Machine Learning

Lecture 4 – Cost Function, Binary Classifier and Performance Measurement

Dr SHI Lei

Last Lecture

- Generalisation
- Training & Test Set
- Representation

Last Lecture

Generalisation

The big picture

- **Goal**: to predict well on new data drawn from (hidden) true distribution.
- **Issue**: we don't see the truth, but we only get to sample from it.
- If it fits current sample well, how can we trust it will predict well on other new samples?

Last Lecture

Generalisation

Three basic assumptions:

1. We draw examples independently and identically (i.i.d.) at random from the distribution.
2. The distribution is stationary - it doesn't change over time.
3. We always pull from the same distribution, including training, validation, and test sets.

Last Lecture

Training & Test Set

Divide into two sets:

- Training set
- Test set

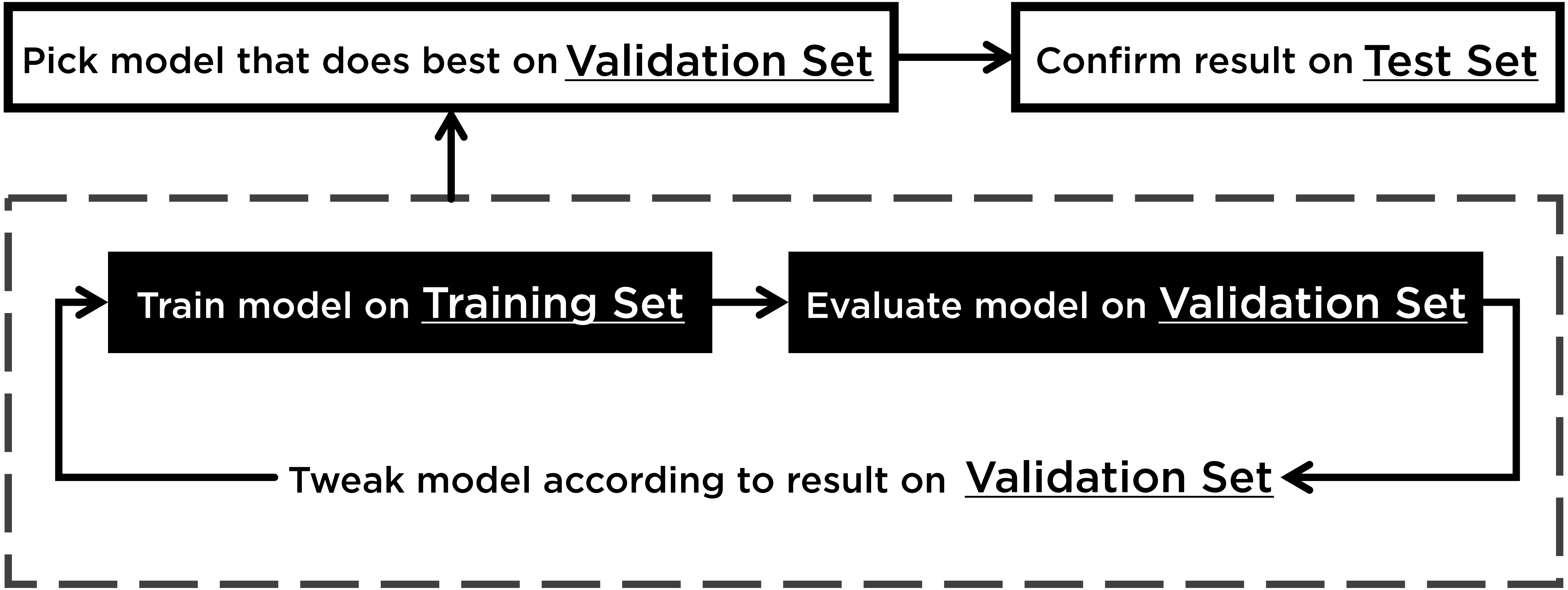
Training Set	Test Set

Do not train on test data

Last Lecture

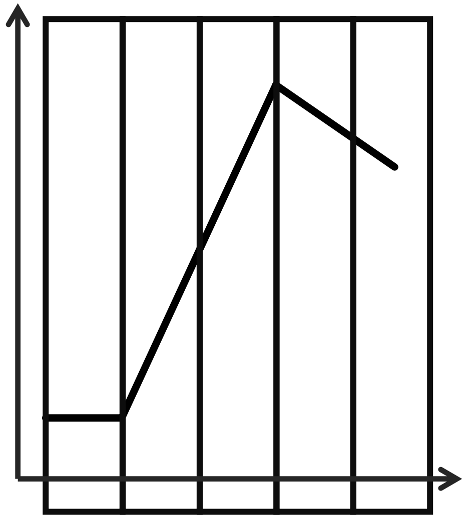
Training & Test Set

Better Workflow: Use a Validation Set

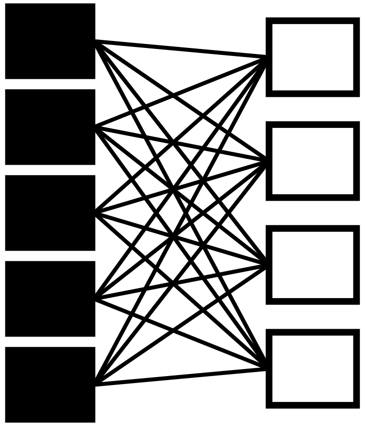


Last Lecture

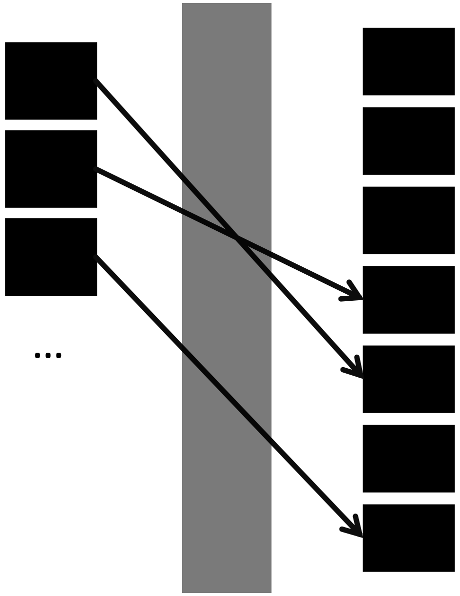
Representation



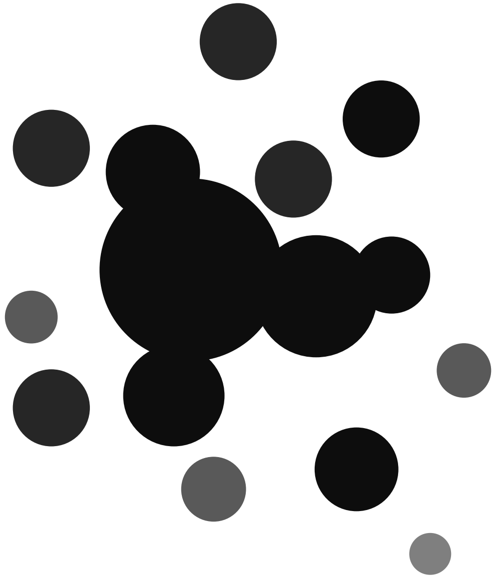
bucketing



crossing



hashing



embedding

Today

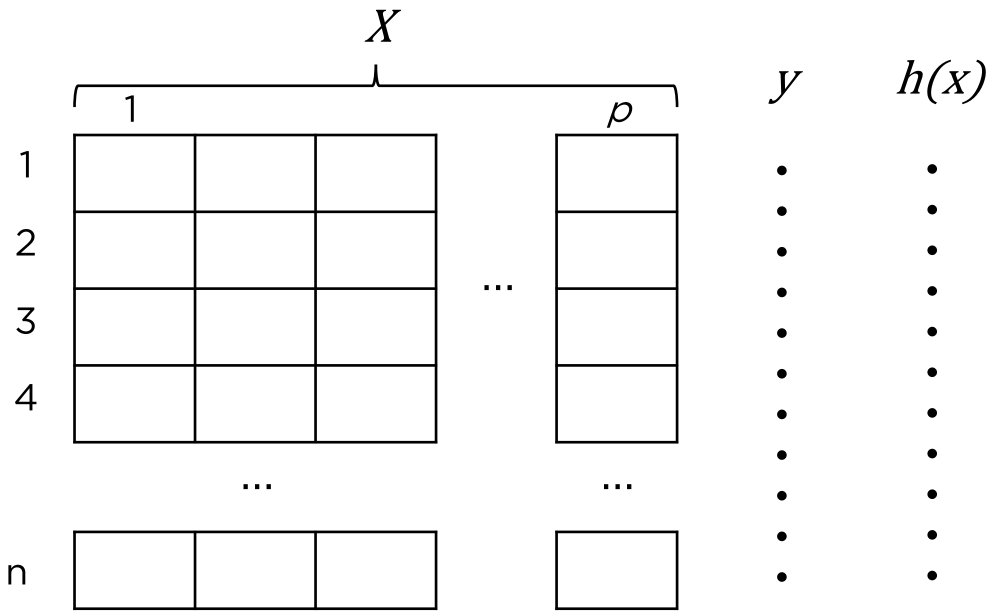
- Cost Functions
- Binary Classifier
- Performance Measures

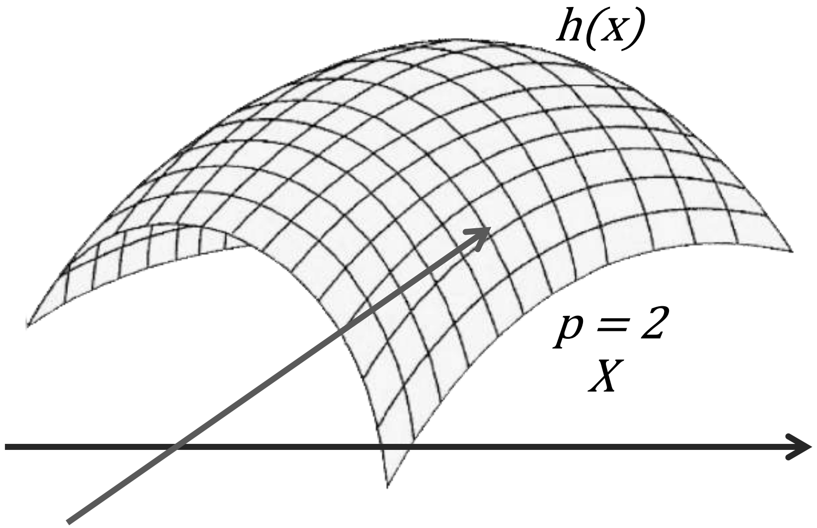
Cost Functions

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: $h(x)$



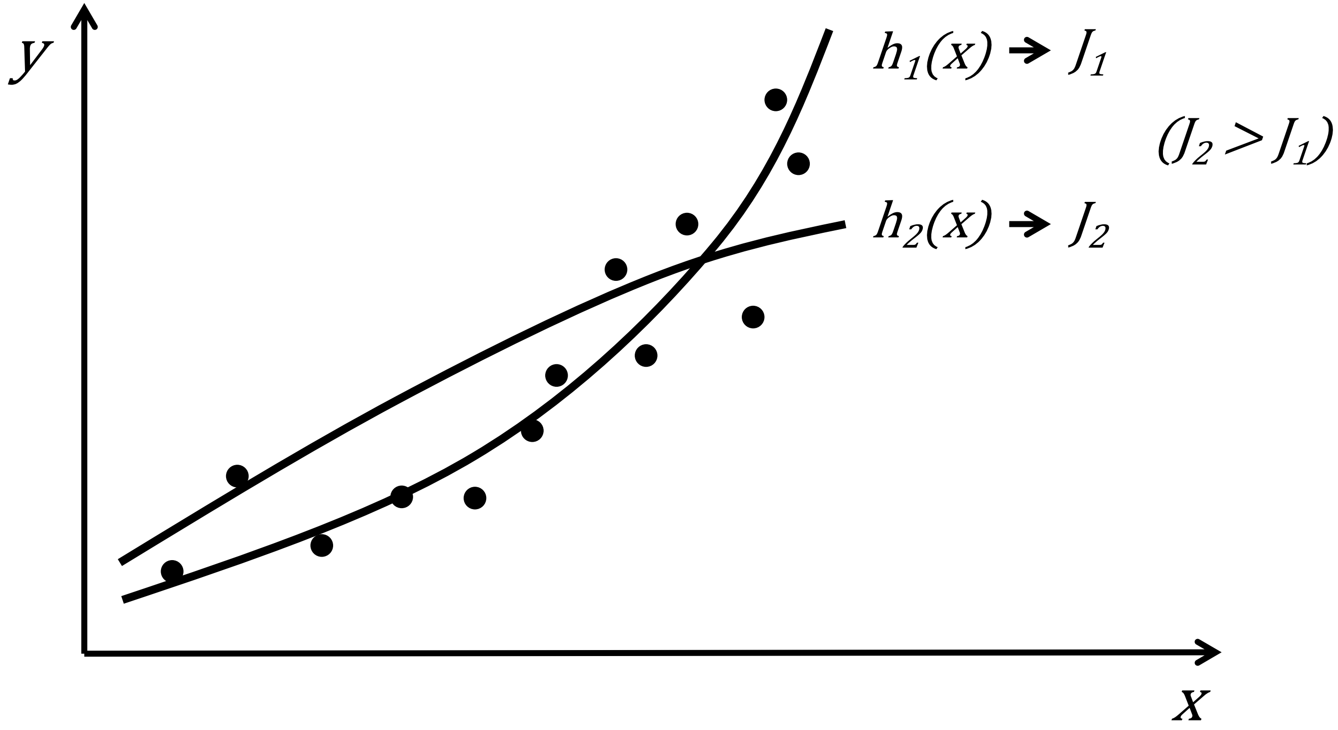


Cost Functions

- Describes how well the current response surface ***h(x)*** fits the available data (on a given data set):

