Precept 3: Evaluation metrics and feature selection for classification

COS424/524/SML302 Spring 2021

Xiaoyan Li

Topics:

- Evaluation metrics for classification
  - Accuracy/error, confusion matrix,
  - ROC curve, precision recall curve
- Feature extraction
- Feature selection
  - Frequency based approaches: document frequency, Categorical Proportional Difference (PD)
  - Chi-square feature selection
  - Mutual information
- Paper for COS524

Is accuracy enough for evaluating classifiers?

- Accuracy
  - the number of correctly classified samples/the total number of samples
- Error:
  - Error = 1-accuracy

Confusion Matrix for binary classification

	True class labels	
	Positive	Negative
Predicted Positive	True Positive(**TP**)	False Positive(**FP**)
Predicted Negative	False Negative(**FN**)	True Negative(**TN**)
	P=TP+FN	**N=FP+TN**

Classifier A: Accuracy = 0.9

TP **45**	**FP** **5**
FN **5**	**TN** **45**

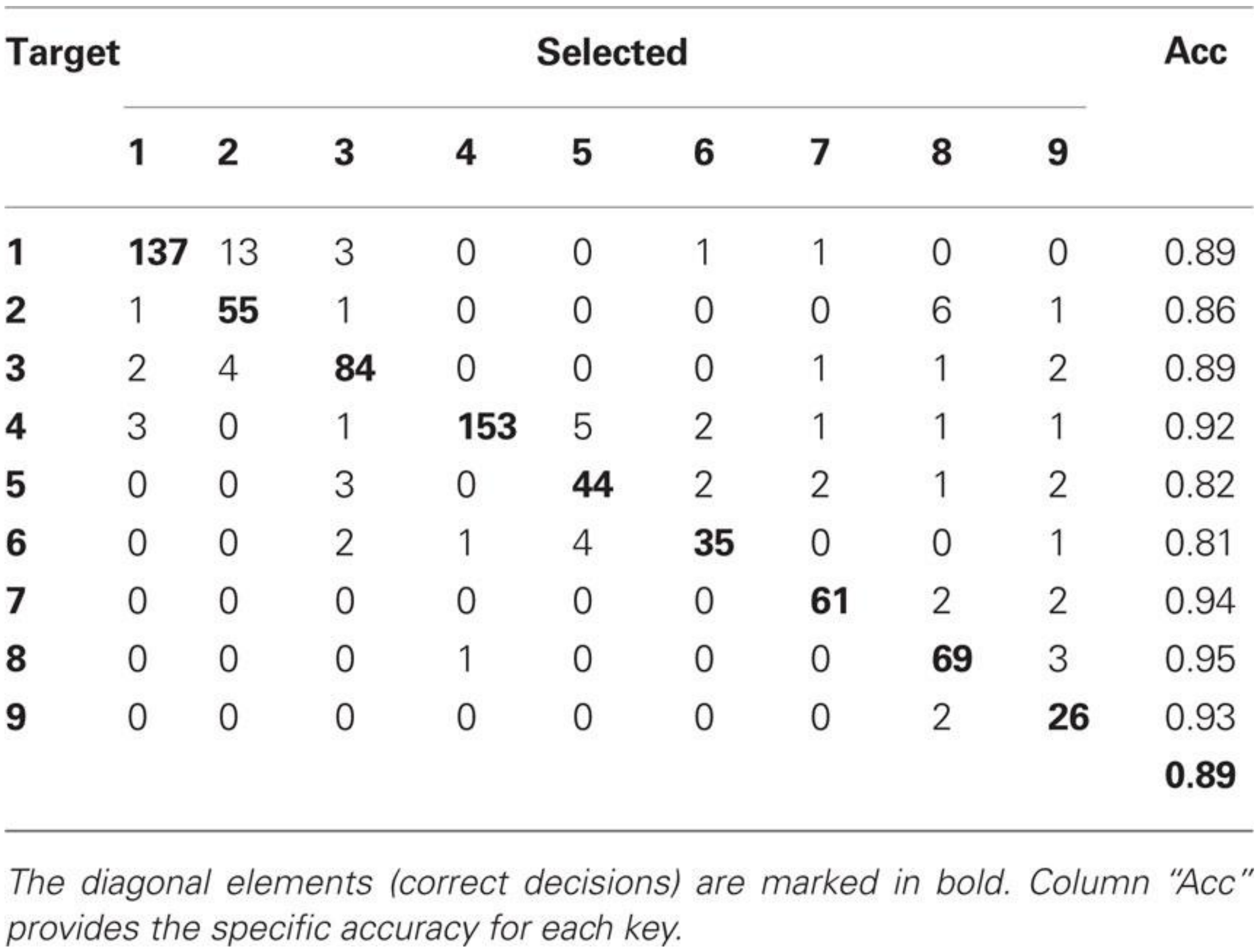
Classifier B: Accuracy = 0.9

TP **49**	**FP** **9**
FN **1**	**TN** **41**

Q1: Which classifier do you prefer for cancer diagnosis, A or B?

