Cost Function, Binary Classifier and Performance Measurement

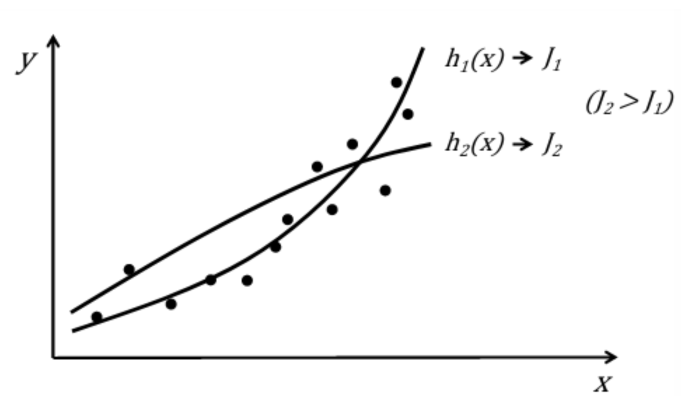
1 Cost Functions

Supervised learning problem

- Collection of n p-dimensional feature vectors: $\{x_i\}, i = 1...n$
- Collection of observed responses $\{y_i\}, i = 1...n$
- Aims to construct a response surface $h(x)$
- Describes how well the current response surface $h(x)$ fits the available data (on a given set) - we use J to represent the cost function

$$J(y_i, h(x_i))$$

- Smaller values of the cost function correspond to a better fit, so in the graph below $J_2 > J_1$
- Machine learning goal: construct $h(x)$ such that J is minimised
- In regression, $h(x)$ is usually directly interpretable as a predicted response

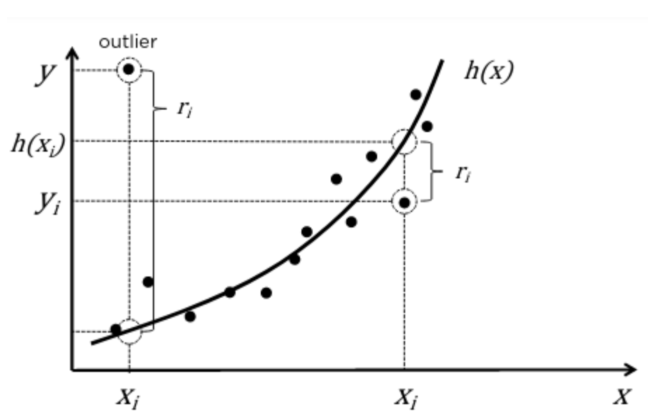


1.1 Least squares deviation cost

$$J(y_i, h(x_i)) = \frac{1}{n}\sum_{i=1}^{n}(y_i - h(x_i))^2$$

r_i is the difference between the real value and the predicted value

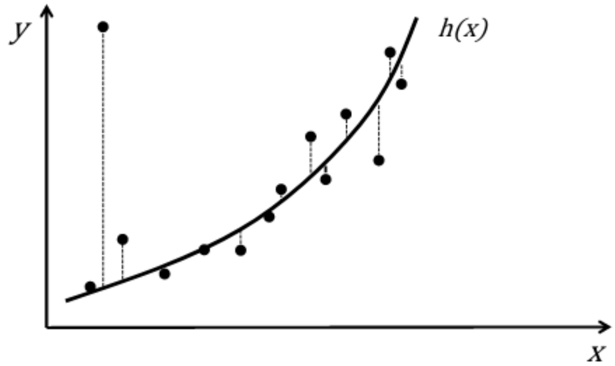
- Nice mathematical properties
- Problem with outliers- when you have a large residual and it is then squared, the impact is large where it should be ignored



1.2 Least Absolute Deviation Cost

$$J(y_i, h(x_i)) = \frac{1}{n}\sum_{i=1}^{n} |(y_i - h(x_i))|$$

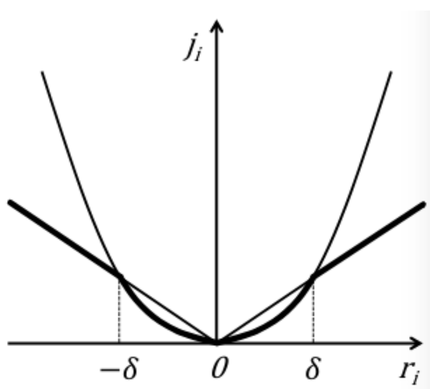
- More robust with respect to outliers - not squared residual so less impact
- May pose computational challenges



1.3 Huber-M Cost

$$J(y_i, h(x_i)) = \frac{1}{n}\sum_{i=1}^{n} \left\{ \begin{array}{ll} 0.5(y_i - h(x_i))^2 & \text{if } |y_i - h(x_i)| < \delta \\ \delta(|y_i - h(x_i)| - 0.5\delta) & \text{otherwise} \end{array} \right\}$$

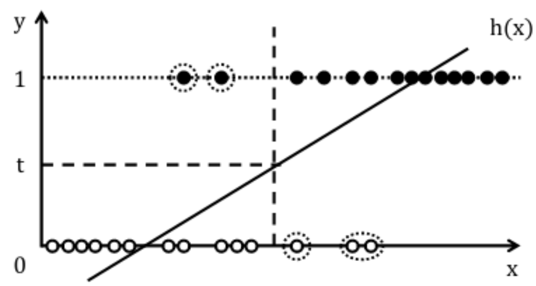
- Combines the best qualities of the LS and LAD losses (basically using one or the other depending on which one is better to use)
- Parameter δ is usually set automatically to a specific percentile of absolute residuals. Calculate all residuals, then for example top 10% is δ



