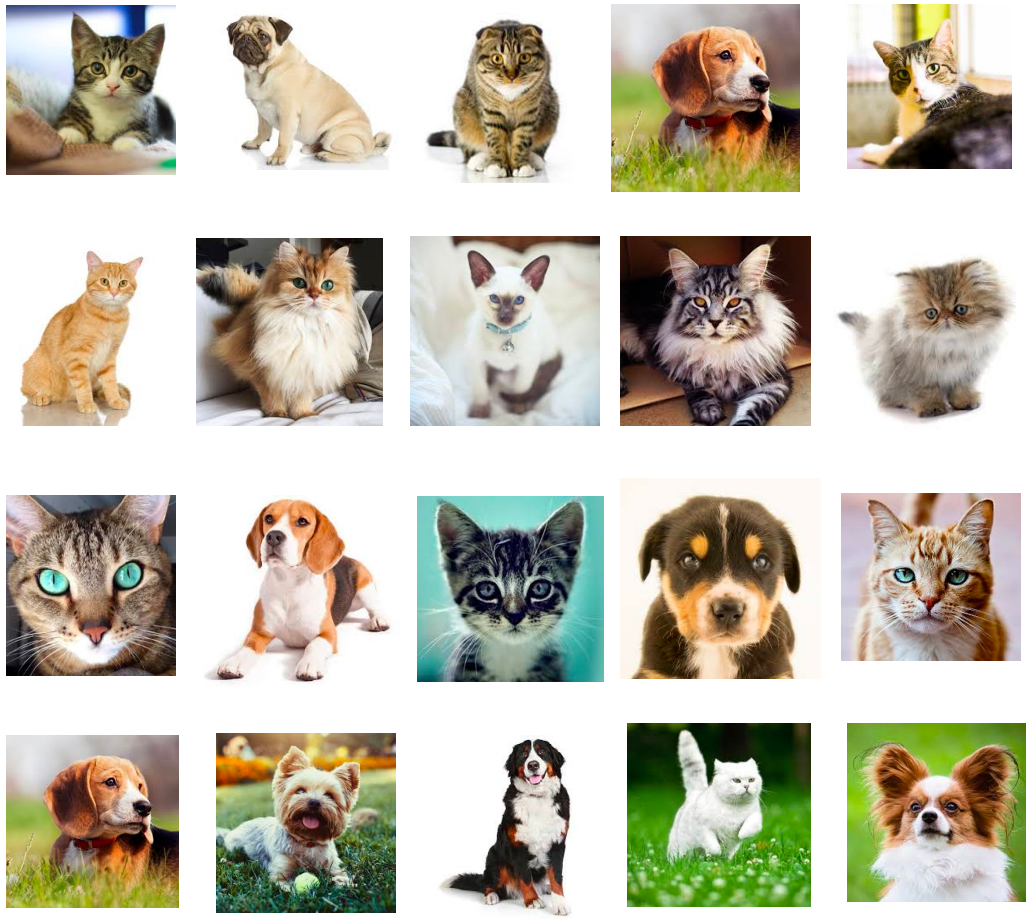
Model Evaluation

Milan Vojnovic

ST445 Managing and Visualizing Data

Binary classification



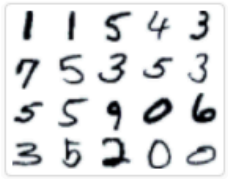
- Given output labels of a classifier, how should we quantify (visualize) its performance?

Example: image classification

- Ranking lists for different image datasets available from the link below
- Uses accuracy metric (portion of correctly classified examples)
- Accuracy can be a misleading metric when the true label distribution is skewed

MNIST

who is the best in MNIST ?



MNIST 50 results collected

Units: error %

> Classify handwriten digits. Some additional results are available on the original dataset page.

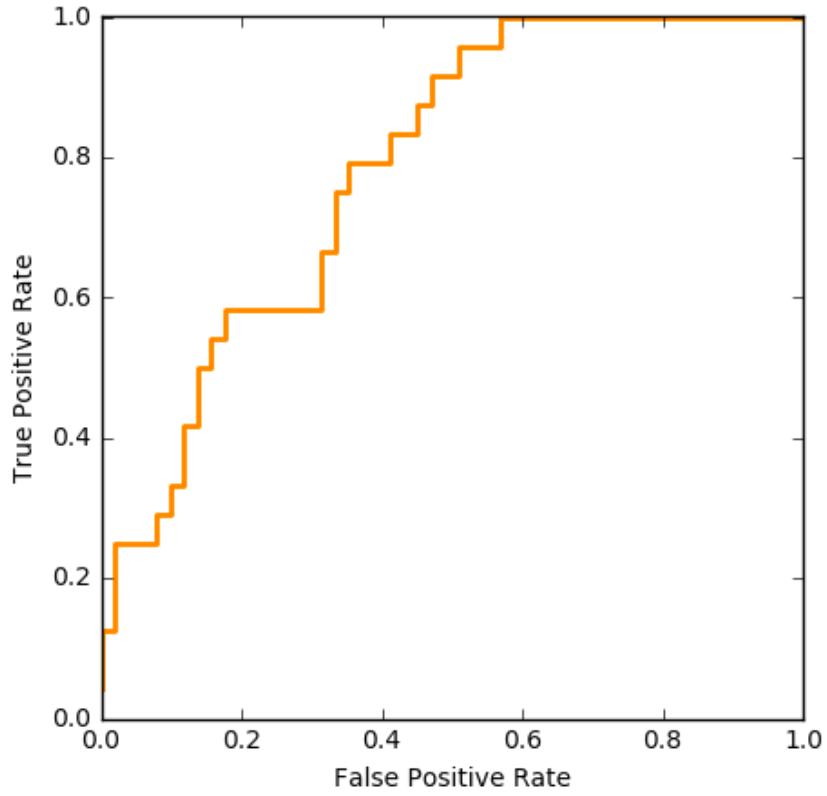
Result	Method	Venue	Details
0.21%	Regularization of Neural Networks using DropConnect	ICML 2013	
0.23%	Multi-column Deep Neural Networks for Image Classification	CVPR 2012	
0.23%	APAC: Augmented PAttern Classification with Neural Networks	arXiv 2015	
0.24%	Batch-normalized Maxout Network in Network	arXiv 2015	Details
0.29%	Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree	AISTATS 2016	Details

http://rodrigob.github.io/are_we_there_yet/build/classification_datasets_results.html

