CCRMA MIR Workshop 2014
Evaluating Information Retrieval Systems

Leigh M. Smith
Humtap Inc.
leigh@humtap.com

Basic system overview

Overview

- MIR Data Preparation
- Training & Test Data
- The Overfitting Problem
- Cross-validation
- Evaluation Metrics
  - Precision, Recall, F-measure
  - ROC
  - AUC

Content Format

- Impacts all levels of system
  - Data volume, storage options, analysis DSP, DB design, etc.
- Systems may or may not maintain original source content (vs. metadata).
- Systems may preserve several formats of source and metadata (n-tier).
- This is typically a given situation, rather than a design option.

Audio Content Formats

- Audio-based
  - Properties/volume of source recordings
  - MP3/AAC/WMA decoders needed?
- MIDI-based
  - Problems with MIDI, assumptions to make.
  - Human-performed vs. “quantized” MIDI
- Score image based
  - Useful, but not treated here – genre specific.
- Formal language-based
  - SCORE, SMDL, Smoke, etc.
  - MusicXML

Database Technology

- Database Designs:
  - Consider Application Requirements and Design
- Relational DB (e.g MySQL/Oracle/PostgreSQL)
  - Fixed table-formatted data
  - Few data types (number, string, date, ...)
  - One or more indices/table (part of DB design, application-specific, impacts performance)
  - Cross-table indexing and joins
- “Schema-less” NoSQL (MongoDB, Cassandra, DynamoDB)
  - Each record can differ.
  - Handling of Large/Variable Feature Vectors
- Graph DB’s (neo4j)
  - Social-Graph oriented
  - Schema-less, but models relationships between entities.
  - Enables fast retrieval of cascaded relationships.

Media data

- Historically images, now video, audio
- Volume (large single items)
- Format
  - Often items of no known, or variable structure.
- Require both content and metadata for usage.
- Scalability of storage.
- “Cloud storage”
  - Accessed via web service (HTTP) API.
- Common online providers:
  - Amazon Simple Storage Service (S3)
  - rackspace.com
  - etc.

Data preparation ("eat your greens")

- Examine your data at every chance (means, max, min, std, "NaN", "Infs").
- Sanity check: Try to visualize data when possible to see patterns and see if it makes sense.
- Eliminate noisy data
- Data preparation
  - Cleaning
    - Open up and examine
    - Handle missing values
  - Relevance / Feature analysis
    - Remove irrelevant or redundant attributes
  - Data Transformation
    - Generalize or normalize data

Training and test data

- An overfit model matches every training example (now it's "overtrained.")
- Training Error AKA "Class Loss"
- Generalization
  - The goal is to classify new, unseen data.
  - The goal is NOT to fit the training data perfectly.
- An overfit model will not be well-generalized, and **will** make errors.
- Rule of thumb: favor simple solutions and more "general" solutions.

A bad evaluation metric

- "How many training examples are classified correctly?"

