San José State University

Math 251: Statistical and Machine Learning Classification

Evaluation Criteria (for binary classification)

Dr. Guangliang Chen

Outline of the presentation:

- **Evaluation criteria**
  - Precision
  - Recall/sensitivity
  - Specificity
  - F1 score
- **Main reference**[1]

[1] `http://cs229.stanford.edu/section/evaluation_metrics_spring2020.pdf`

Motivation

The two main criteria we have been using for evaluating classifiers are

- Accuracy or error (both overall)
- Running time

The overall accuracy does not reflect the classwise accuracy scores, and can be dominated by the largest class(es).

For example, with two imbalanced classes $(80 : 20)$, the constant prediction with the dominant label will achieve 80% accuracy overall.

In the setting of binary classification, where the data points only have two different labels, more performance metrics can be defined based on the confusion matrix.

Interpretation of the confusion matrix

The confusion table summarizes the 4 different combinations of true conditions and predicted labels:

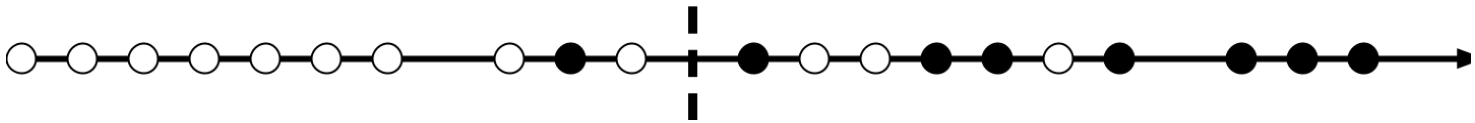
Predicted	Actual	
	Positive	Negative
Positive	TP	FP
Negative	FN	TN

(H_0: Test point is negative)
(H_1: Test point is positive)
FP: Type-I error
FN: Type-II error

The overall accuracy of the classifier is

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Example



Predicted	Actual	
	Positive	Negative
Positive	TP=7	FP=3
Negative	FN=1	TN=9

The overall accuracy of this classifier is

$$\frac{7+9}{7+3+1+9} = \frac{16}{20} = 0.8$$

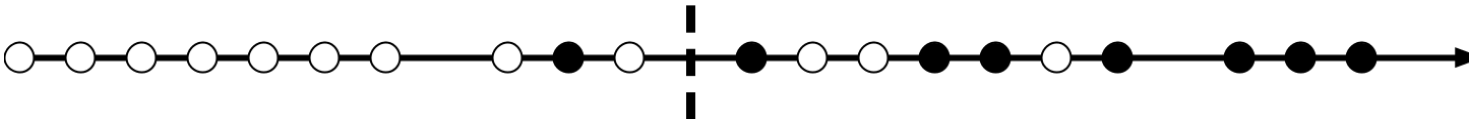
Remark. The overall accuracy is not a good measure when the classes are imbalanced.

Example. *In a cancer detection example with 100 people, only 5 people has cancer. Let's say our model is very bad and predicts every case as No Cancer. In doing so, it has classified those 95 non-cancer patients correctly and 5 cancerous patients as Non-cancerous. Now even though the model is terrible at predicting cancer, the accuracy of such a bad model is also 95%.*

More performance metrics

- Precision
- Recall or Sensitivity
- Specificity
- F1 score

Precision

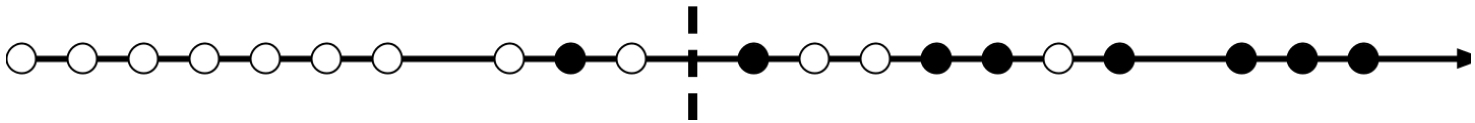


Predicted	Actual	
	Positive	Negative
Positive	TP=7	FP=3
Negative	FN=1	TN=9

The precision of this classifier is defined as

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{7}{7+3} = \frac{7}{10}$$

Recall or sensitivity



Predicted	Actual	
	Positive	Negative
Positive	TP=7	FP=3
Negative	FN=1	TN=9

The recall/sensitivity of this classifier is defined as

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{7}{7+1} = \frac{7}{8}$$

Evaluation criteria

When to use precision and when to use recall:

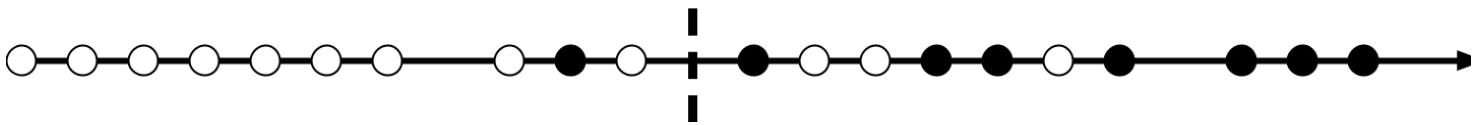
Precision is about being precise. So even if we managed to capture only one positive case, and we captured it correctly, then we are 100% precise.