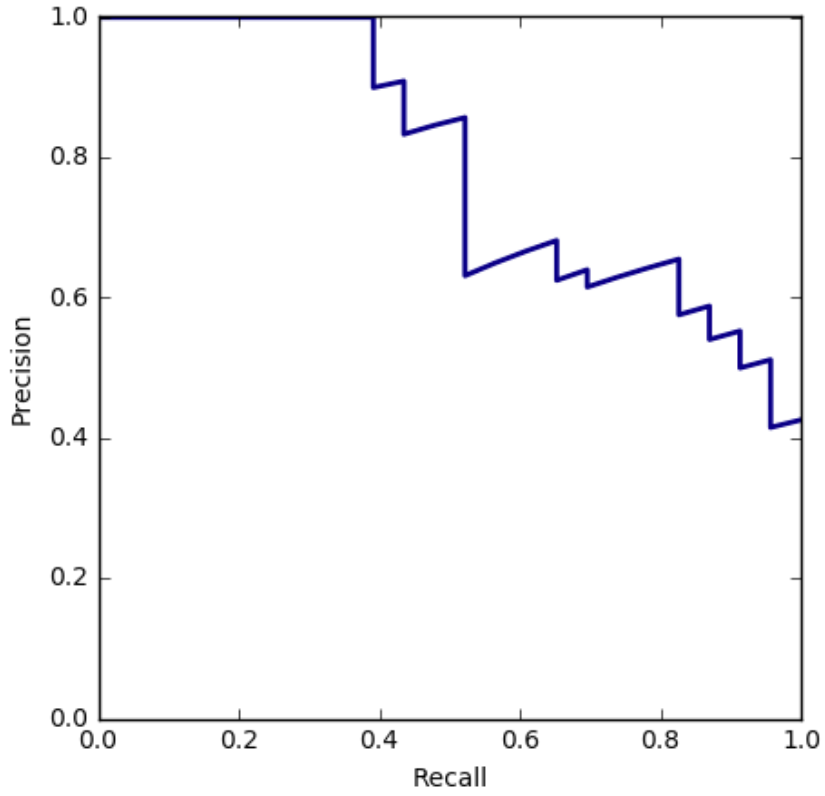
Outline

- Binary classification errors
- Accuracy metric
- Receiver operating characteristic curves
- Area under curve
- Precision-recall curves

ROC and PR curves



ROC curve
(Receiver Operating Characteristic)



PR curve
(Precision-Recall)

Software libraries

- Model evaluation in scikit-learn:

http://scikit-learn.org/stable/modules/model_evaluation.html

Performance of binary classifiers

- Input: for a classifier f

  true labels $y = (y_1, y_2, \dots, y_n)$

  and the corresponding predicted labels $\hat{y} = (\hat{y}_1, \hat{y}_2, \dots, \hat{y}_n)$

- The labels are assumed to be binary, taking values 0 or 1

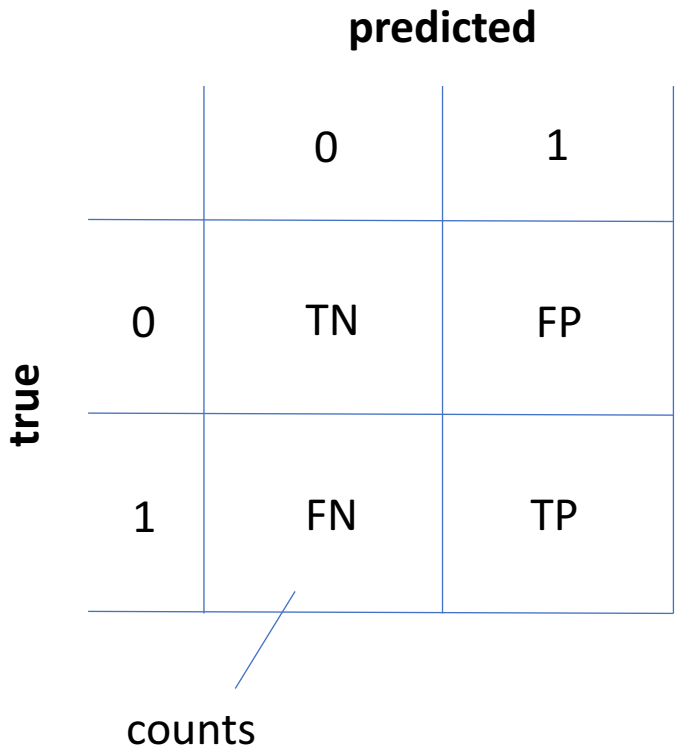
  - 0 referred as a negative label
  - 1 referred as a positive label

- Q: How should we measure the performance of classifier f ?

Scoring classifiers

- A scoring classifier outputs a real-valued score for each example
- A score may correspond to the probability of a label
- For a binary classifier, the predicted distribution is fully specified by the predicted probabilities of positive examples: $\hat{p} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n)$

Confusion matrix



$$\mathrm{TN}(y, \hat{y}) = \sum_{i=1}^{n} (1 - y_i)(1 - \hat{y}_i)$$

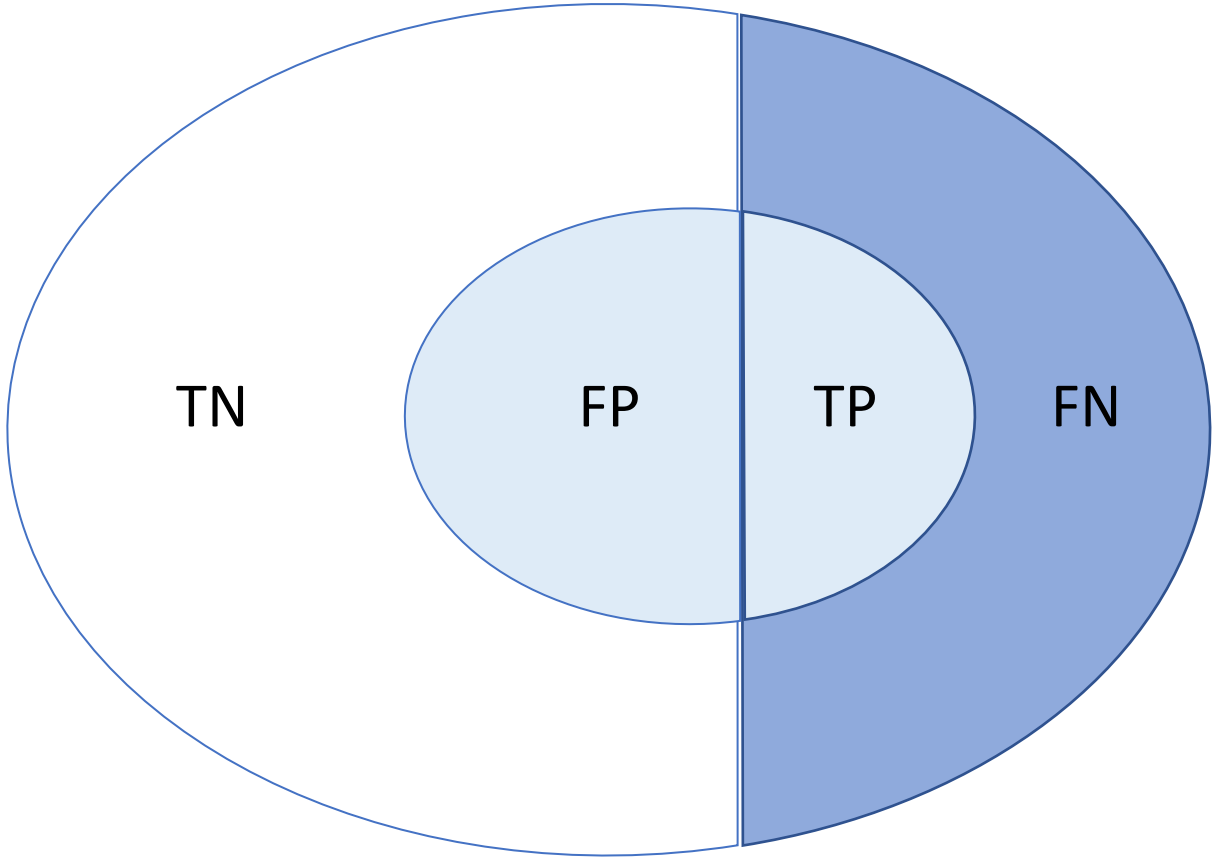
$$\mathrm{FP}(y, \hat{y}) = \sum_{i=1}^{n} (1 - y_i)\, \hat{y}_i$$