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

observed predicted



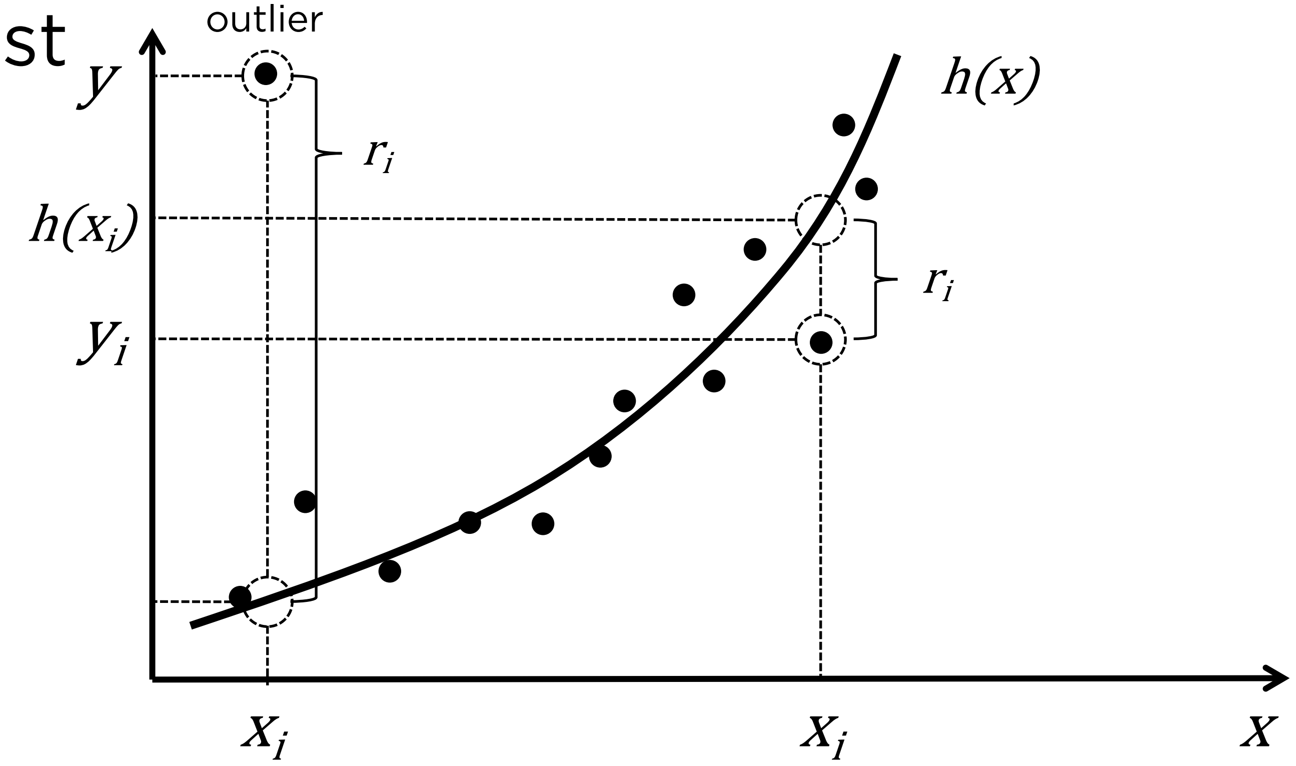
- Smaller values of the cost function correspond to a better fit.
- Machine learning goal: construct ***h(x)*** such that ***J*** is minimised.
- In regression, ***h(x)*** is usually directly interpretable as predicted response.

Cost Functions

Least Squares Deviation Cost

- Defined as

$$J(y_i, h(x_i)) = \boxed{\frac{1}{n}} \sum_{i=1}^{n} \underbrace{(y_i - h(x_i))}_{r_i \text{ (residual)}}{}^{\boxed{2}}$$



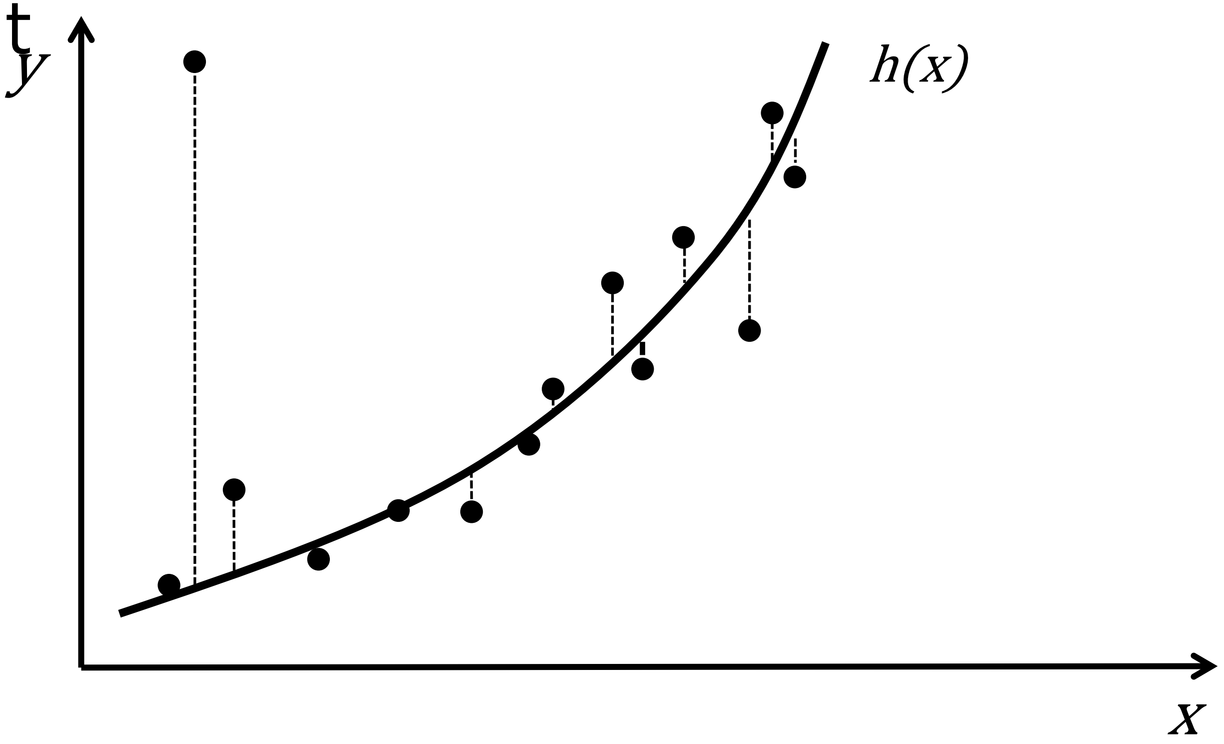
- Nice mathematical properties
- Problem with outliers

