CLASSIFICATION

JELENA JOVANOVIĆ

Email: jeljov@gmail.com

Web: http://jelenajovanovic.net

OUTLINE

- What is classification?
- Binary and multiclass classification
- Classification algorithms
- Performance measures for classification models

WHAT IS CLASSIFICATION?

- A supervised learning task of determining the class of an instance; it is assumed that:
  - feature values for the given instance are known
  - the set of possible classes is known and given
- Classes are given as nominal values; for instance:
  - classification of email messages: spam, not-spam
  - classification of news articles: politics, sport, culture i sl.

BINARY AND MULTICLASS CLASSIFICATION

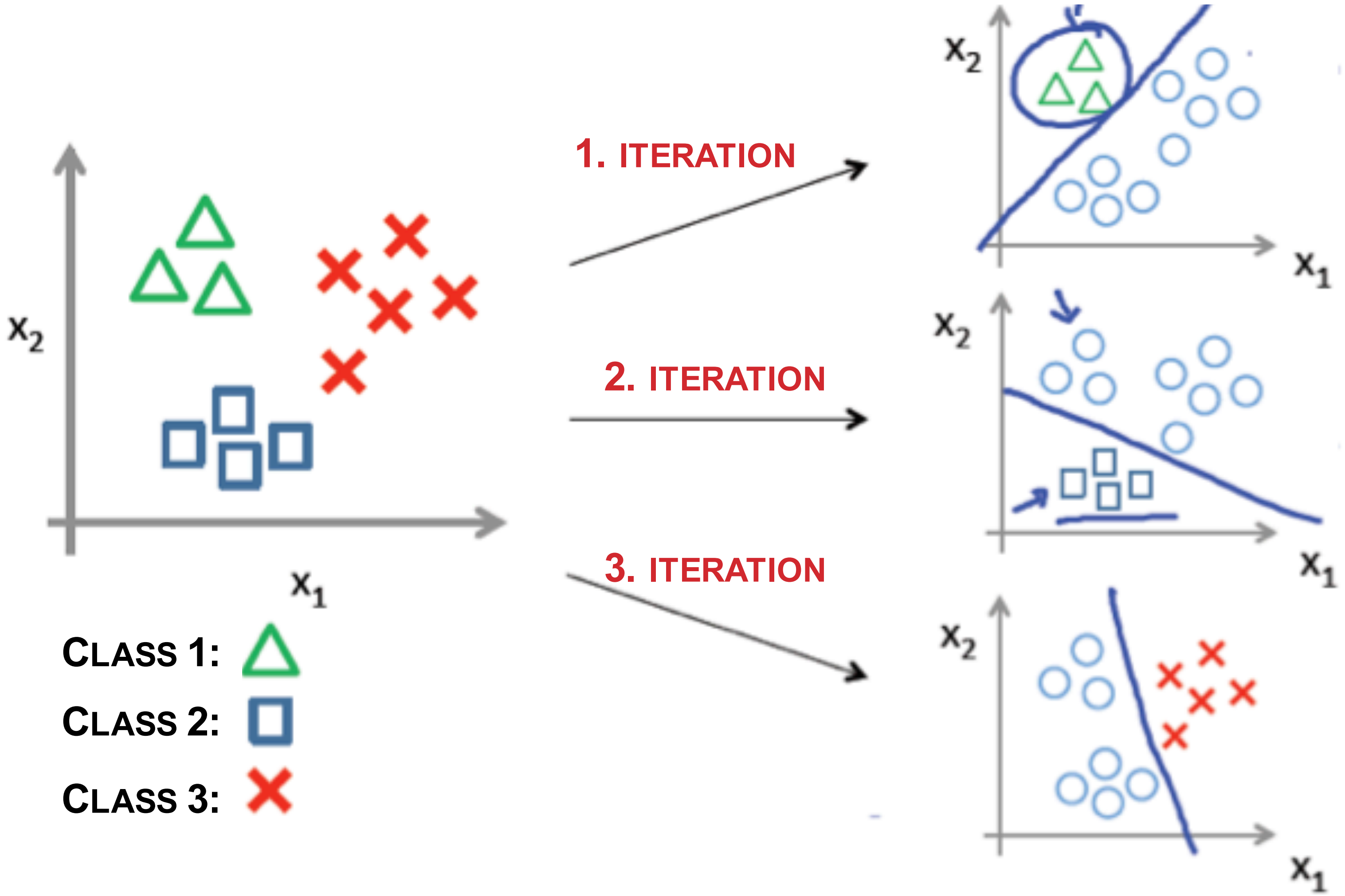
Based on the number of classes, classification can be:

- *binary* – instances should be classified into 2 classes
- *multiclass* – more than 2 classes are used for classifying instances

In both cases, a classifier works in a rather similar manner:

In multiclass classification, the classifier learns iteratively, so that in each iteration, it learns to differentiate instances of one class from all the other instances

MULTICLASS CLASSIFICATION



Source: https://www.coursera.org/learn/machine-learning

CLASSIFICATION ALGORITHMS

There are numerous classification models/algorithms:

- Logistic regression
- Naïve Bayes
- Algorithms from the Decision trees family
- Algorithms from the Neural networks family
- k-Nearest Neighbor (kNN)
- Support Vector Machines (SVN)
- …

PERFORMANCE MEASURES

The most frequently used metrics:

- Confusion Matrix
- Accuracy
- Precision and Recall
- F measure
- Area Under the ROC Curve

Confusion Matrix

Serves as the basis for calculating other performance measures

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

TP = True Positive

FP = False Positive

TN = True Negative

FN = False Negative

ACCURACY

Accuracy is the percentage of correctly classified instances

Accuracy = (TP + TN) / N

where:

- TP – True Positive; TN – True Negative
- N – the total number of instances in the dataset

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

ACCURACY

In the case of highly unequal distribution of instances across classes (so called *skewed* classes), this measure is unreliable

An example:

- in the case of message classification as spam vs. not-spam, the training set might contain 0.5% of spam messages
- if we apply a biased classifier that classifies each message as not-spam, we get very high accuracy – 99.5%
- obviously, this metric is unreliable and in the case of skewed classes, other metrics are needed

Precision and Recall

Precision = TP / # predicted positive = TP / (TP + FP)

Example: out of all the messages *marked as spam*, the percentage of those that are *really spam* messages

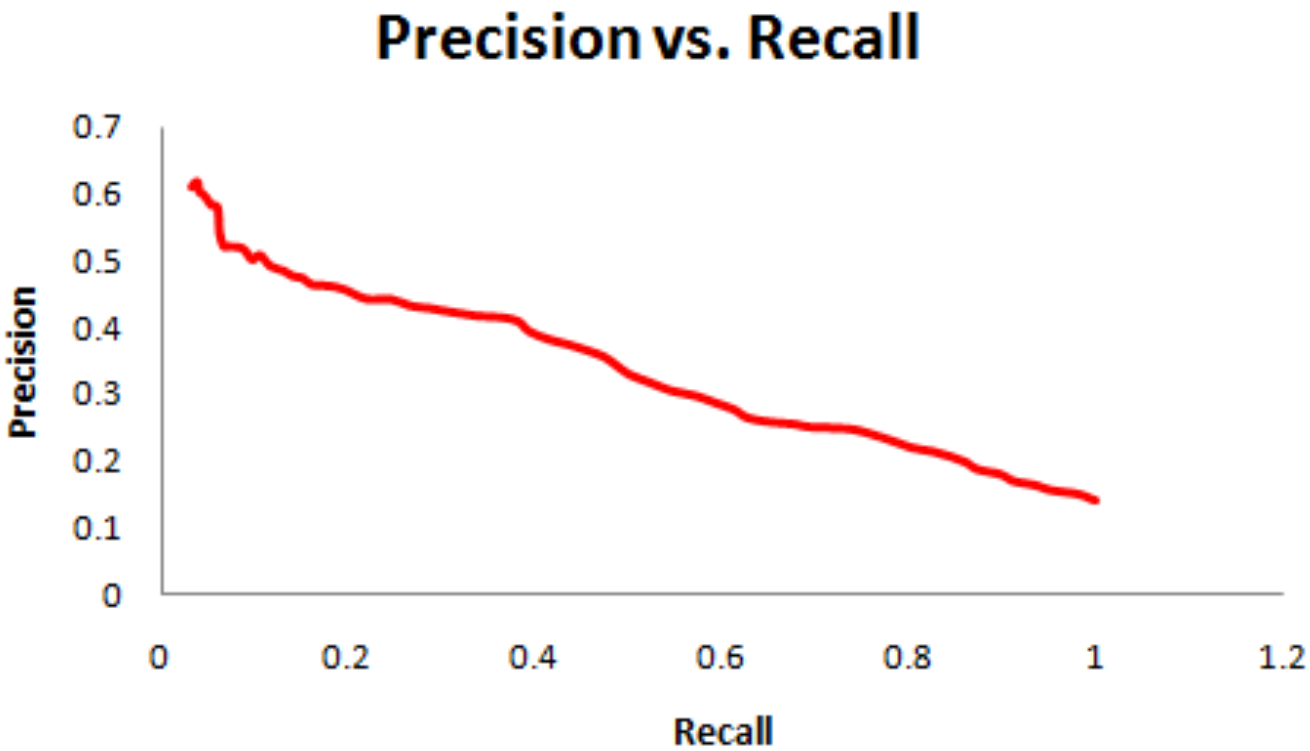
Recall = TP / # actual positive = TP/ (TP + FN)