Q2: Total number of samples? # of positive samples? # of negative samples?

Confusion Matrix for a multiclass classification

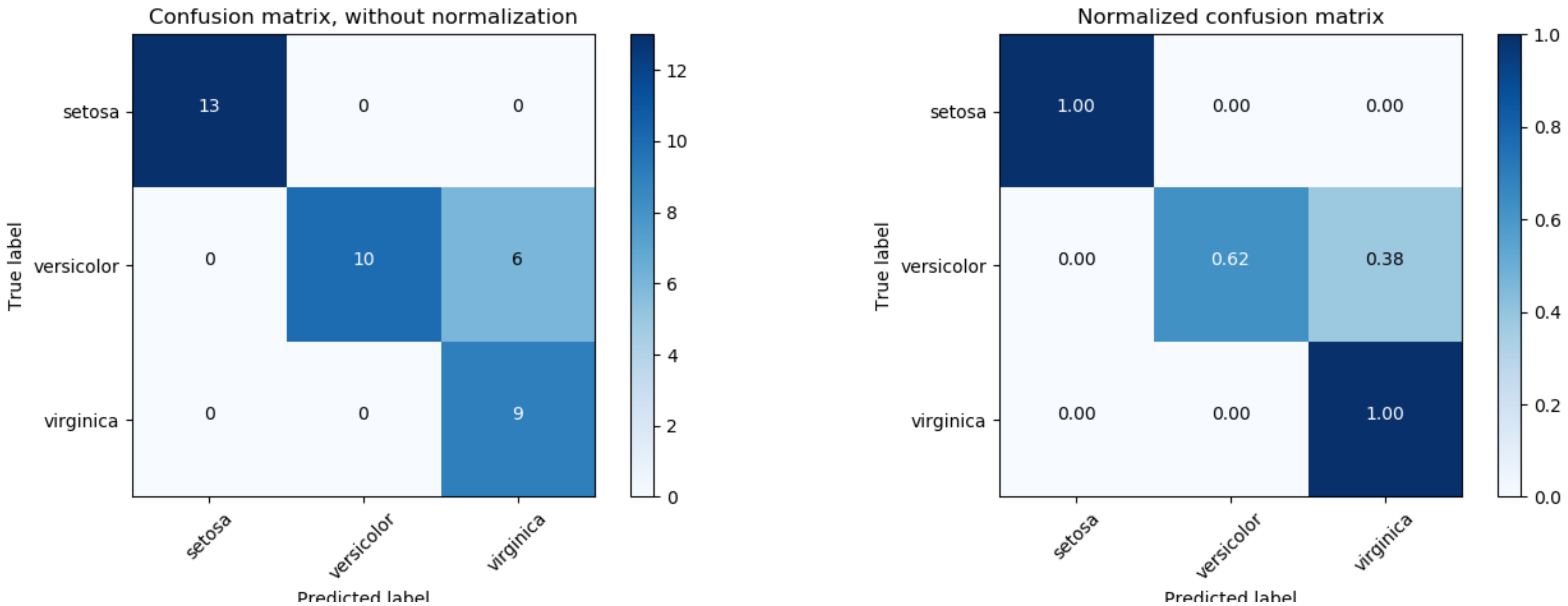


Target	**Selected**									**Acc**
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
1	**137**	13	3	0	0	1	1	0	0	0.89
2	1	**55**	1	0	0	0	0	6	1	0.86
3	2	4	**84**	0	0	0	1	1	2	0.89
4	3	0	1	**153**	5	2	1	1	1	0.92
5	0	0	3	0	**44**	2	2	1	2	0.82
6	0	0	2	1	4	**35**	0	0	1	0.81
7	0	0	0	0	0	0	**61**	2	2	0.94
8	0	0	0	1	0	0	0	**69**	3	0.95
9	0	0	0	0	0	0	0	2	**26**	0.93
										0.89

The diagonal elements (correct decisions) are marked in bold. Column "Acc" provides the specific accuracy for each key.

https://stats.stackexchange.com/questions/179835/how-to-build-a-confusion-matrix-for-a-multiclass-classifier

Confusion Matrix for a multiclass classification



Example of confusion matrix usage to evaluate the quality of the output of a classifier on the iris data set. (sklearn.metrics.confusion_matrix)
http://scikit-learn.org/stable/auto_examples/model_selection/plot_confusion_matrix.html#sphx-glr-auto-examples-model-selection-plot-confusion-matrix-py