Cost Functions

Least Absolute Deviation Cost

- Defined as

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

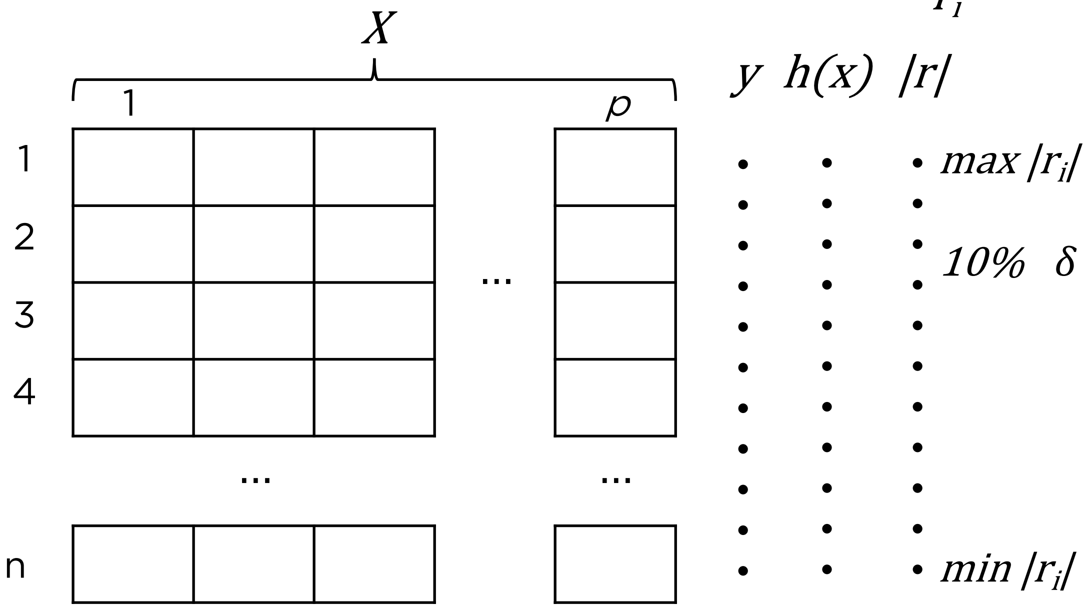


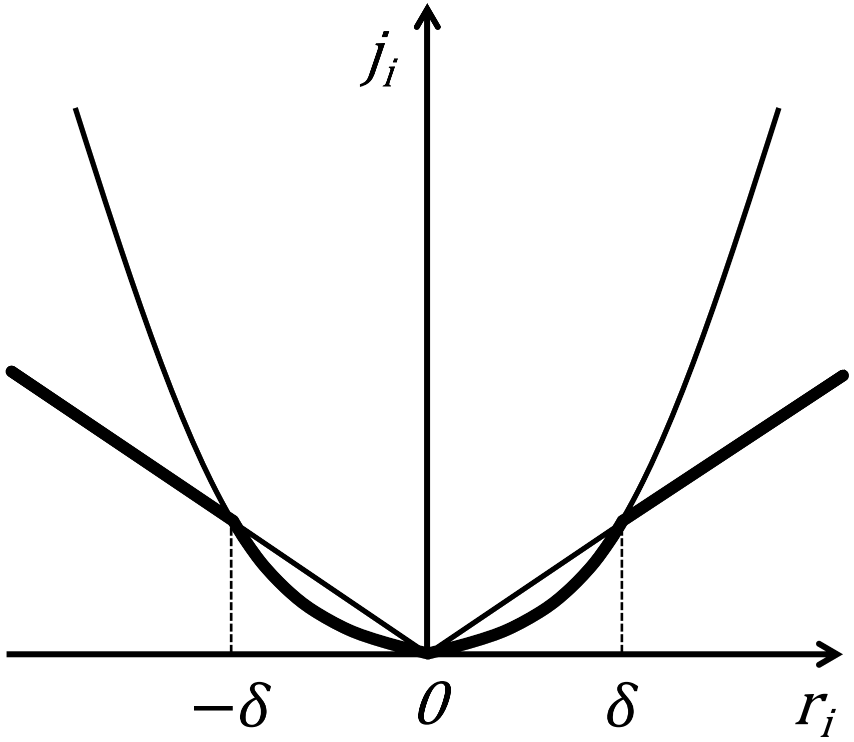
- More robust with respect to outliers
- May pose computational challenges

Huber-M Cost

- Defined as

$$J(y_i, h(x_i)) = \frac{1}{n}\sum_{i=1}^{n} \overbrace{\begin{cases} 0.5(\underline{y_i - h(x_i)})^2, if\, |\underline{y_i - h(x_i)}| < \delta \\ \delta(|\underbrace{y_i - h(x_i)}_{r_i}| - 0.5\delta), otherwise \end{cases}}^{J_i}$$





- Combines the best qualities of the LS and LAD losses
- Parameter δ is usually set automatically to a specific percentile of absolute residuals

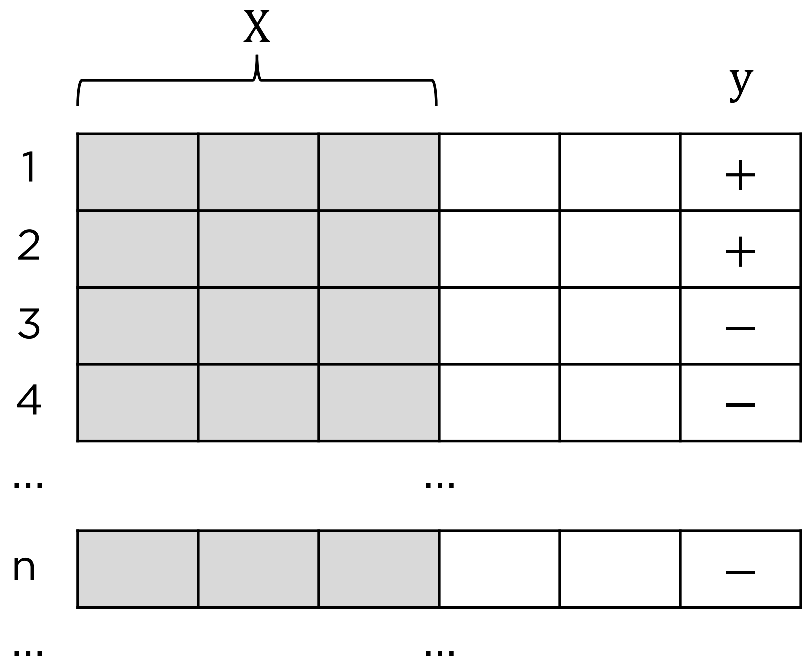
Today

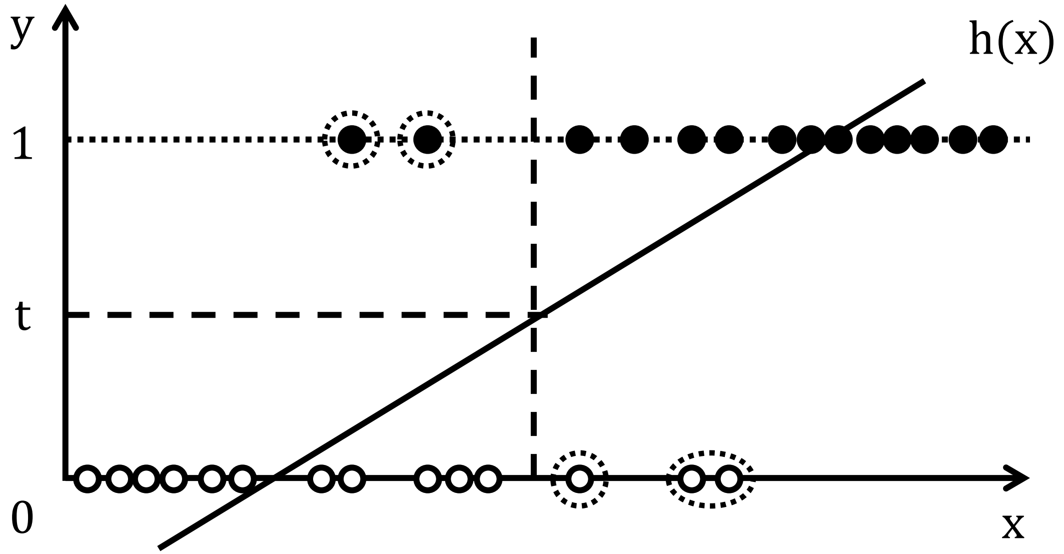
- Cost Functions
- Binary Classifier
- Performance Measures

Binary Classifier

Binary Classifier

- Observed response **y** takes only two possible values **+** and **–**
- Define relationship between ***h(x)*** and ***y***
- Use the decision rule: $\hat{y} = \begin{cases} +, & h(x) \geq t \\ -, & otherwise \end{cases}$