$$\mathrm{FN}(y, \hat{y}) = \sum_{i=1}^{n} y_i\, (1 - \hat{y}_i)$$

$$\mathrm{TP}(y, \hat{y}) = \sum_{i=1}^{n} y_i\, \hat{y}_i$$

Also referred to as the contingency table

Confusion matrix (cont’d)



Confusion matrix (cont'd)

- The elements of any confusion matrix sum up to the input sample size n

  $\Rightarrow$ for given sample size, the confusion matrix has three degrees of freedom

- In other words, for given sample size, the confusion matrix characterizes the performance of a binary classifier with three integer numbers

False positive rate

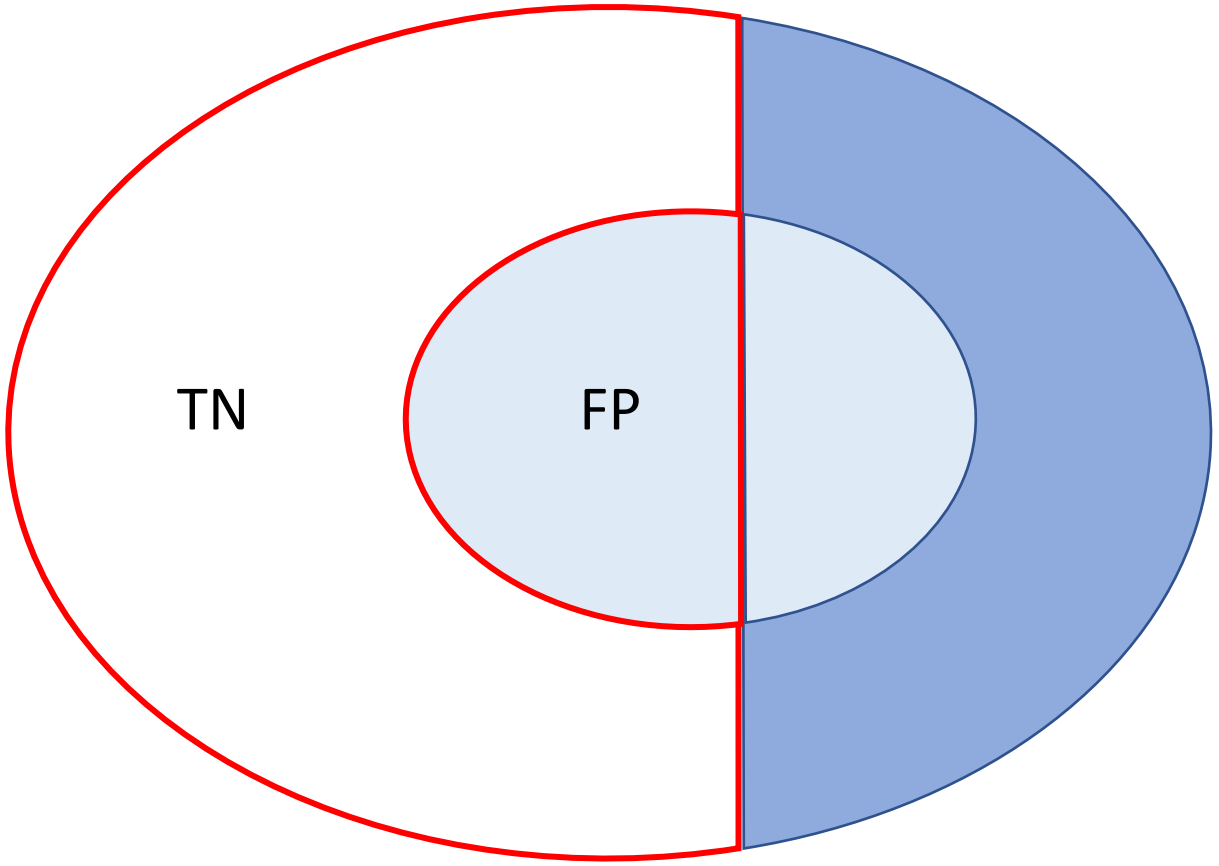
- Defined as the portion of misclassified true negative examples:

$$\mathrm{FPR}(y, \hat{y}) = \frac{\mathrm{FP}(y, \hat{y})}{\mathrm{FP}(y, \hat{y}) + \mathrm{TN}(y, \hat{y})} = \frac{\sum_{i=1}^{n}(1 - y_i)\hat{y}_i}{\sum_{i=1}^{n}(1 - y_i)}$$

- Also referred to as a fall-out

False positive rate (cont'd)

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}}$$



True negative rate

- Defined as the portion of correctly classified negative examples:

$$\mathrm{TNR}(y, \hat{y}) = \frac{\mathrm{TN}(y, \hat{y})}{\mathrm{FP}(y, \hat{y}) + \mathrm{TN}(y, \hat{y})} = \frac{\sum_{i=1}^{n}(1 - y_i)(1 - \hat{y}_i)}{\sum_{i=1}^{n}(1 - y_i)}$$

- Also referred to as a specificity

- Specificity quantifies avoiding of false positives

- Note: FPR = 1- specificity

True positive rate

- Defined as the portion of correctly classified true positive examples:

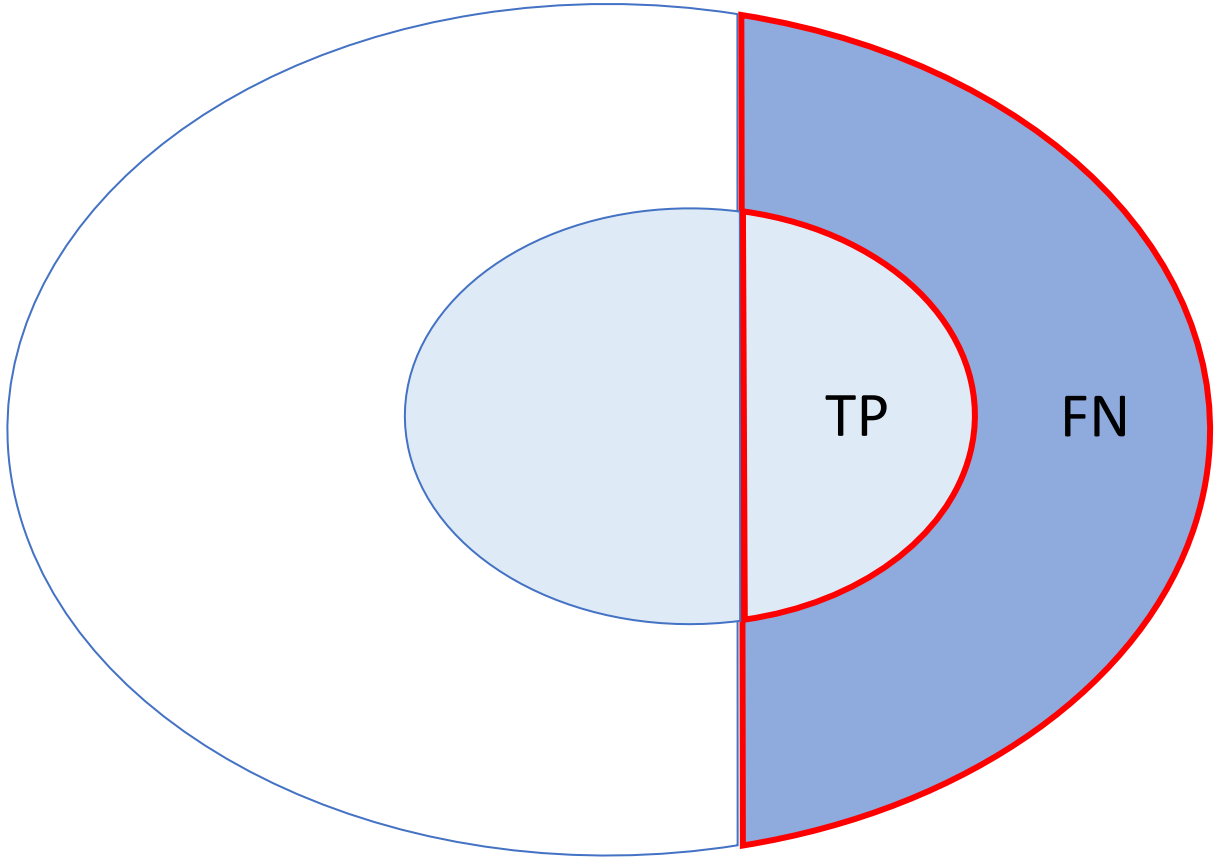
$$\mathrm{TPR}(y, \hat{y}) = \frac{\mathrm{TP}(y, \hat{y})}{\mathrm{TP}(y, \hat{y}) + \mathrm{FN}(y, \hat{y})} = \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} y_i}$$

- Also referred to as a sensitivity

- Sensitivity quantifies avoiding of false negatives

True positive rate (cont'd)

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$



Accuracy

- Defined as the portion of correctly classified examples:

$$\mathrm{ACC}(y, \hat{y}) = \frac{1}{n}(\mathrm{TP}(y, \hat{y}) + \mathrm{TN}(y, \hat{y}))$$

$$= \frac{1}{n}\sum_{i=1}^{n} y_i \hat{y}_i + \frac{1}{n}\sum_{i=1}^{n}(1 - y_i)(1 - \hat{y}_i)$$

$$\approx \mathbf{P}[Y = 1]\mathbf{P}\left[\hat{Y} = 1 \middle| Y = 1\right] + \mathbf{P}[Y = 0]\mathbf{P}[\hat{Y} = 0 | Y = 0]$$

class probability

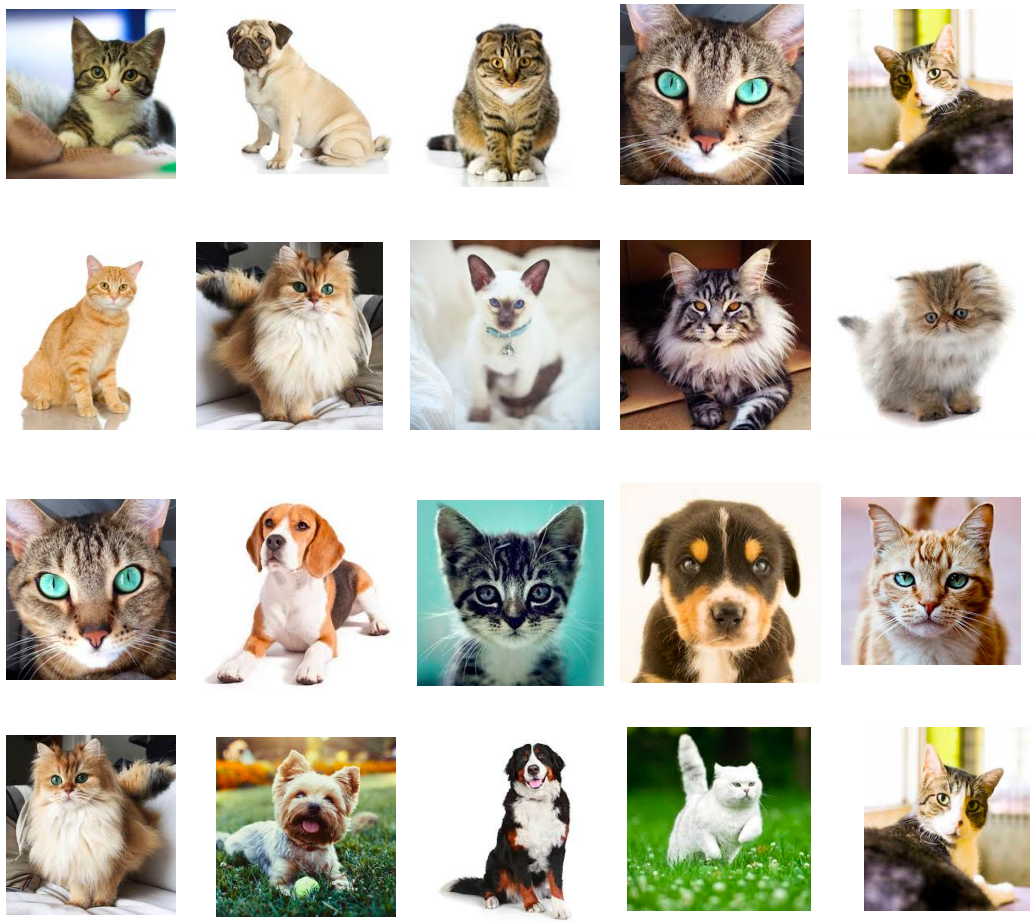
probability of correct classification

Pitfalls of the accuracy metric

- Sensitive to the true label distribution
  - The true label distribution is typically unknown
  - The true label distribution can be skewed