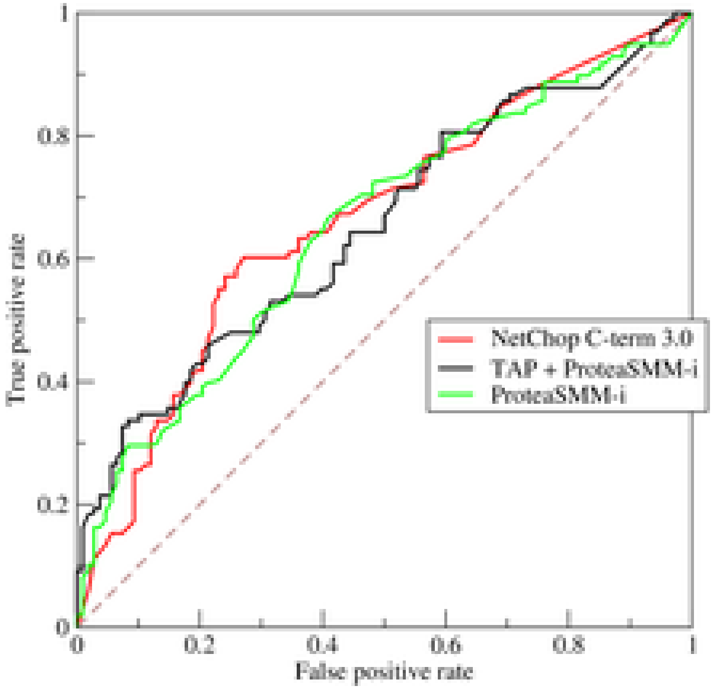
Other evaluation metrics

True class labels

	Positive	Negative
Predicted Positive	True Positive(**TP**)	False Positive(**FP**)
Predicted Negative	False Negative(**FN**)	True Negative(**TN**)
	P=TP+FN	**N=FP+TN**

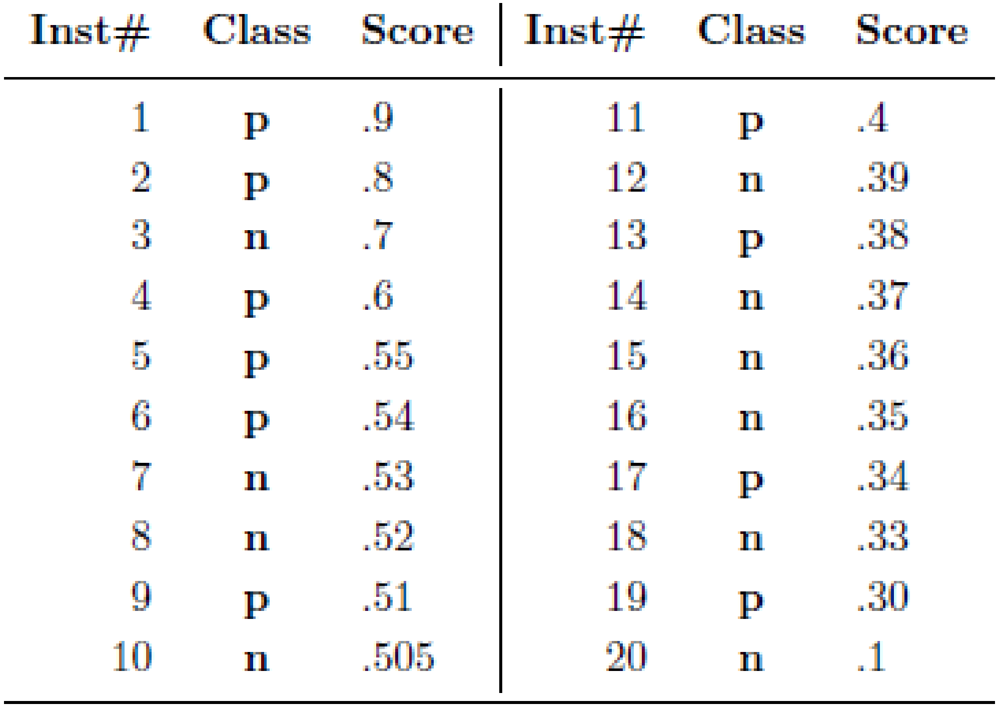
- Precision = TP/(TP+FP),
- Recall=Sensitivity=True Positive Rate = TP/P
- Specificity = True Negative Rate = TN/N
- False Positive Rate = FP/N,
- False Positive Rate = 1-Specificity
- F1 score = 2*precision*recall/(precision + recall)

Receiver Operating Characteristic(ROC) curves



- Plot the false positive rate(x-axis) vs. the true positive rate(y-axis) as the prediction threshold varies across the full range of possible values. (Graph taken from wikipedia)

How to create a ROC curve?