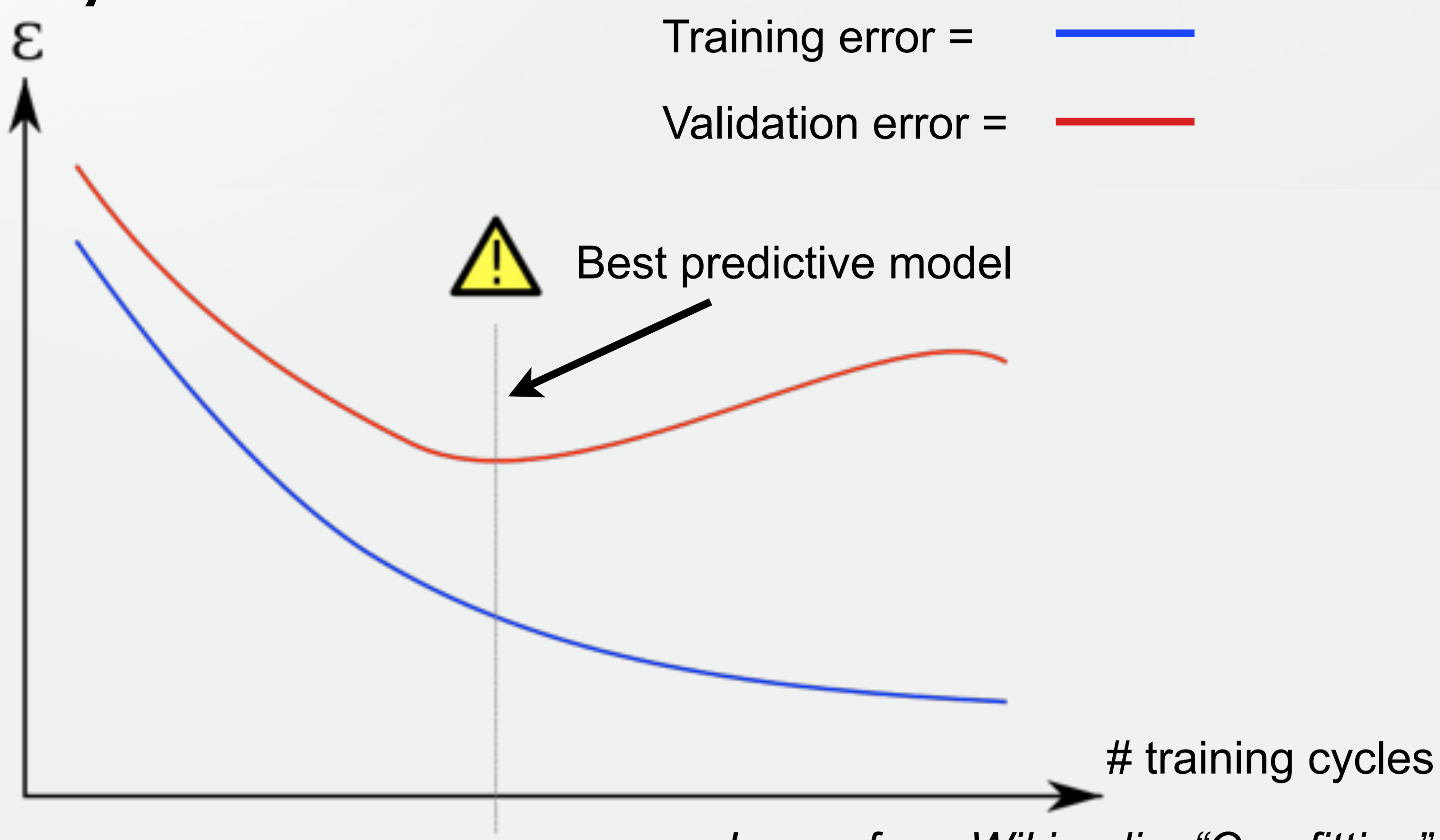


Image from Wikipedia, "Overfitting"