Recall is about capturing all positive cases with the answer as positive. So if we simply always say every case as being positive, we have 100% recall.

Some common patterns:

- High precision is hard constraint, do best recall (search engine results, grammar correction): Intolerant to FP
- High recall is hard constraint, do best precision (medical diagnosis): Intolerant to FN

The precision-recall tradeoff



As we move the threshold from left to right, how do the precision and recall change?

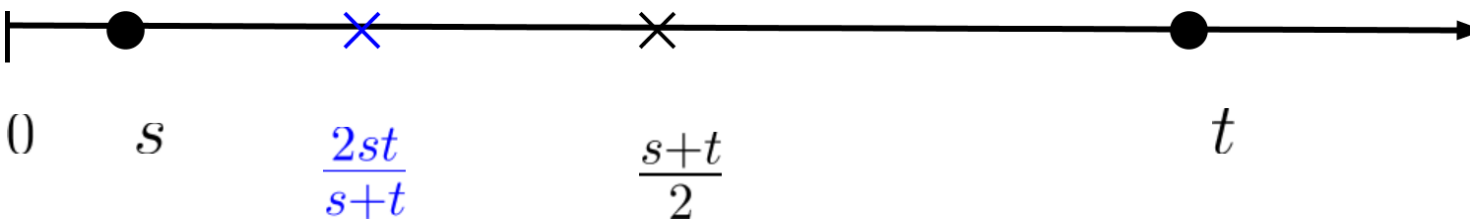
F1 score

Predicted	Actual		
	Positive	Negative	
Positive	TP=7	FP=3	Precision $= \frac{7}{10}$
Negative	FN=1	TN=9	
	Recall $= \frac{7}{8}$		

The F1 score is defined as the harmonic mean of precision and recall:

$$\text{F1 score} = \frac{1}{\frac{1}{2}\left(\text{precision}^{-1} + \text{recall}^{-1}\right)} = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} = \frac{14}{18}$$

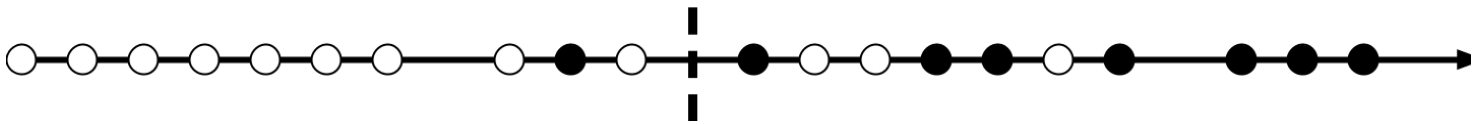
Remark. The harmonic mean of two different positive numbers s, t



is closer to the smaller number than to the larger number:

$$s = 0.2, t = 0.8: \quad \frac{1}{2}(s+t) = 0.5, \quad \frac{2st}{s+t} = 0.32$$

Specificity



Predicted	Actual	
	Positive	Negative
Positive	TP=7	FP=3
Negative	FN=1	TN=9

The specificity of this classifier is defined as

$$\text{Specificity} = \frac{TN}{FP + TN} = \frac{9}{3 + 9} = \frac{9}{12}$$

ROC and AUC (to be defined later)

The kNN classifier outputs a discrete label (there is no threshold).

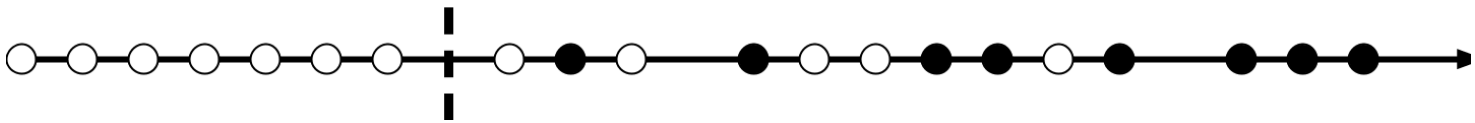
When the classifier outputs a continuous score (e.g., Bayes classification, logistic regression, and SVM), we can use the sensitivity and specificity to further define

- Receiver operating characteristic (ROC) curves
- Area under the curve (AUC)

by continuously changing the threshold.

We will introduce these two additional measures when we get to those methods.