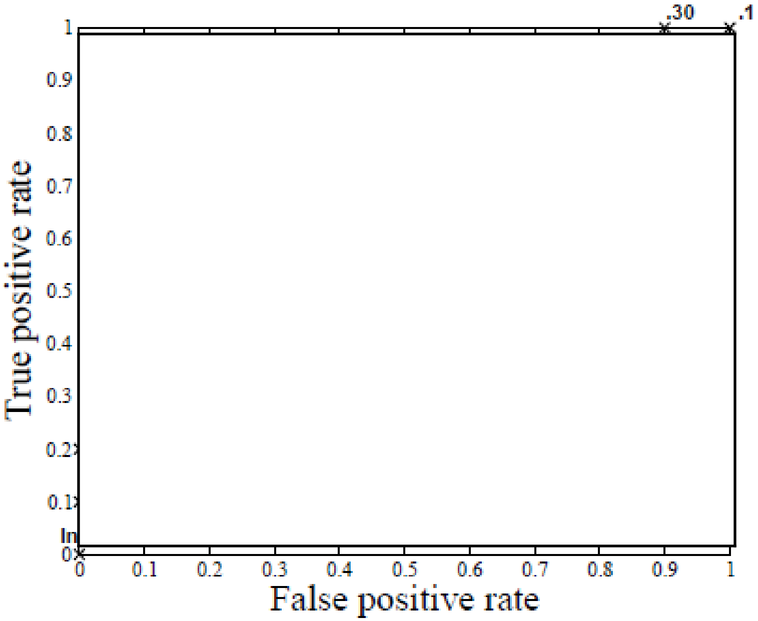


Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1

1. Sort samples by prediction scores generated by some classifier.
2. Vary the prediction threshold to make predictions and draw a point on the graph for each threshold.
3. Connect all the points.

How to create a ROC curve?

Inst#	Class	Score	Inst#	Class	Score
1	p	.9	11	p	.4
2	p	.8	12	n	.39
3	n	.7	13	p	.38
4	p	.6	14	n	.37
5	p	.55	15	n	.36
6	p	.54	16	n	.35
7	n	.53	17	p	.34
8	n	.52	18	n	.33
9	p	.51	19	p	.30
10	n	.505	20	n	.1



Q: Where is the point located for the following threshold:

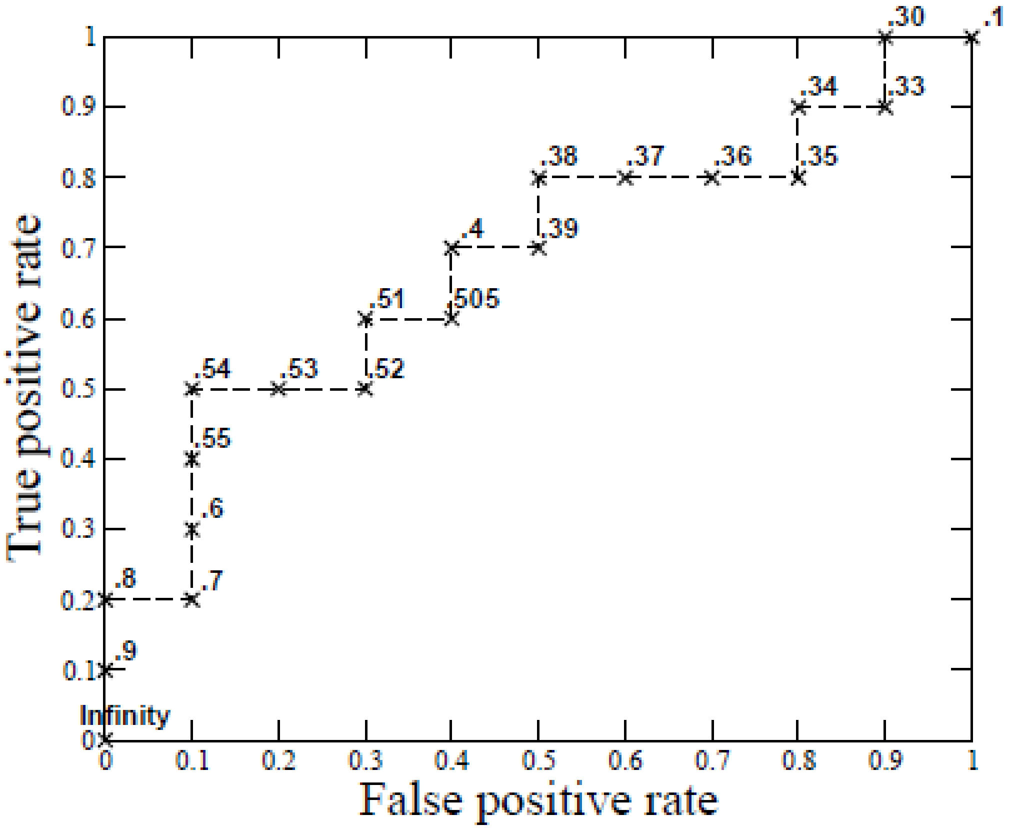
Threshold = 1 ?