- Assumes equal misclassification costs
  - For false positives and false negatives
  - This is problematic because typically one type of classification error is much more expensive than other
  - Ex. fraud detection: the cost of missing a fraud event is quite different from the cost of a false alarm

Issue of skewed true label distribution



- If one class is predominant, then even a poor classifier can have a large accuracy metric

Misclassification cost

- Misclassification cost defined as

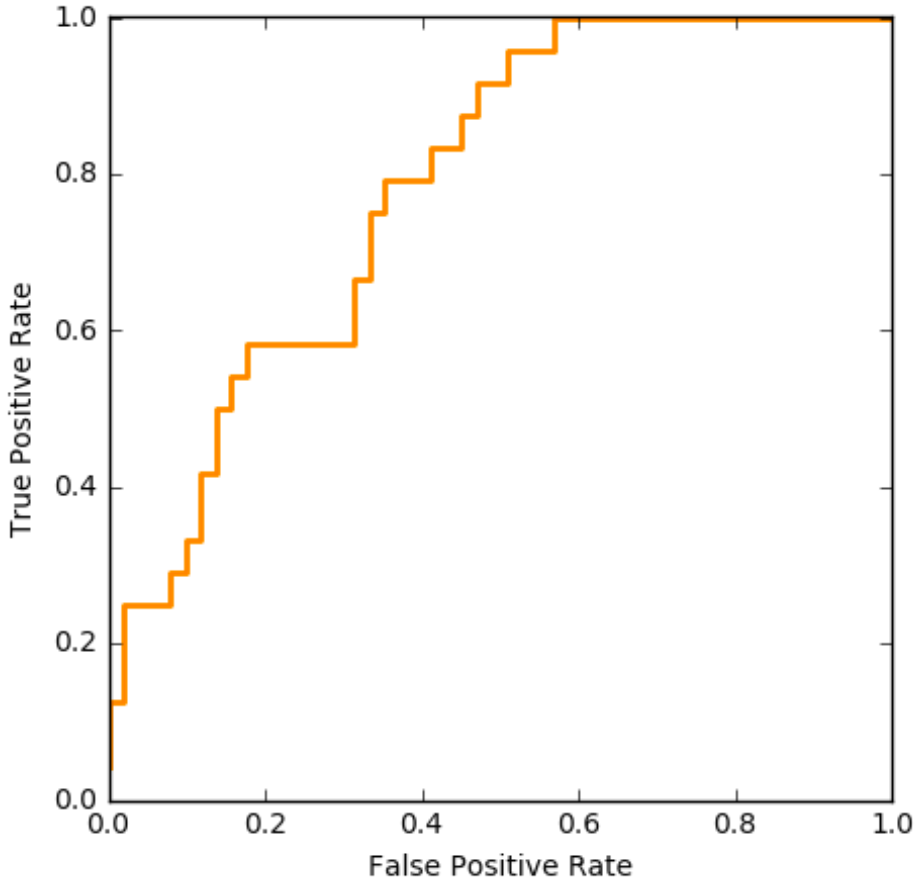
$$C_1 \frac{\mathrm{FP}(y,\hat{y})}{n} + C_2 \frac{\mathrm{FN}(y,\hat{y})}{n}$$

  where C_1 and C_2 are cost parameters

- The cost parameters are typically unknown in practice

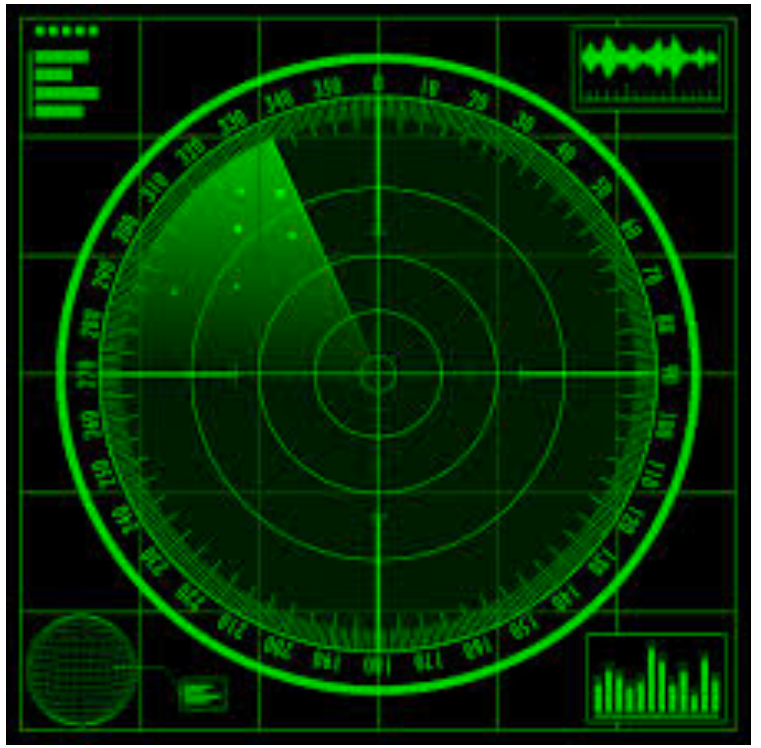
Receiver operating characteristic (ROC)

- For any classifier there is usually a trade-off between the sensitivity and the specificity
- This trade-off can be graphically represented by the ROC curve
- A ROC curve shows how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples



Historical remarks

- The term receiver operating characteristic (ROC) originates from the use of radar during World War II
- ROC analysis was developed as a standard methodology to quantify a signal receiver's ability to correctly distinguish objects of interest from the background noise in the system
- Originates from signal detection theory



Historical remarks (cont'd)

- Provides means to set a decision threshold or an operating point for the receiver to detect presence or absence of a signal

- Two types of events:
  - Hit: detection of the signal when the signal is actually present
  - False alarm: detection of the signal when the signal is actually absent