Example: out of all the messages that are *really spam*, the percentage of those that have been *detected/classified as spam*

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

PRECISION VS. RECALL

In practice, one always needs to make a compromise between these two metrics: by increasing Recall, we decrease (though unwillingly) Precision, and vice versa



Source: http://groups.csail.mit.edu/cb/struct2net/webserver/images/prec-v-recall-v2.png

F MEASURE

F measure combines Precision and Recall and allows for easier comparison of two or more algorithms

$$F = (1 + \beta^2) * \text{Precision} * \text{Recall} / (\beta^2 * \text{Precision} + \text{Recall})$$

Parameter β controls the extent to which we want to favor Recall over Precision

In practice, F1 measure is typically used; it is called “balanced“ F measure as it equally weights Precision and Recall:

$$F1 = 2 * \text{Precision} * \text{Recall} / (\text{Precision} + \text{Recall})$$

AREA UNDER THE ROC* CURVE (AUC)

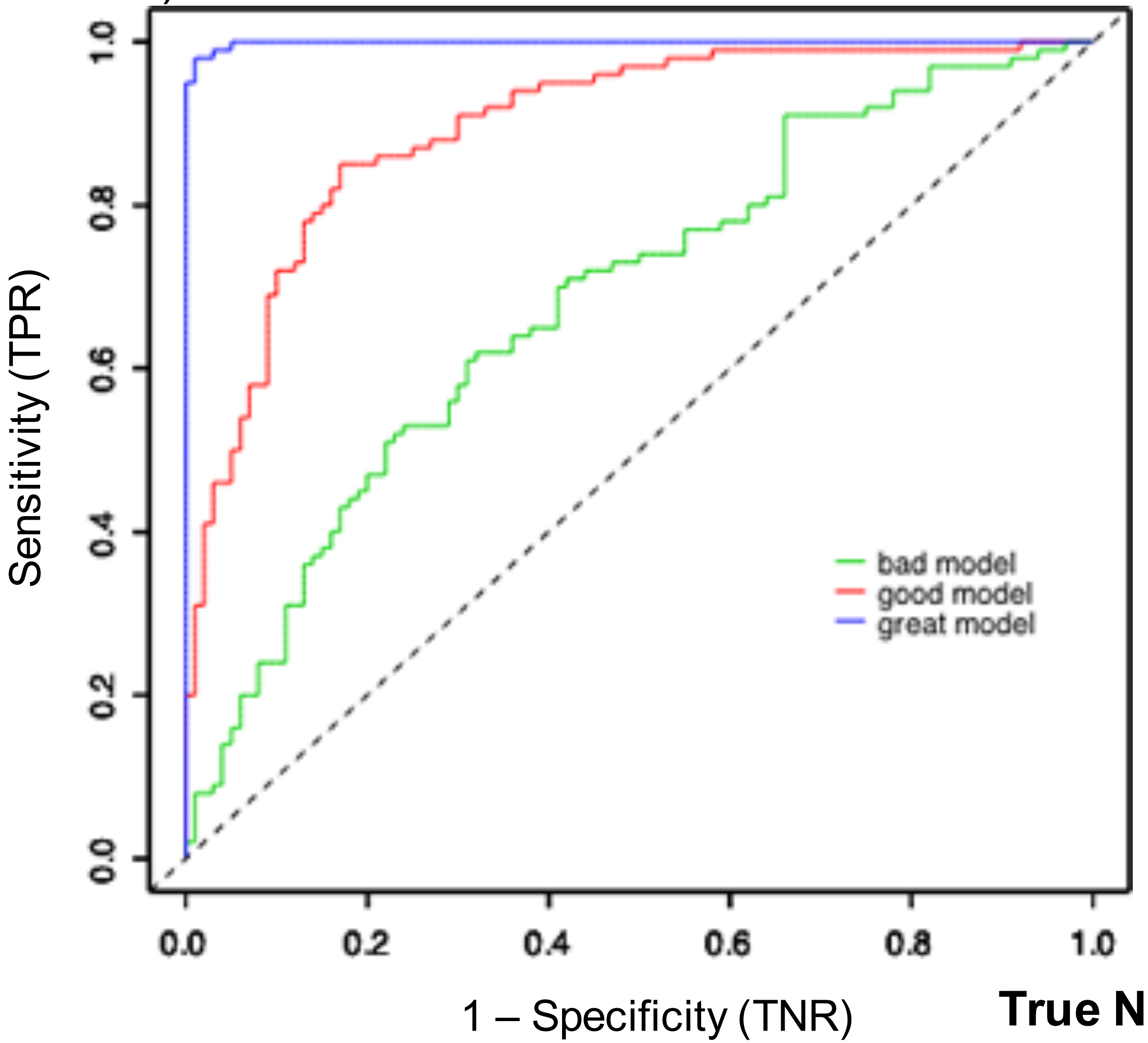
- It measures discriminatory power of a classifier, i.e., its ability to correctly differentiate instances of different classes
- It is used for measuring performance of binary classifiers
- It takes values from the 0-1 interval
- In the case of random classification, AUC = 0.5; so, as the AUC value is greater than 0.5, the classifier is better