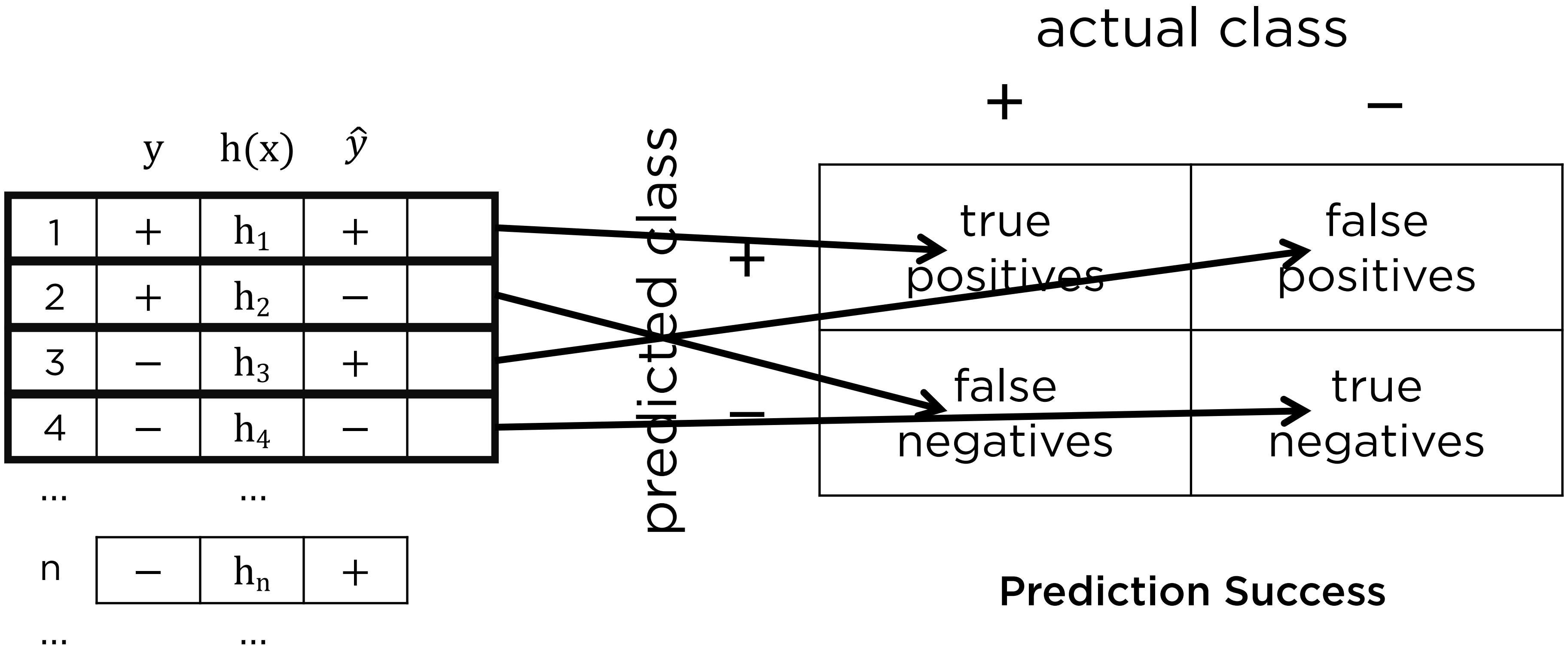
Performance Measures

Performance Measures

- Precision & Recall
- ROC Curve
- Gains & Lift

Performance Measures - Precision & Recall

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



Prediction Success

(Confusion Matrix)

Performance Measures - Precision & Recall

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

actual class

predicted class		+	−
	+	tp	fp
	−	fn	tn

Prediction Success

(Confusion Matrix)

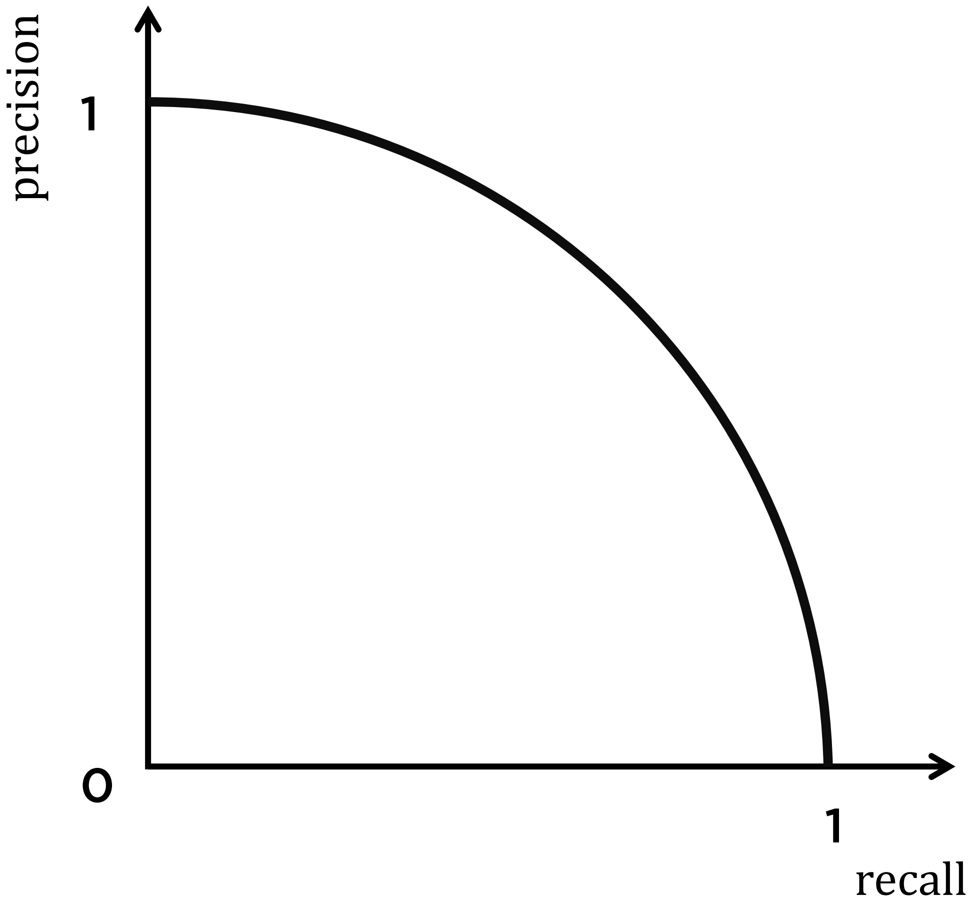
- Precision $\frac{tp}{tp+fp}$ ❯ 1

- Recall (Sensitivity) $\frac{tp}{tp+fn}$ ❯ 1

Performance Measures - Precision & Recall

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

  - Precision $\frac{tp}{tp + fp}$
  - Recall (Sensitivity) $\frac{tp}{tp + fn}$



Performance Measures

- Precision & Recall
- ROC Curve
- Gains & Lift

Performance Measures - ROC Curve

		actual class +	actual class –
predicted class	+	true positives	false positives
	–	false negatives	true negatives

- Recall (Sensitivity) $\frac{\boxed{tp}}{\boxed{tp + fn}}$
- Specificity $\frac{\boxed{tn}}{\boxed{tn + fp}}$