Overfitting

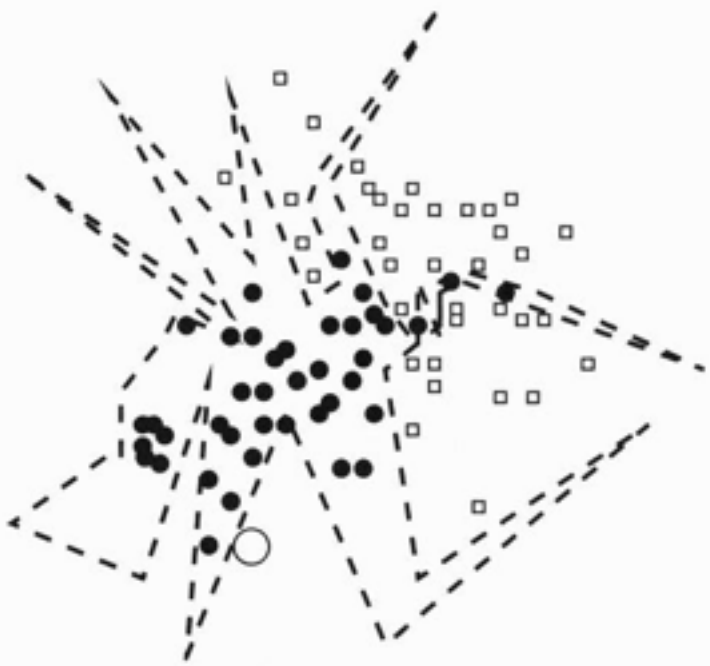


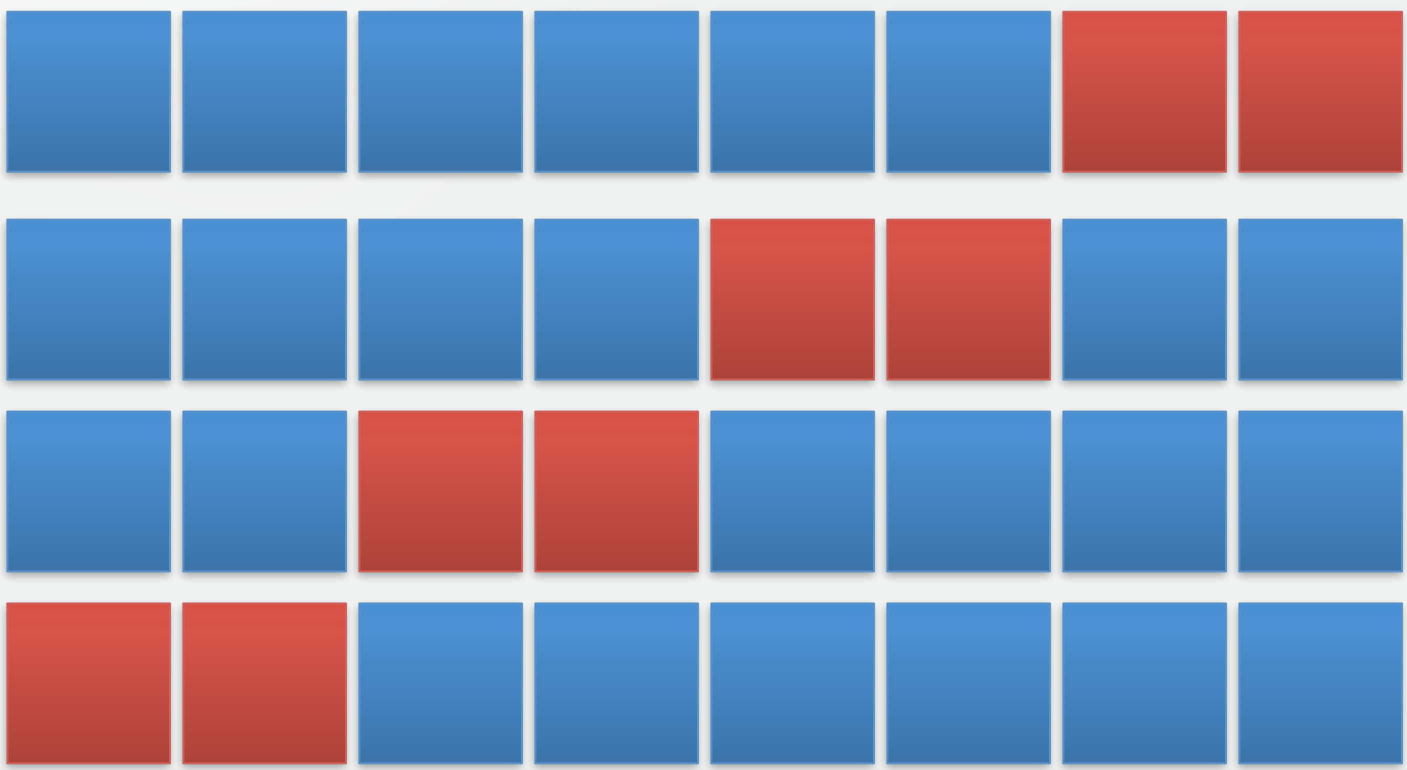
Fig. 2.13. Supervised classification into two classes with 2-dimensional data. In the training set (X, Y), data with label $y = -1$ are represented with dots, whereas data with label $y = 1$ are represented with squares. The dotted line is a classification function F such that $R^{emp}_{(X,Y)}[F] = 0$. Though it achieves zero empirical risk, F is not a good classification function, as it makes an error for a new datum which is not in the training set (circle at the bottom, with the true label $y = -1$).

Training and test data

- Training, Validation, and Test sets
  - Partition randomly to ensure that relative proportion of files in each category was preserved for each set
    - Weka or Netlab has sampling code
  - "Cross-validation"
    - Repeated partitioning.
    - Reduces false measures from data variability within sets.
- Warnings:
  - Don't test (or optimize, at least) with training data!
  - Don't train on test data!