Threshold = 0.9?

Threshold = 0.5?

Threshold = 0.1

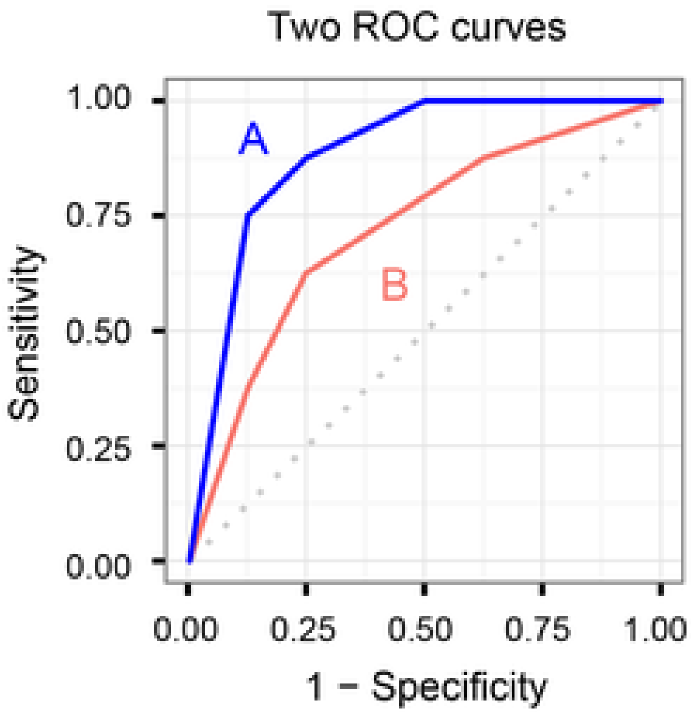
How to create a ROC curve?



Q1: What does the curve look like if predicting at random?

Q2: Where is the ROC cure for a perfect classifier?

Compare two classifiers using ROC curves

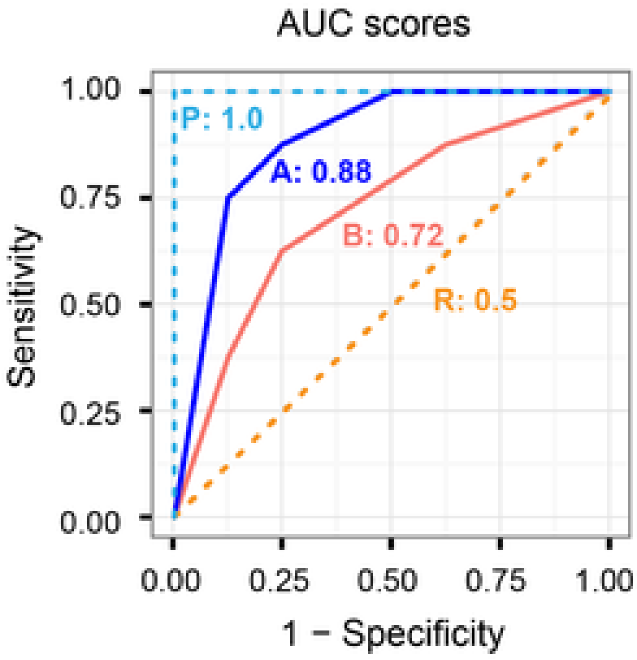


Two ROC curves represent the performance levels of two classifiers A and B. Classifier A clearly outperforms classifier B in this example.

when no curves cross each other. Curves close to the perfect ROC curve have a better performance level than the ones closes to the baseline.

Q: What if two ROC curves cross each other?