Performance Measures - ROC Curve

- Recall (Sensitivity) $\dfrac{tp}{tp + fn}$

- Specificity $\dfrac{tn}{tn + fp}$

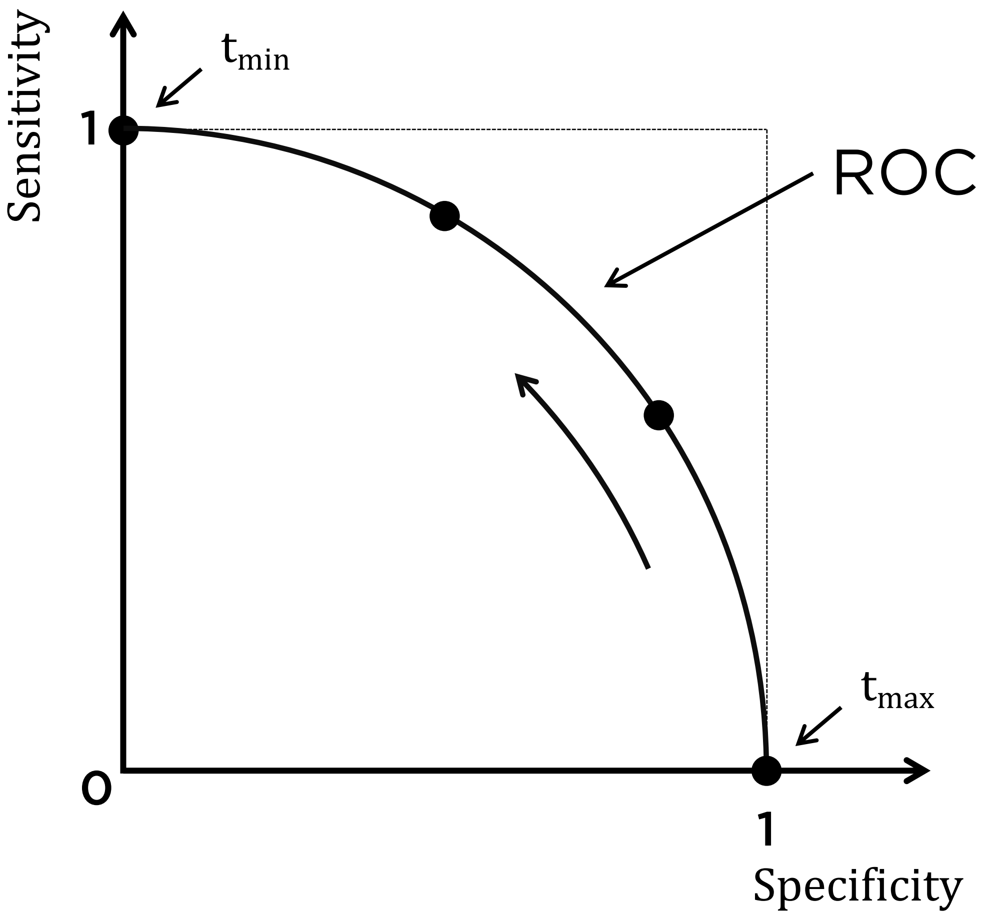
y	$h(x)$	$\hat{y}$	
$+$	h_1		max
$+$	h_2		
$-$	h_3		
$-$	h_4		
$+$	h_5		
$-$	h_6		
$+$	h_7		
$-$	h_8		
$-$	h_9		min

Performance Measures - ROC Curve

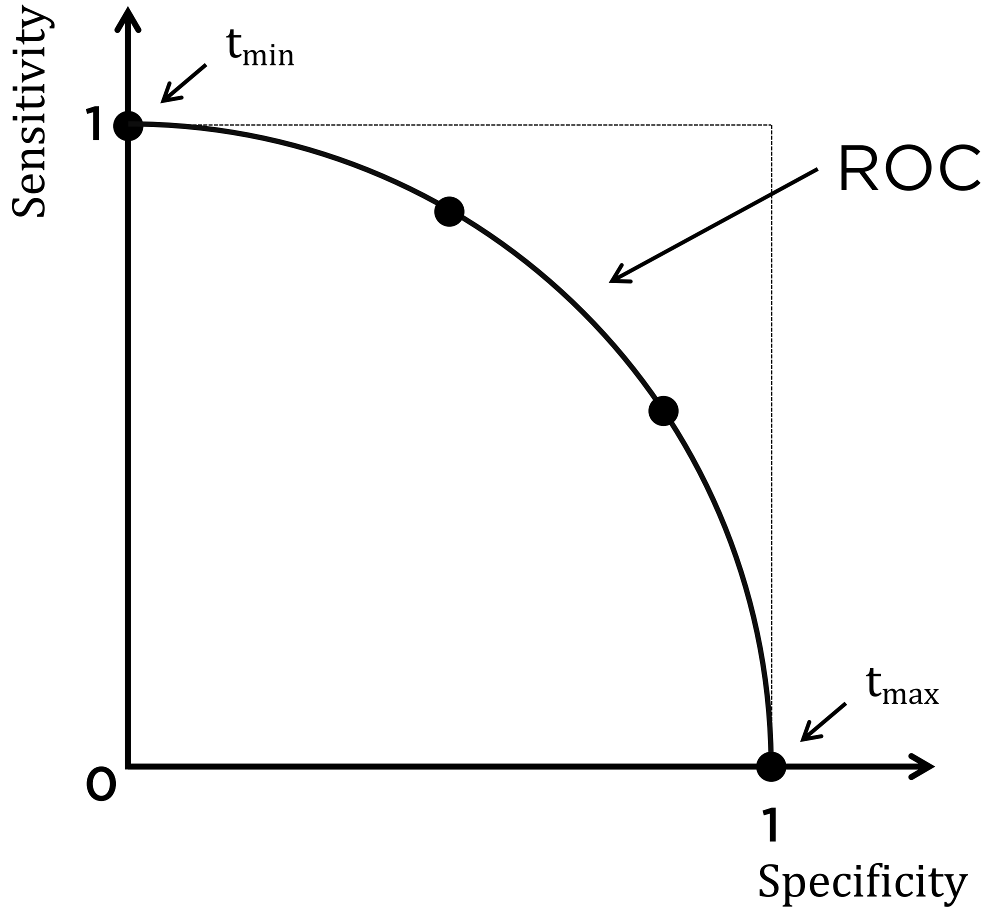
- Recall (Sensitivity) $\dfrac{tp}{tp + fn}$
- Specificity $\dfrac{tn}{tn + fp}$

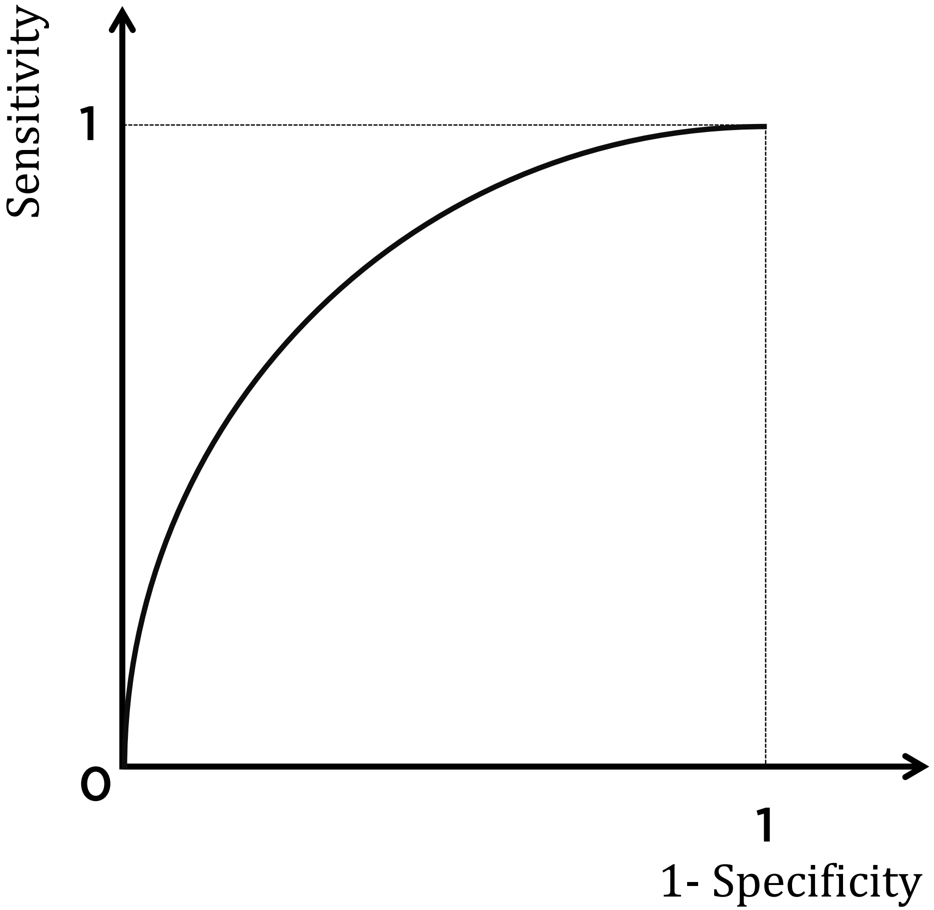
y	h(x)	$\hat{y}$
+	h_1	+
+	h_2	+
−	h_3	+
−	h_4	+
+	h_5	−
−	h_6	−
+	h_7	−
−	h_8	−
−	h_9	−