Cross-validation:

- Accuracy on held-out (“test”) examples
- Cross-validation: repeated train /test iterations:



Evaluation Measures

True +ve	Correct	Classifier correctly predicted something in it's list of known positives.
False +ve, Type I error	Incorrect, False alarm	Classifier said that something was positive when it's actually negative. e.g. Error light flashes, but no error actually occurred. Rejecting the null hypothesis
True –ve	Correct	Classifier correctly rejected something when it’s actually negative.
False –ve, Type II error	Absent	Classifier did not hit, for a known positive result. e.g Error actually occurred, but no error light flashed. Failed to reject the null hypothesis, when the null

Confusion Matrix/Contingency Table

Hypothesized class		True class: p	True class: n
	Y	True Positives	False Positives
	N	False Negatives	True Negatives

Evaluation Measures (C. V. van Rijsbergen 1979)

“**Accuracy**”
↑ is good

Precision – “Positive Predictive Value”, “Specificity”
↓ = high F+ rate, the classifier is hitting all the time
↑ = low F+ rate, no extraneous hits

Recall – “Missed Hits”, “Sensitivity”
↓ = high F– rate, the classifier is missing good hits
↑ = low F– rate, great at negative discrimination – always returns a negative when it should

F–Measure – a blend of precision and recall (harmonic–weighted mean)
↑ is good

Precision

- Metric from information retrieval: How relevant are the retrieved results?

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

⇒ # TP / (# TP + # FP)

In MIR, may involve precision at some threshold in ranked results.

Mnemonic: **P**recision = **P**rediction measure = false **P**ositive

Recall

- How complete are the retrieved results?

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

⇒ # TP / (# TP + # FN)

⇒ $\frac{\text{Number actually correct}}{\text{Number annotated (i.e. known to be correct)}}$

⇒ determines deletions (ratio of false negatives).

F-measure

- A combined measure of precision and recall (harmonic mean)
  - Treats precision and recall as equally important

$$F = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

Accuracy metric summary

		True class: p	True class: n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		P	N

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

From T. Fawcett, “An introduction to ROC analysis”

Example Results – Confusion Matrix

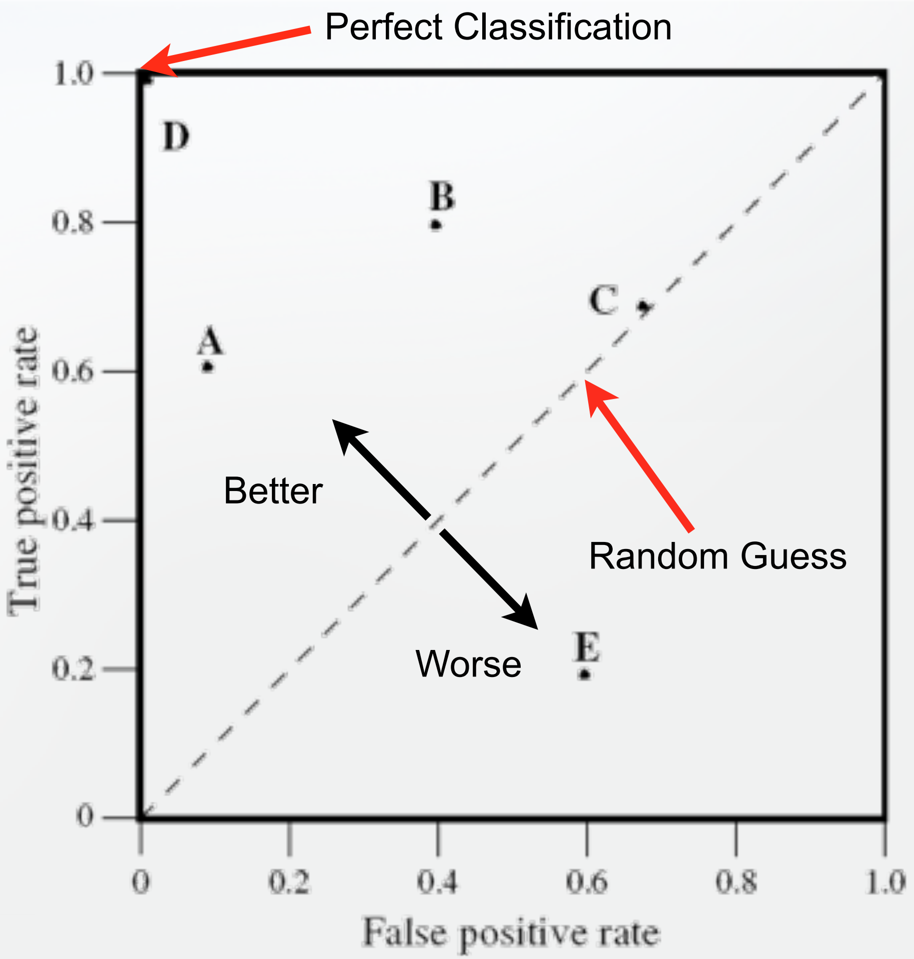
- Music/Speech/Other classification

```
Score: 2163/2450 Correct, (0 additional partial matches) of 2761 files attempted to read.
Precision = 0.8814, Recall = 0.8829, F-Measure = 0.8821

Confusion Matrix (rows = ground truth, columns = classification):
                   Other  Music Speech
           Other:    431     68    110
           Music:     17    775     18
          Speech:     46     28    957

Recall by class:
           Other: 0.7077
           Music: 0.9568
          Speech: 0.9282            Mean class recall: 0.8642
```