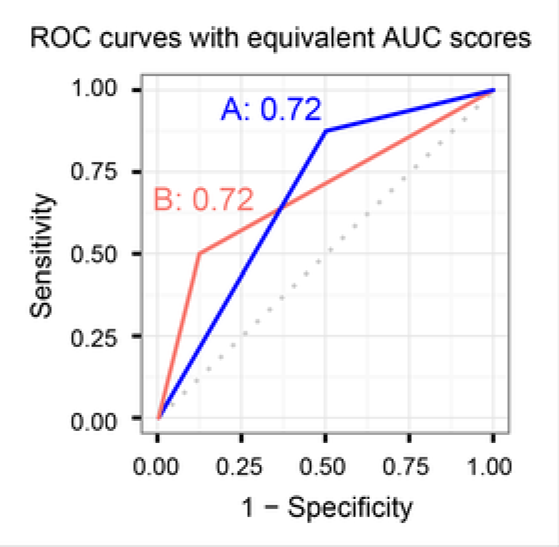
Area under ROC curves(AUC)



It shows four AUC scores. The score is 1.0 for the classifier with the perfect performance level (P) and 0.5 for the classifier with the random performance level (R). ROC curves clearly shows classifier A outperforms classifier B, which is also supported by their AUC scores (0.88 and 0.72).

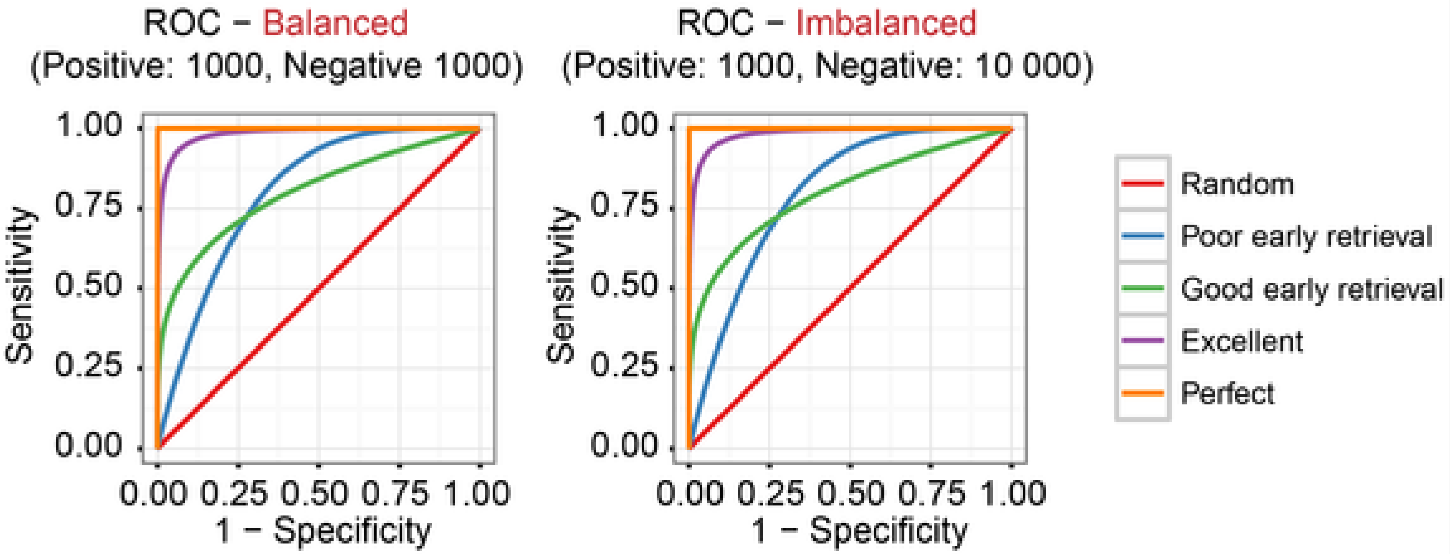
AUC is the probability that a classifier will rank a randomly chosen positive sample higher than a randomly chosen negative one.

Rely not only on AUC scores

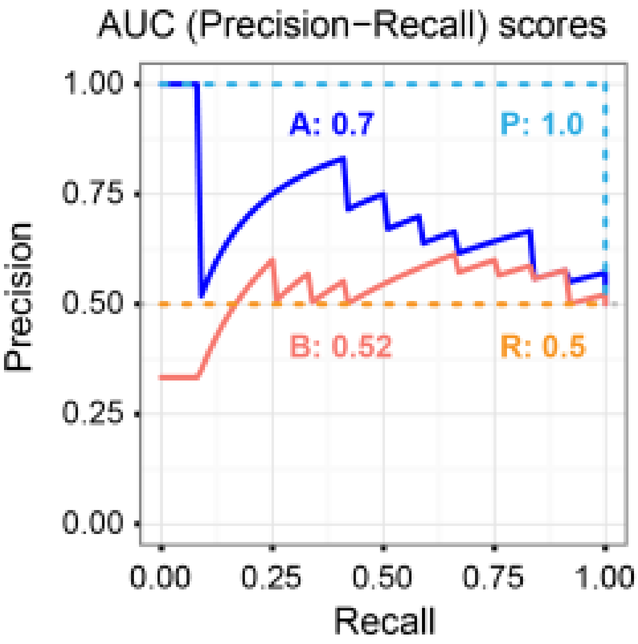


Two classifiers A and B have the same AUC scores, but their ROC curves are different.