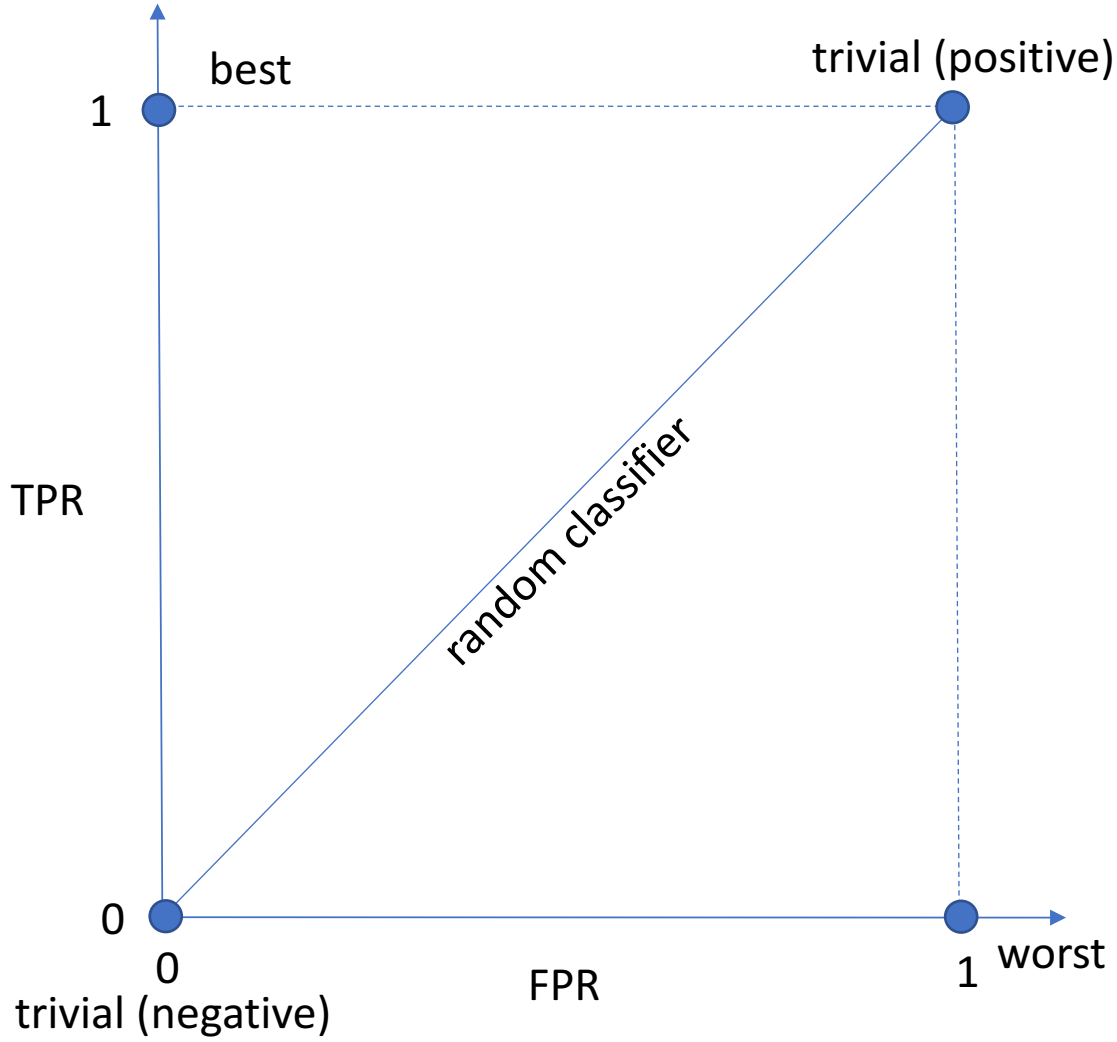
- The selection of the best operating point is about finding a trade-off between the hit rate and the false-alarm rate of a receiver

Use cases

- Signal detection theory
- Identification of optimal behavior regions
- Model selection
- Comparative evaluation of learning algorithms

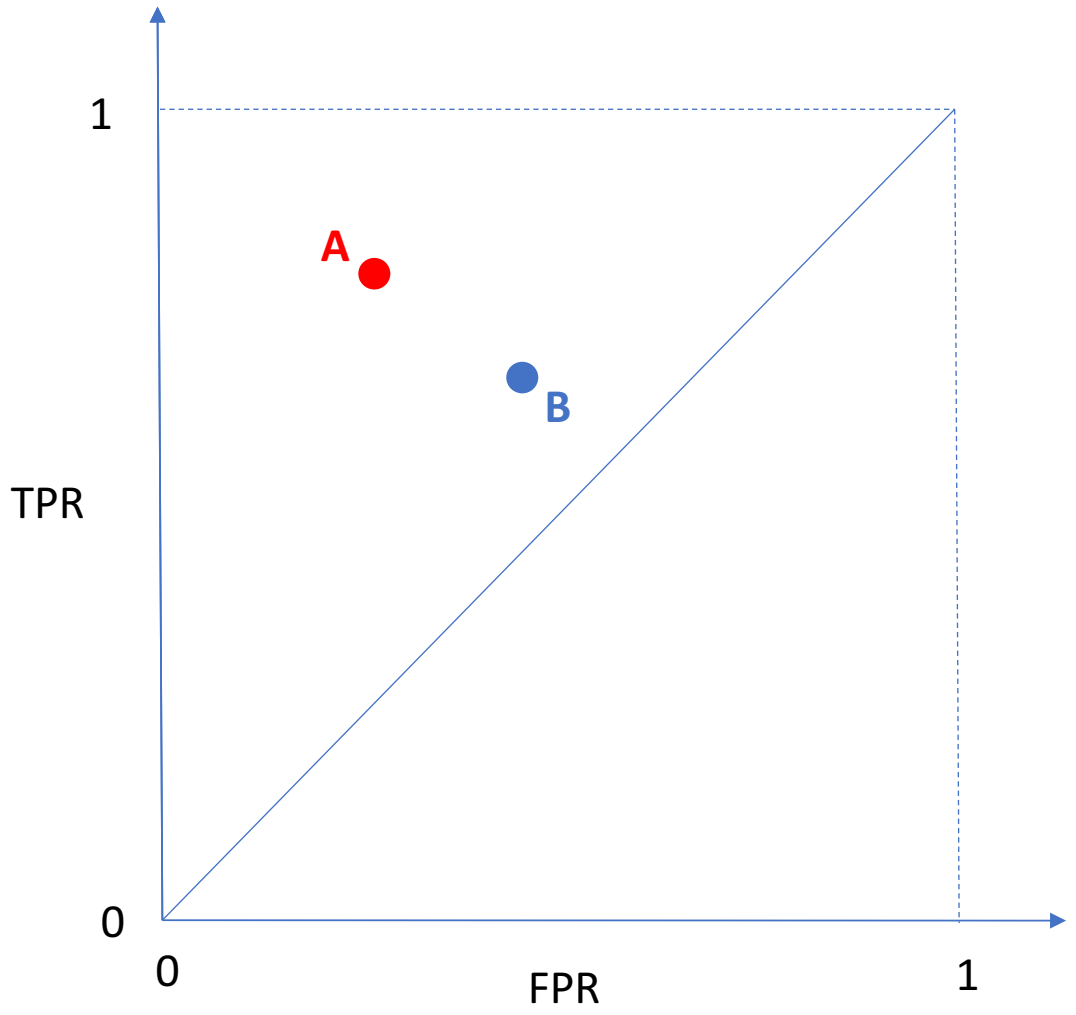
ROC space

- $[0,1]^2$ space
- $(0,0)$ trivial classifier that classifies all the examples as negative
- $(1,1)$ trivial classifier that classifies all the examples as positive
- Points on the diagonal line connecting $(0,0)$ and $(1,1)$
  - TPR= FPR
  - Considered as a random classifier
- Classifiers lying above (below) the diagonal perform better (worse) than random classifier