ROC Graph

- "Receiver operating characteristics" curve.
- A richer method of comparing model performance than classification accuracy alone.
- Plots true positive rate vs. false positive rate for different classifier threshold parameter settings.
- Depicts relative trade–offs between true positive (benefits) and false positive (costs).

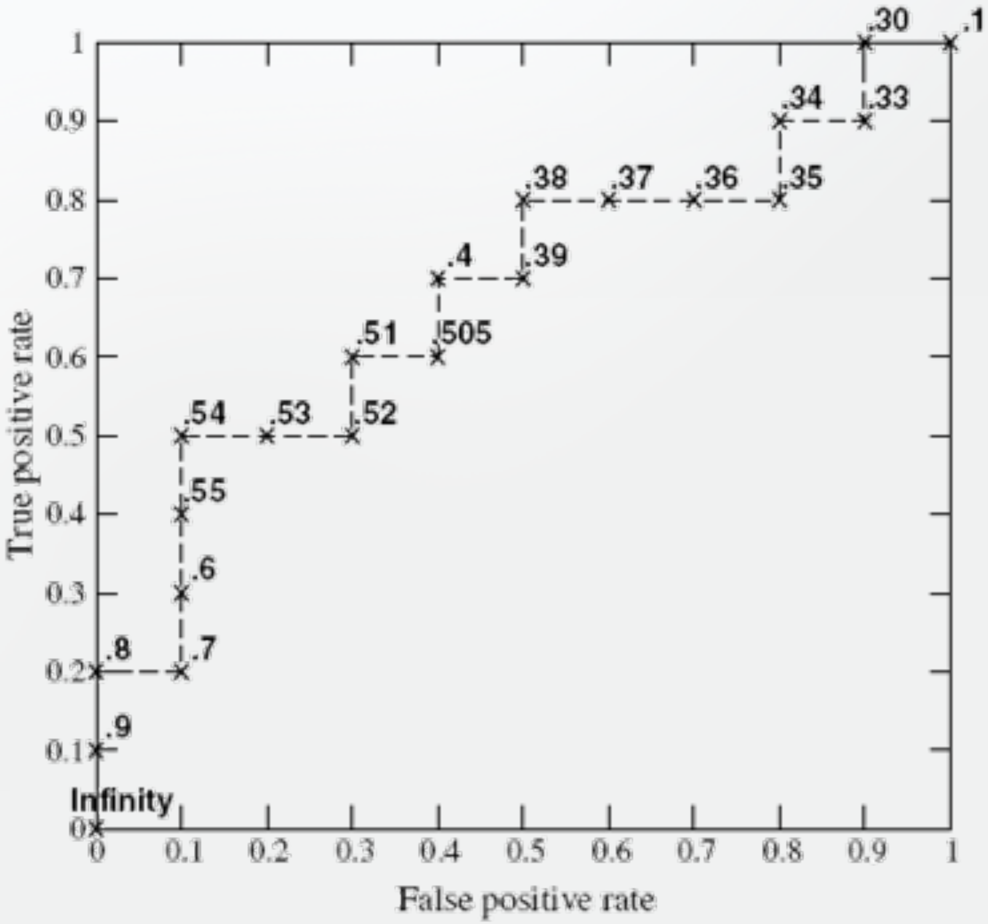
ROC plot for discrete (binary) classifiers