ROCs on imbalanced data

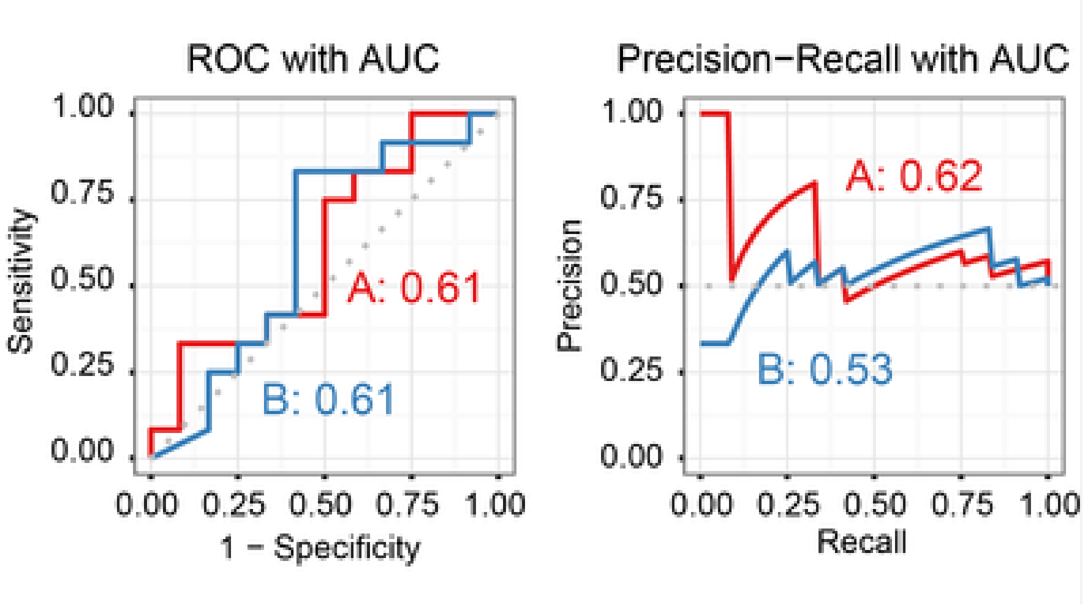


Precision recall curve



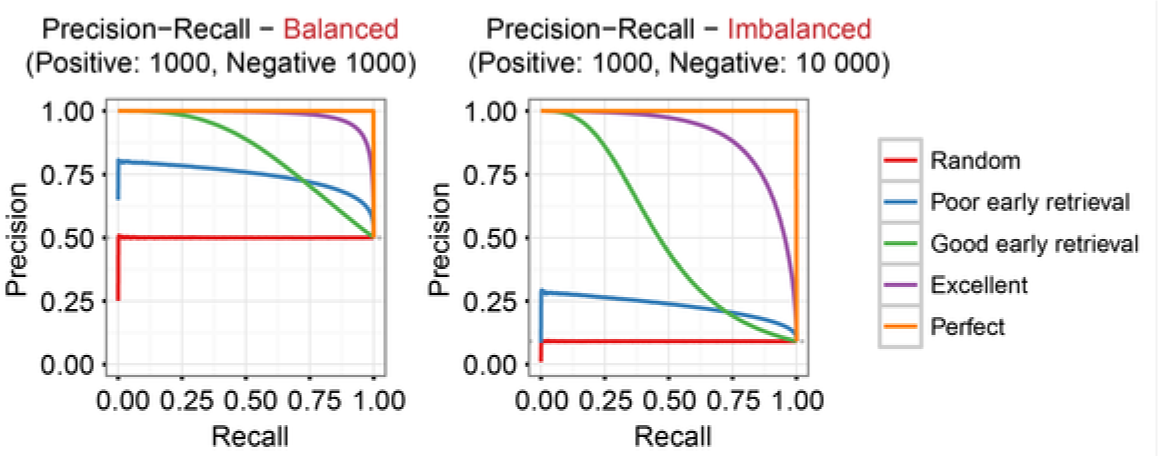
The score is 1.0 for the classifier with the perfect performance level (P) and 0.5 for the classifier with the random performance level (R). The plot clearly shows classifier A outperforms classifier B, which is also supported by their AUC scores (0.7 and 0.52).

Classifier evaluation with imbalanced data sets



- ROC shows the same AUC score for A (0.61) and B (0.61), but precision-recall shows different scores for A (0.62) and B (0.53).
- Note: AUC under precision recall curve and AUC under ROC curve have different meanings.