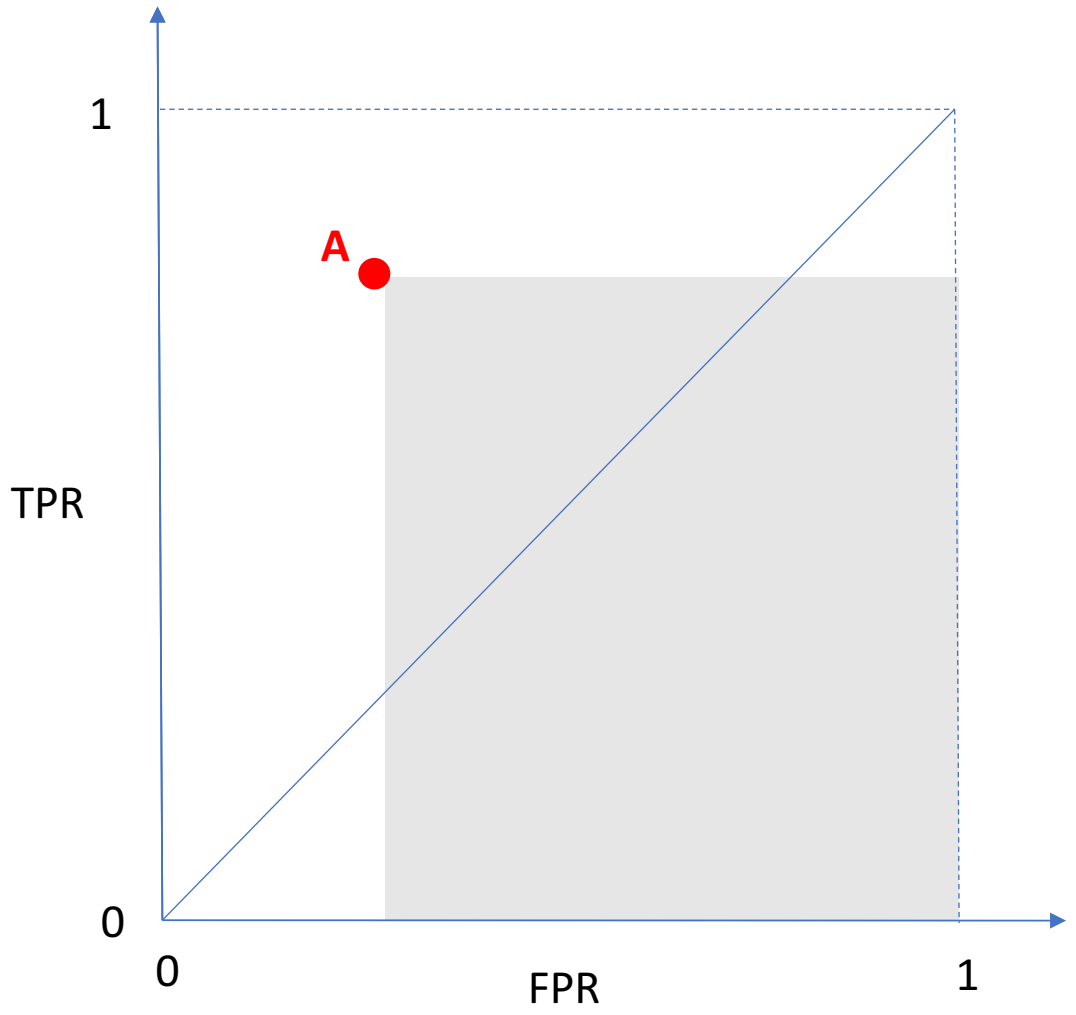
ROC space (cont'd)

- **A** represents a better classifier than **B** if **A** is on the left and higher than **B**



ROC space (cont'd)

- **A** dominates all points that are on the right and lower
- Classifiers that are more to the left-hand side are more conservative in classification of negative examples



Comparing different classifiers

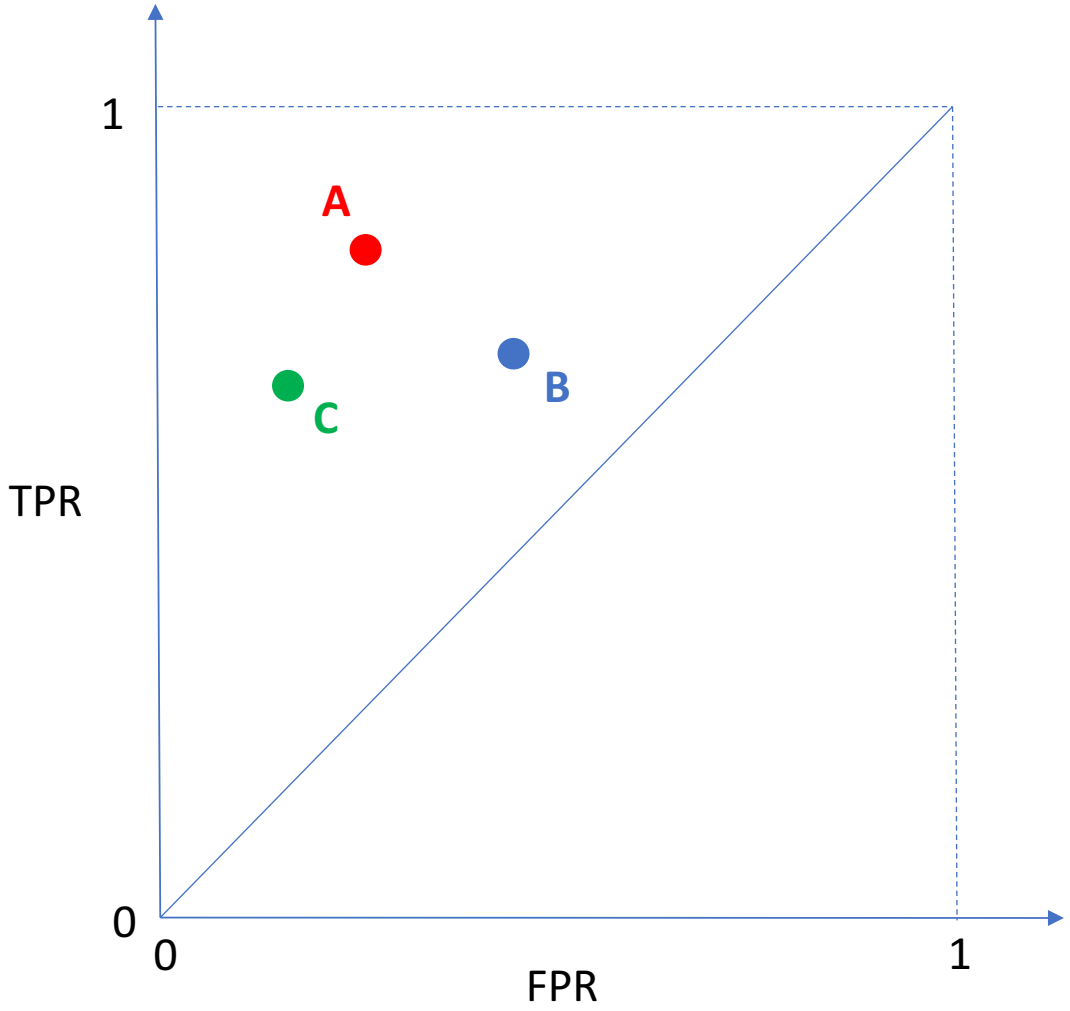
- Different operating points might be desired in different applications