$\leftarrow t_{intermedia}$



Performance Measures - ROC Curve





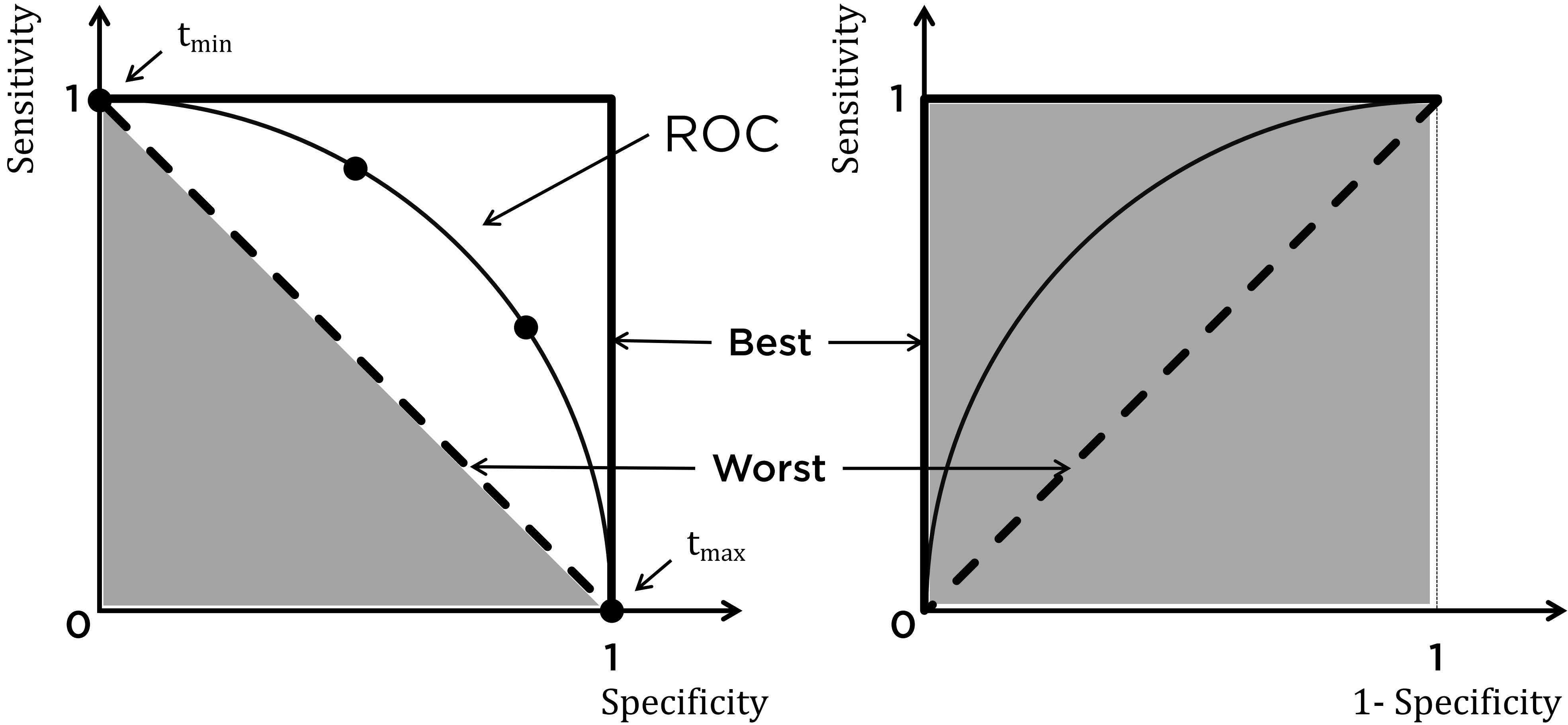
Performance Measures - ROC Curve

$$0.5 \leq AUC \leq 1.0$$

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

t_{max}

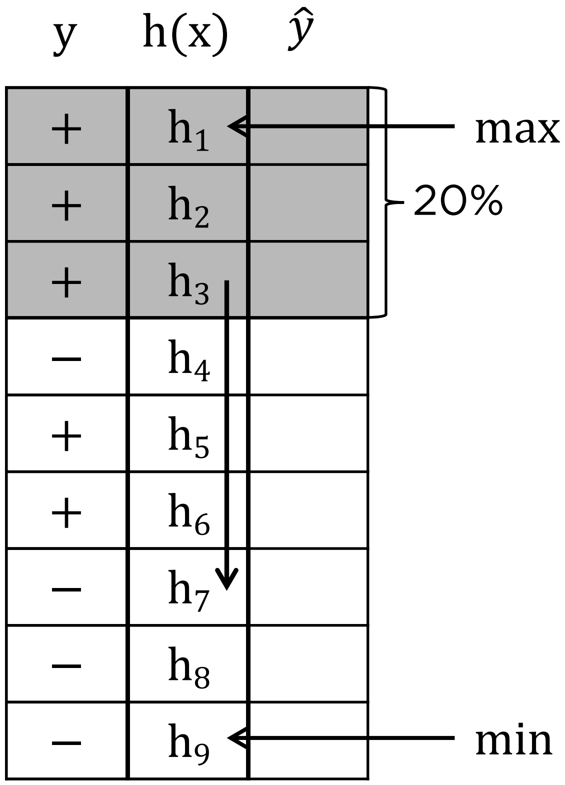
t_{min}



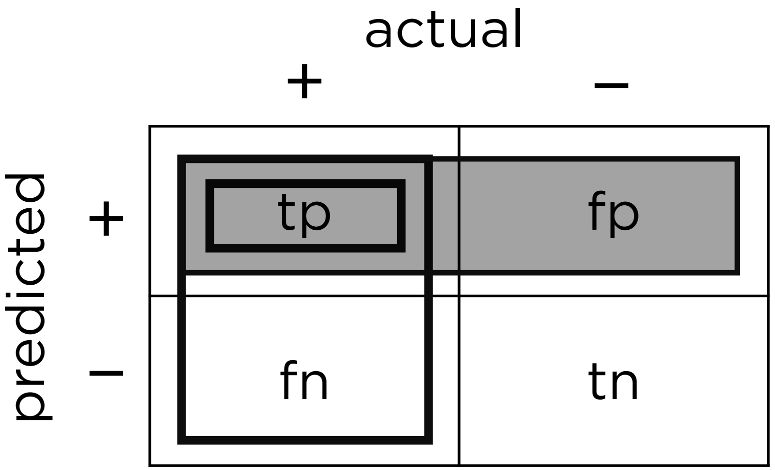
Performance Measures

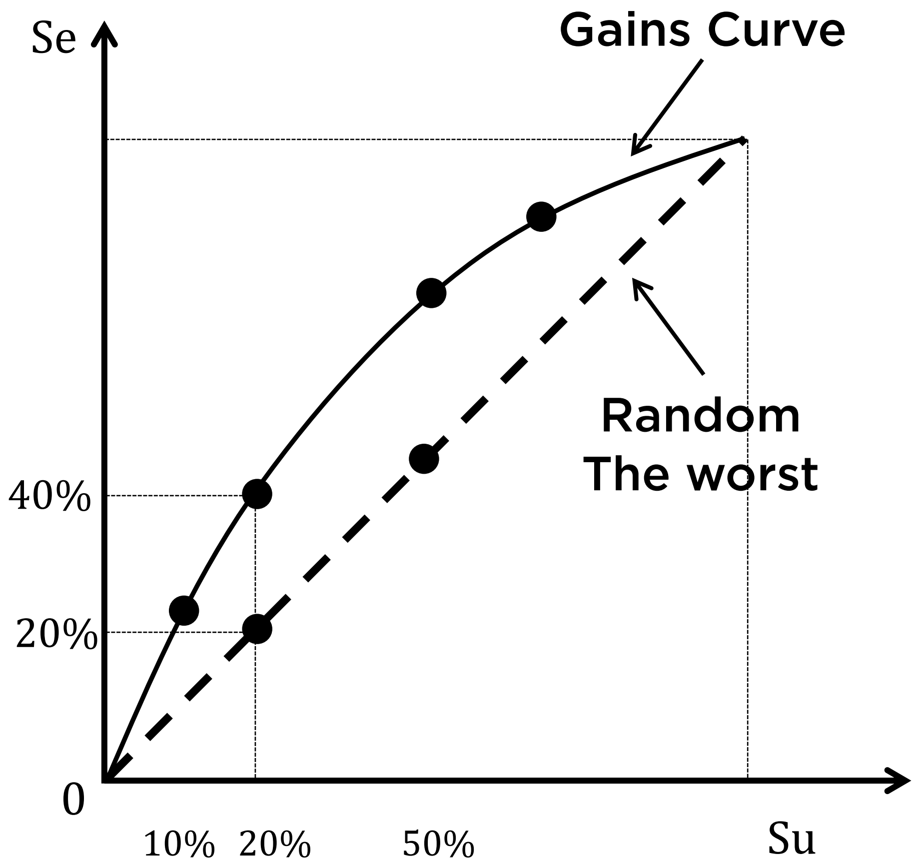
- Precision & Recall
- ROC Curve
- Gains & Lift

Performance Measures - Gains & Lift