Precision recall curves



Two precision-recall plots show different curves for different positive and negative ratios. Both plots have five curves with different performance levels under balanced and imbalanced scenarios.

Topics:

- Evaluation metrics for classification
  - Accuracy/error, confusion matrix,
  - ROC curve, precision recall curve
- Feature extraction
- Feature selection
  - Frequency based approaches: document frequency, Categorical Proportional Difference (PD)
  - Chi-square feature selection
  - Mutual information
- Paper for COS524

Feature extraction

- Extract features to represent data:
  - Unigram: (each word in a sentence)
    - Binary representation: (0, 1,…,)
    - Vector of word counts:
    - Vector of TF-IDF scores:
  - Bigram: (each word pair in a sentence)
    - e.g. “study computer science princeton”
    - Bigrams: study computer, computer science, science princeton
  - Unigram + bigram
    - Consider all words and word pairs.

Feature selection:

- Select a subset of the extracted features to train a model
- In order to reduce computation time and avoid overfitting
- The basic idea of feature selection:
  - Assign a value to each feature according to some scoring function
  - Select the K features with the highest values

Frequency-based feature selection

- Document frequency
  - The number of documents that contain the word/term t
  - More appropriate for the Bernoulli model
- Collection frequency
  - The number of occurrences of term t in the collection
- Select some frequent terms without considering class information
  - Simple, fast method, but may select useless terms.
  - i.e. “movie”, frequent term in both positive and negative class.

Categorical Proportional Difference (PD)

- Try to find terms that occur mostly in one class of documents or the other
- Calculate the difference between the positive document frequency and negative document frequency of a term
- Score = $\frac{|\text{positveDF} - \text{negativeDF}|}{\text{positiveDF} + \text{negativeDF}}$
- Worked well for sentiment analysis reported by Tim O’Keefe and Irena Koprinska(see resources on the last slide)

Mutual information

- mutual information (MI) is a measure of the mutual dependence between two random variables.

$$\text{I(X;Y)}=\sum_{x,y} p(x,y)\log(\frac{p(x,y)}{p(x)p(y)})$$

- MI measures the amount of information shared between a term and a class.
- I(X,Y)=? If X and Y and independent.
- How to compute MI given a data set?

Observed table

Term/class	Class 1	Class 0	total
1(presence)	150	75	225
0(absence)	50	125	172
Total	200	200	400

- Mutual information can be calculate based from the observed table
- Consider X is a feature and Y is the class variable
- Compute MI for all features and select K highest values.
- sklearn.metrics.mutual_info_score

Chi-square feature selection

- Test the independence of two categorical variables: a term/feature variable and the class variable.
- $X^2 = \sum_{i,j} \frac{(O_{ij} - Eij)^2}{E_{ij}}$
  - where O_{ij} is the observed count and E_{ij} is the expected count
- *Can compute X^2 from the observed table.*
- *Higher score indicates they are dependent.*
- sklearn.feature_selection.chi2
  - Compute chi-squared stats between each non-negative feature and class.

Paper: On Discriminative vs. Generative Classifier: A Comparison of Logistic Regression and Naive Bayes

- Group discussion for COS524 in breakout rooms (~15 minutes)
- Share your opinions in the shared google doc:
  - Will send you the link in chat
  - You can focus on one of the questions
- Come back for precept wrap up

Paper: On Discriminative vs. Generative Classifier: A Comparison of Logistic Regression and Naive Bayes

- Discriminative learning has lower asymptotic error, a generative classifier may also approach its asymptotic error much faster

- Would be able to design hybrid classifiers that enjoy the best properties of either.

- How will this paper affect your choices of classifiers for HW1?

Wrap up for Precept3

- Evaluation metrics for classification
  - Accuracy/error, confusion matrix,
  - ROC curve, precision recall curve
- Feature extraction
- Feature selection
  - Frequency based approaches: document frequency, Categorical Proportional Difference (PD)
  - Chi-square feature selection
  - Mutual information

Resources:(many graphs and texts are taken from the following resources.)

- “Feature Selection and Weighting Methods in Sentiment Analysis” by Tim O’Keefe and Irena Koprinska
  - http://es.csiro.au/adcs2009/proceedings/oral-presentation/09-okeefe.pdf
- “ROC Graphs: Notes and Practical Considerations for Researchers” by Tom Fawce
  - http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.10.9777&rep=rep1&type=pdf
- Introduction to ROC plot, and precision recall plot
  - https://classeval.wordpress.com/introduction/introduction-to-the-roc-receiver-operating-characteristics-plot
  - https://classeval.wordpress.com/introduction/introduction-to-the-precision-recall-plot/
- Text classification and Naive Bayes
  - http://nlp.stanford.edu/IR-book/html/htmledition/text-classification-and-naive-bayes-1.html