2 Binary Classifier

- Observed response y takes only two possible values + and -
- Define relationship between $h(x)$ and y
- If larger than the threshold, then set to 1, if less then set to 0
- Use the decision rule:

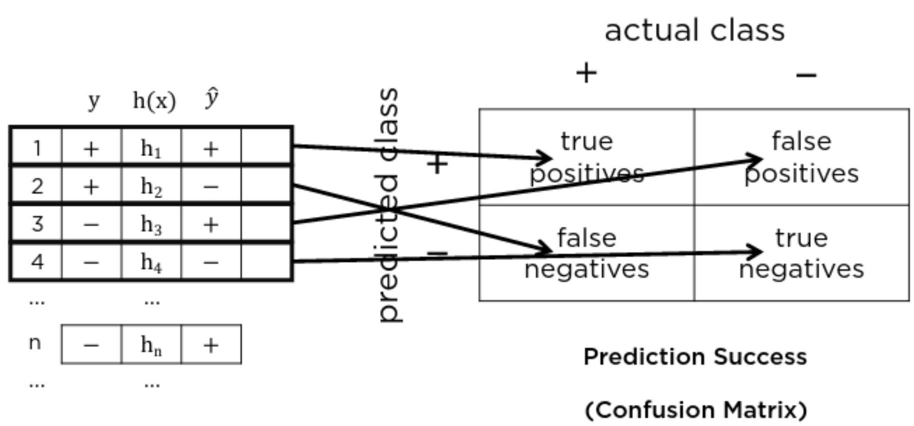
$$\hat{y} = \begin{cases} +, & h(x) \geq t \\ -, & \text{otherwise} \end{cases}$$

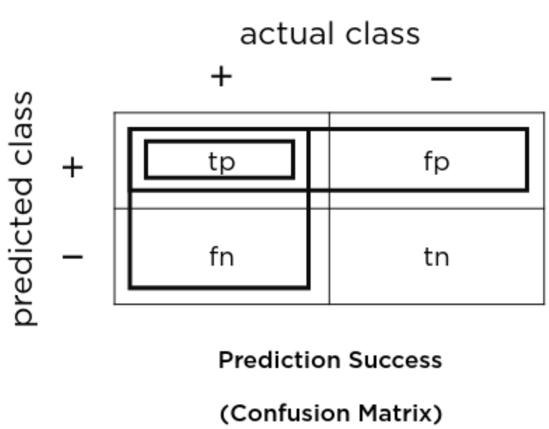


3 Performance Measures

3.1 Precision and Recall

How well did we capture the + group for the given threshold



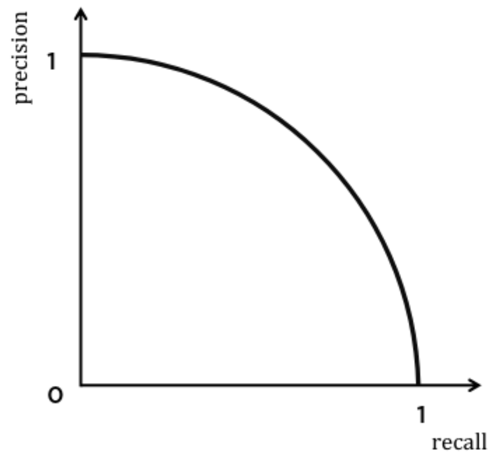


Precision:

$$\frac{tp}{tp+fp} > 1$$

Recall (Sensitivity)

$$\frac{tp}{tp+fn} > 1$$



3.2 ROC Curve

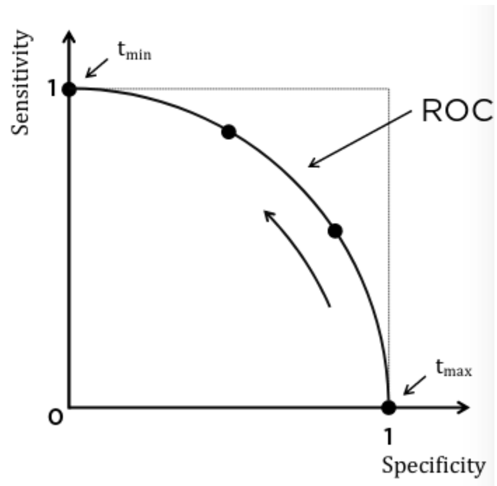
Recall (sensitivity)

$$\frac{tp}{tp+fn}$$

Specificity

$$\frac{tn}{tn+fp}$$

y	h(x)	$\hat{y}$	
+	h_1		max
+	h_2		
−	h_3		
−	h_4		
+	h_5		
−	h_6		
+	h_7		
−	h_8		
−	h_9		min



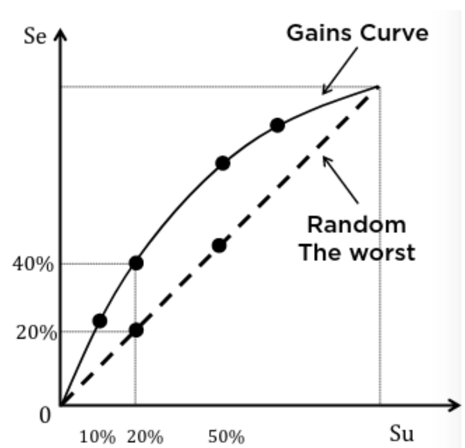
3.3 Gains and Lift

Sensitivity (recall)

$$Se = \frac{tp}{tp + fn}$$

Support (% pop)

$$Su = \frac{tp + fp}{n}$$



Base rate

$$Br = \frac{tp + fn}{n}$$

Gains

$$\{Su, Se\}$$

Lift

$$\{Su, \frac{Se}{Su}\}$$

ROC

$$\{\frac{Su - Br \cdot Se}{1 - Br}, Se\}$$