- Each classifier output is either right or wrong
  - Discrete classifier has single point on ROC plot.
  - Each point a confusion matrix.
- The "Northwest" is better!
- Best sub-region may be task-dependent (conservative or liberal may be better)

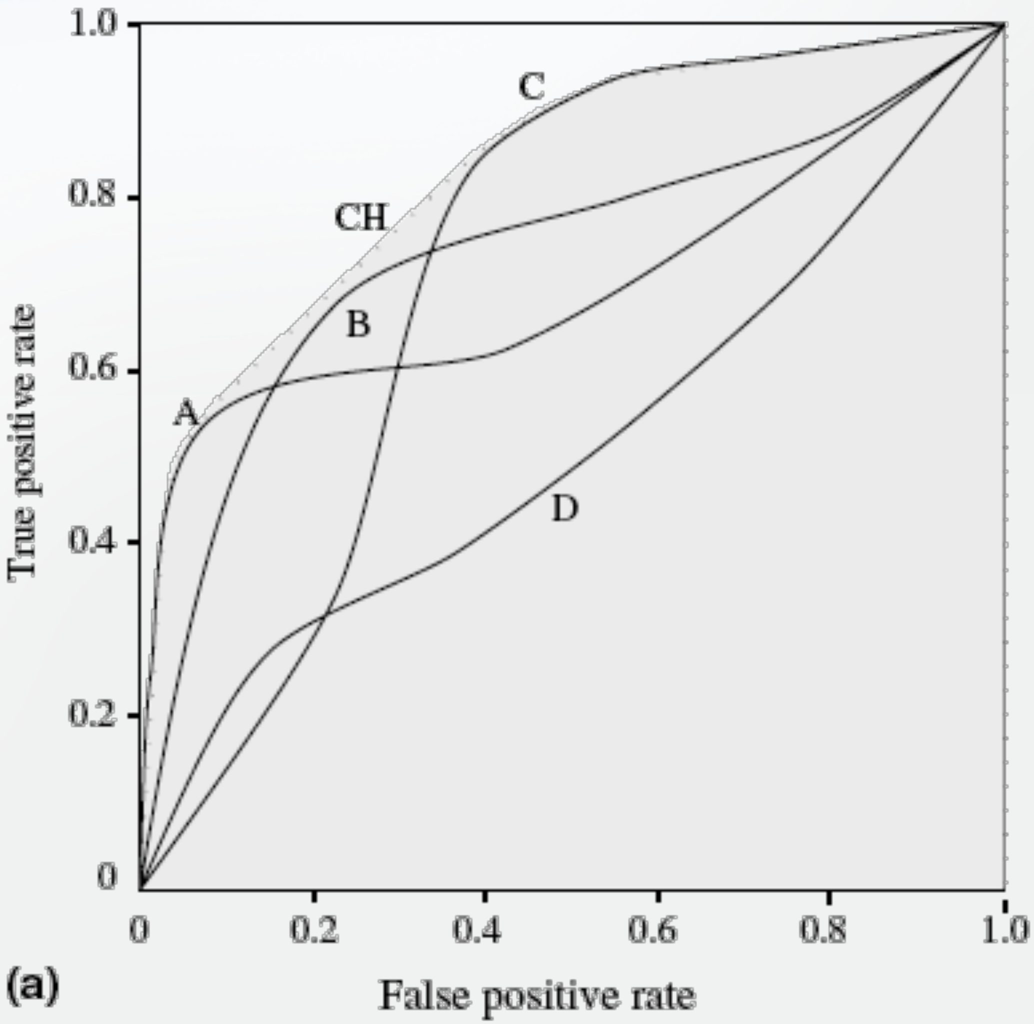
Comparing Classifiers: C < B ≦ A < D

ROC curves for probabilistic/ tuneable classifiers



- Plot TP/FP points for different thresholds of **one** classifier
  - Here, indicates that threshold of .505 is not optimal (0.54 is better)

Area under ROC (AUC)



- Compute AUC to compare different classifiers across parameter spaces.
- AUC = probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance.
- AUC not always ⇒ “better” for a particular problem.