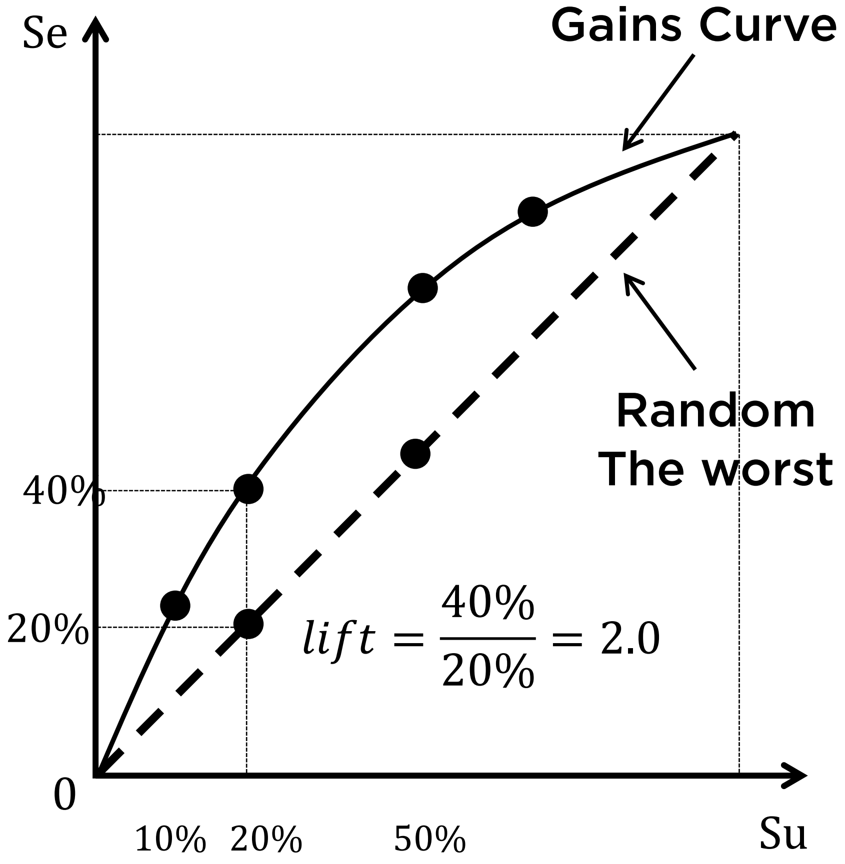
- Sensitivity (Recall) $$Se = \frac{tp}{tp + fn}$$
- Support (% pop) $$Su = \frac{tp + fp}{n}$$

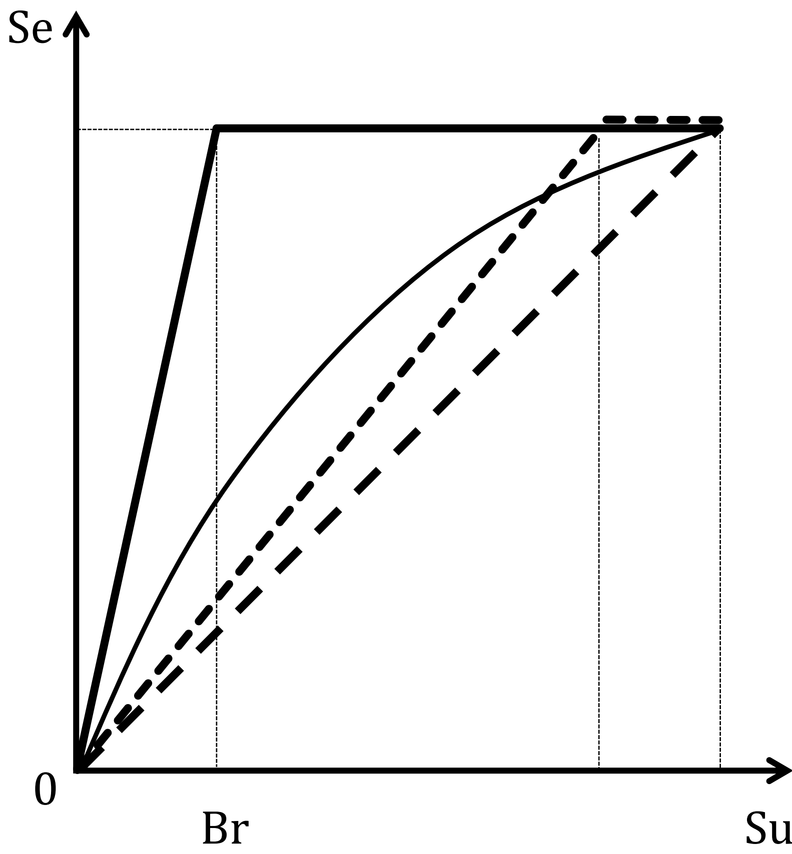




Performance Measures - Gains & Lift

- Sensitivity (Recall) $Se = \dfrac{tp}{tp + fn}$
- Support (% pop) $Su = \dfrac{tp + fp}{n}$
- Base Rate $Br = \dfrac{tp + fn}{n}$
- Gains: $\{Su, Se\}$
- Lift: $\{Su, \dfrac{Se}{Su}\}$
- ROC: $\{\dfrac{Su - Br \cdot Se}{1 - Br}, Se\}$





No summary