Changing threshold



Predicted	Actual		
	Positive	Negative	
Positive	TP=8	FP=5	Precision $= \frac{8}{13}$
Negative	FN=0	TN=7	
	Recall $= 1$ (Sensitivity $= 1$)	Specificity $= \frac{7}{12}$	

Accuracy $= \frac{15}{20}$, and F1 score $= \frac{2 \cdot 1 \cdot \frac{8}{13}}{1+\frac{8}{13}} = \frac{16}{21}$

Receiver operating characteristic (ROC) curves

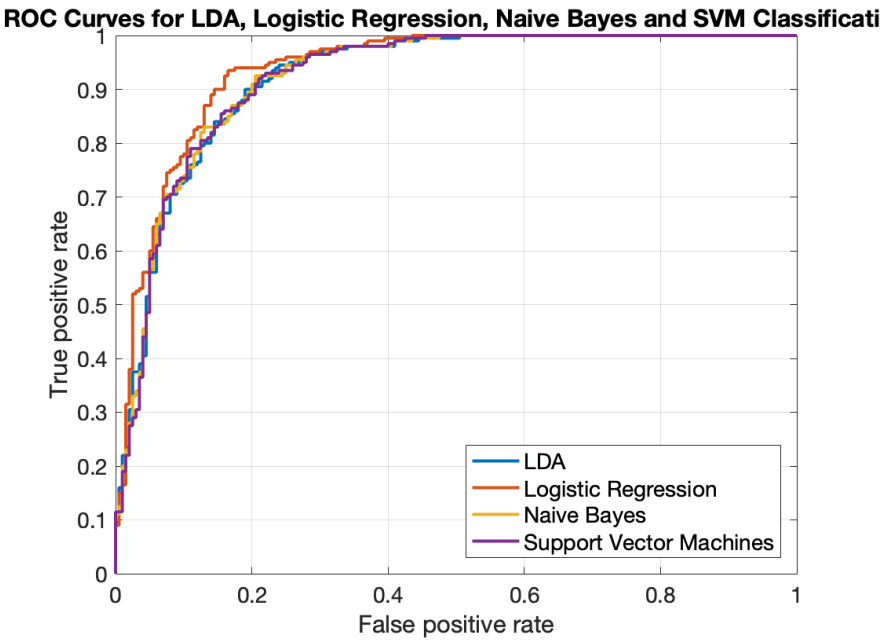
An ROC curve is a graphical plot of the true positive rate

$$TPR = \frac{TP}{TP + FN} = \text{Sensitivity}$$

against the false positive rate

$$FPR = \frac{FP}{FP + TN} = 1 - \text{Specificity}$$

at various threshold settings. It illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

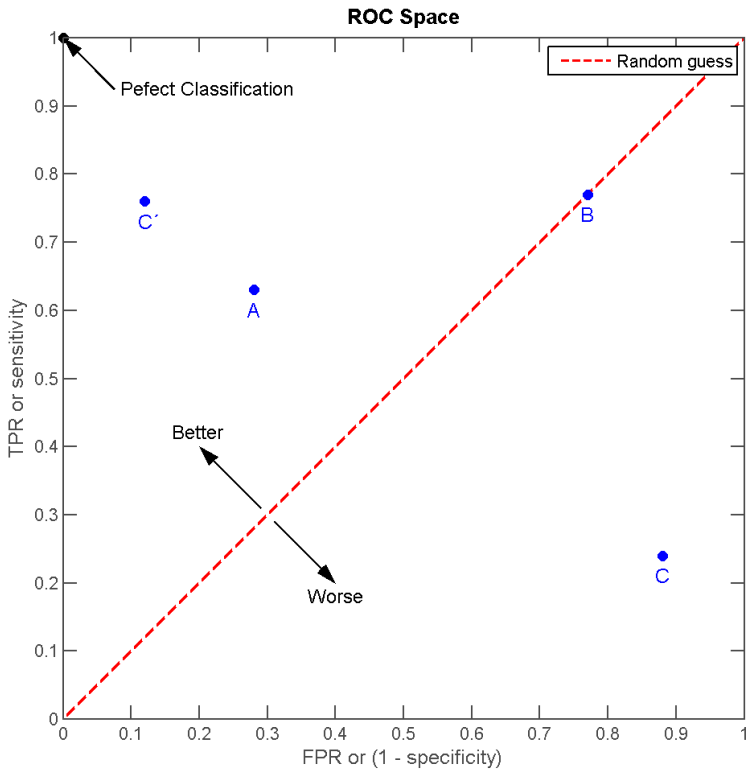


"power vs the Type I error"

Evaluation criteria

Each point in the ROC space corresponds to a unique confusion table (due to a different threshold used in making predictions):

- $(0, 1)$: **perfect classification**
- Diagonal: random guess
- Points above the diagonal represent good classification results (better than random); points below represent bad results (worse than random).



(source: Wikipedia)

Area under the curve (AUC)

The area under the curve (often referred to as simply the AUC) is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one (assuming 'positive' ranks higher than 'negative'):

$$Area = P(X_1 > X_0)$$

where X_1 is the score for a positive instance and X_0 is the score for a negative instance.

The ROC AUC statistic is commonly used for model comparison in the machine learning community. $\longleftarrow$ The larger, the better

Multiclass classifications

Still has the confusion table (with large diagonal entries preferred).

Most metrics (except accuracy) generally analyzed as multiple 1-vs-many.

Assignment 2 (cont'd)

4. Consider the fashion MNIST data set. Pick one training class as your target and merge the other nine training classes to form the second, much bigger class. Apply the kNN classifier with $k = 1 : 12$ with the now-binary training data to classify the test data points correspondingly into the two classes. Plot the test accuracy and F1 score curves, both against k, together in one graph and comment on the plot.