  - 0.7–0.8 is considered fair; 0.8–0.9 good; > 0.9 excellent

*ROC = Receiver Operating Characteristic;
http://en.wikipedia.org/wiki/Receiver_operating_characteristic

AREA UNDER THE ROC CURVE

Sensitivity or **True Positive Rate (TPR)**

TPR = TP/(TP + FN)



Specificity or **True Negative Rate (TNR)**

TNR = TN/(FP + TN)

Source: http://goo.gl/Aeauuh

ACKNOWLEDGEMENTS AND RECOMMENDATIONS

Acknowledgements and Recommendations

Machine Learning @ Stanford

- Coursera: https://www.coursera.org/learn/machine-learning
- Stanford YouTube channel: http://www.youtube.com/view_play_list?p=A89DCFA6ADACE599

RECOMMENDATIONS

- [article] Visual Introduction to Machine Learning: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- [blog post] Choosing a Machine Learning Classifier: http://blog.echen.me/2011/04/27/choosing-a-machine-learning-classifier/
- [article] IU scientists use Instagram data to forecast top models at New York Fashion Week (http://goo.gl/ovepjx)
- [podcast] Data Stories podcast #27; topic: "Big Data Skepticism" (http://goo.gl/KKPGuW)
  - the podcast mentioned a study that was aimed at the prediction of demographic characteristics of Facebook users based on their Likes (http://goo.gl/fykOyt)

(Anonymous) questionnaire for your comments, suggestions, critiques:

http://goo.gl/cqdp3I