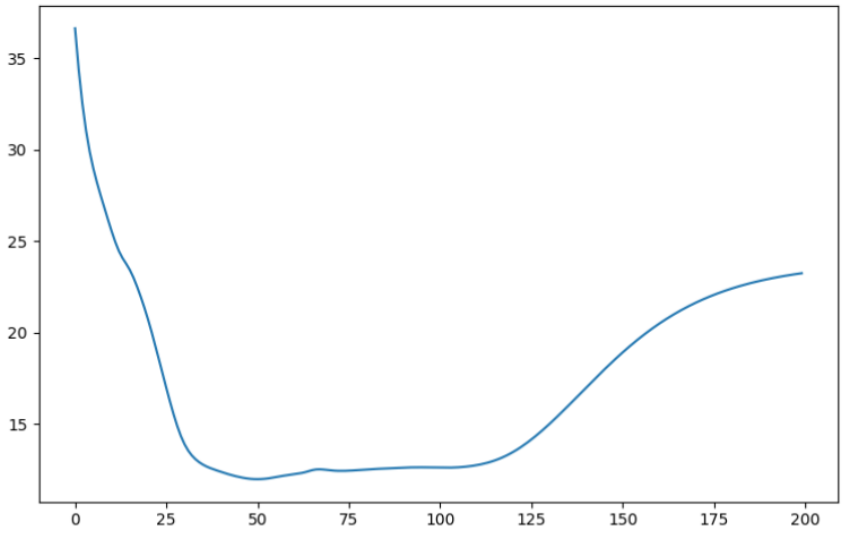
$$\nabla_w L(w) = -\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w} + \lambda\boldsymbol{w}$$

Set to 0 and solve for w^*:

$$\begin{aligned}0 &= -\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} + \boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X}\boldsymbol{w^*} + \lambda\boldsymbol{w^*} \\ &= -\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y} + (\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} + \lambda\boldsymbol{I})\boldsymbol{w^*} \\ \boldsymbol{w^*} &= (\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^T\boldsymbol{A}\boldsymbol{y}\end{aligned}$$

c)

The validation loss should look like this:



(y-axis: validation loss, x-axis: $\log \tau$)

d)

When $\tau \to \infty$ the model behaves like standard linear regression - each training point is equally weighted. When $\tau \to 0$ the model becomes very sensitive to points in the neighbourhood of the test point. This leads to poor generalization (can be viewed as a form of overfitting).