Evaluation

Division into training and test sets

- Fixed
  - Leave out random N% of the data
- k-fold Cross-Validation
  - Select K folds without replace
- Leave-One-Out Cross Validation
  - Special case
- Related: Bootstrap
  - Generate new training sets by sampling with replacement

Leave-one-out cross-validation (LOOCV)

For each training example (x_i, y_i)

Train a classifier with all training data except (x_i, y_i)

Test the classifier's accuracy on (x_i, y_i)

LOOCV accuracy = average of all n accuracies

Twitter Assignment:

To evaluate your term-sentiment model, one recommendation was to
1) leave out one term from AFINN.txt,
2) let your model assign it a sentiment score, and
3) compare your score with the AFINN score.
Do this for many terms, and you get a pretty good idea of the effectiveness of your model

Accuracy

Confusion Matrix

	Predicted +	Predicted -
True +	a	b
True -	c	d

Accuracy = (a+d)/(a+b+c+d)

Evaluation: Accuracy isn't always enough

- How do you interpret 90% accuracy?
  - You can't; it depends on the problem
- Need a baseline:
  - Base Rate
    - Accuracy of trivially predicting the most-frequent class
  - Random Rate
    - Accuracy of making a random class assignment
    - Might apply prior knowledge to assign random distribution
  - Naïve Rate
    - Accuracy of some simple default or pre-existing model
    - Ex: "All females survived"

Confusion Matrix

	Predicted +	Predicted -
True +	**a**	**b**
True -	**c**	**d**

$$lift = \frac{a/(a+b)}{(a+c)/(a+b+c+d)}$$

$$lift = \frac{\%\ positives > threshold}{\%\ dataset > threshold}$$

Confusion Matrix

	Predicted +	Predicted -
True +	a	b
True -	c	d

Precision: *a/(a+c)*

Recall: *a/(a+b)*

ROC Plot

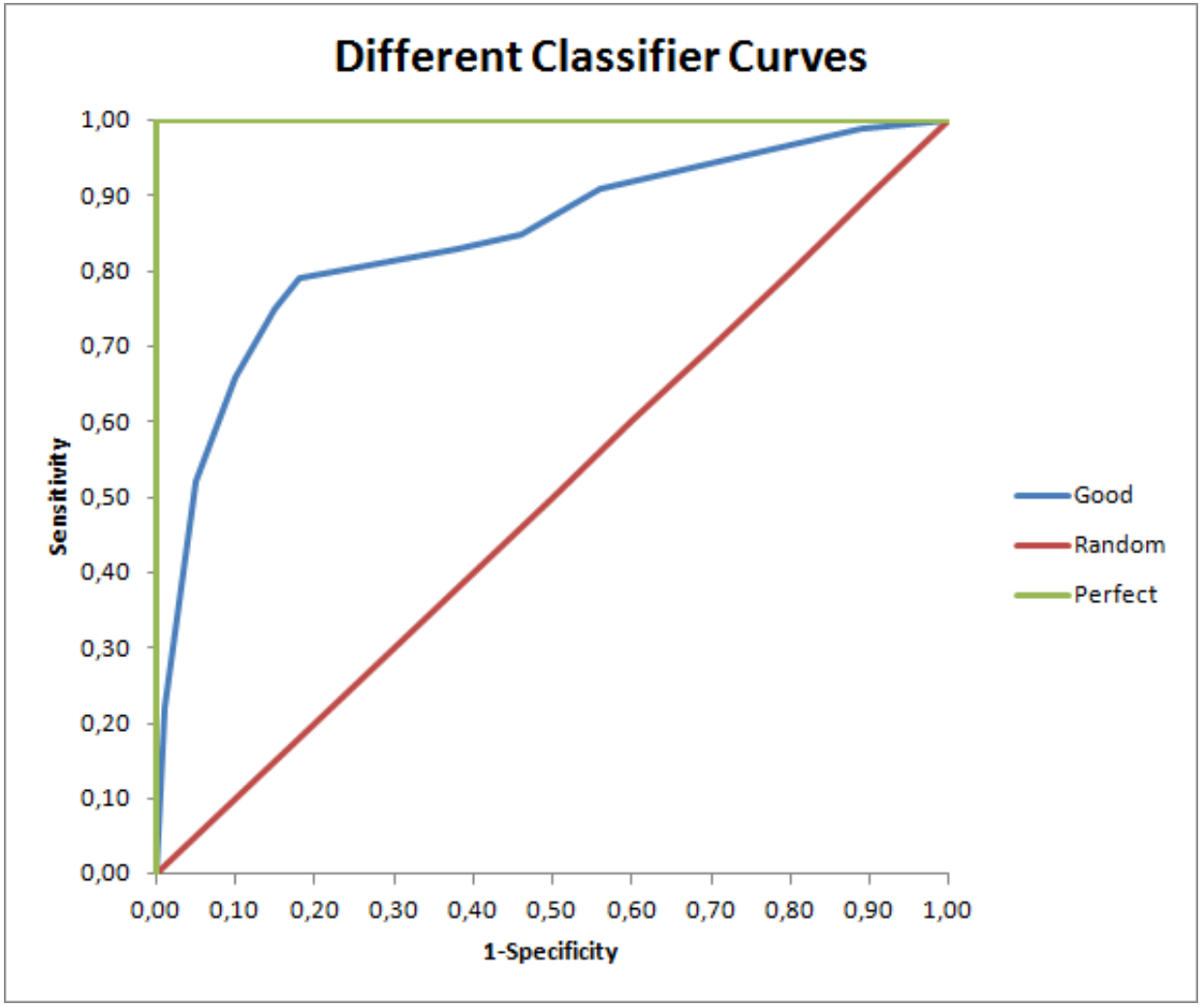
- "Receiver Operator Characteristic"
  - Historical term from WW2
  - Used to measure accuracy of radar operators

Confusion Matrix

	Predicted +	Predicted -
True +	a	b
True -	c	d

Sensitivity = a/(a+b)

1 - Specificity = 1 - d/(c+d)



src: Cesar Souza