- Ex. 1 cancer detection: labeling a benign growth as cancer (false positive) leads to fewer negative consequences than missing to recognize a cancerous growth (false negative)

- Ex. 2 information retrieval: false negatives are often not so serious

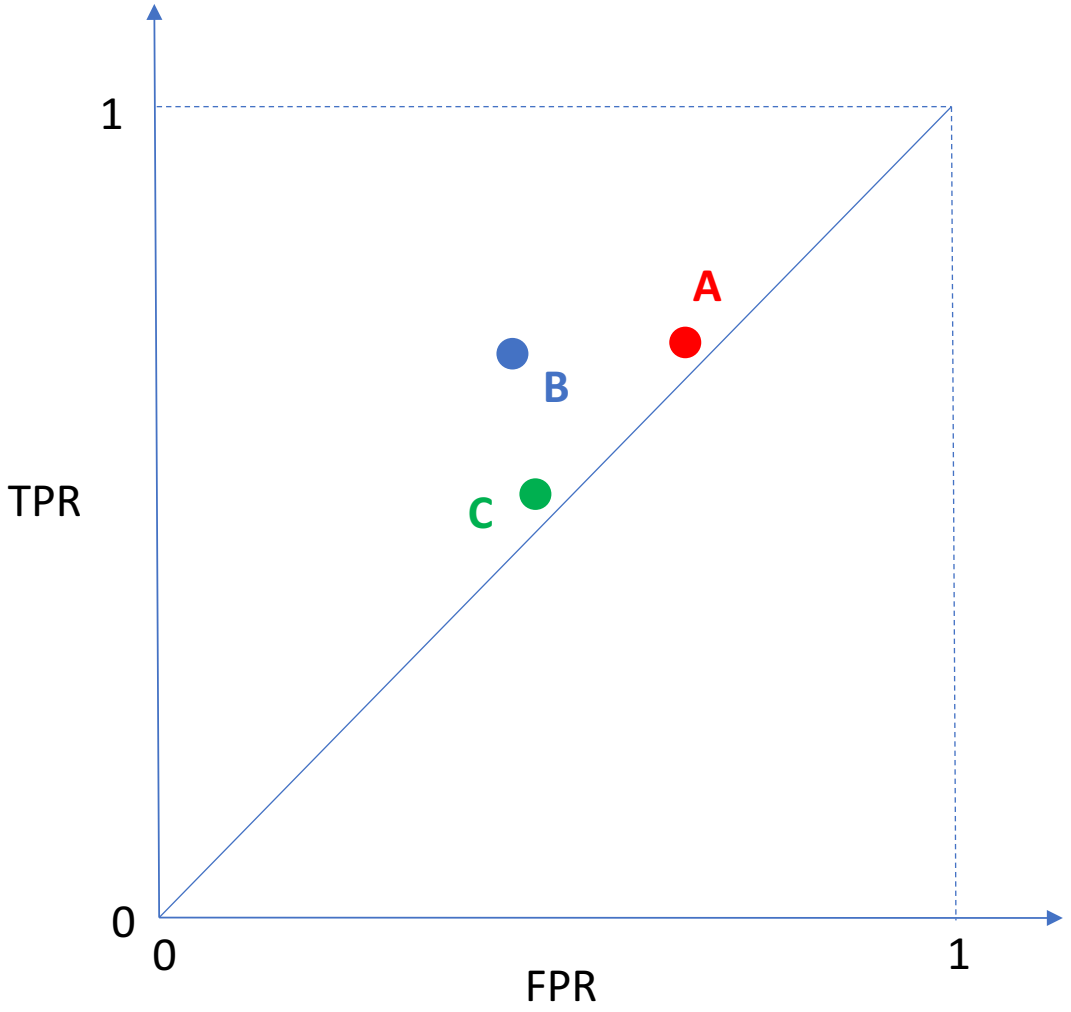
Comparing different classifiers (cont'd)

- **A** may be deemed the best because it is closest to (0,1)
- However, **C** may be preferred in cases when having small false positive rate is important
- There is not reason why **B** would be preferred over **A**



Comparing different classifiers (cont'd)

- It is non trivial to compare performance of classifiers that are just slightly better than those of a random classifiers (points just above the diagonal line)
- A statistical significance is required in such instances



ROC curve

- Input: true labels $y = (y_1, y_2, \dots, y_n)$ and predicted probabilities of positive examples $\hat{p} = (\hat{p}_1, \hat{p}_2, \dots, \hat{p}_n)$

- Let $\hat{y}(\theta) = (\hat{y}_1(\theta), \dots, \hat{y}_n(\theta))$, for parameter $\theta \in [0,1]$,

$$\hat{y}_i(\theta) = \begin{cases} 1 & \text{if } \hat{p}_i > \theta \\ 0 & \text{otherwise} \end{cases}$$

- ROC curve is defined as the parametric function:

$$(\mathrm{FPR}(y, \hat{y}(\theta)), \mathrm{TPR}(y, \hat{y}(\theta))) \text{ for } \theta \in [0,1]$$

Properties of ROC curves

- ROC curves are piece-wise constant and increasing functions
  - Check: both $\mathrm{FPR}(y, \hat{y}(\theta))$ and $\mathrm{TPR}(y, \hat{y}(\theta))$ are decreasing in θ
- The closer the curve is to the upper-left corner the better
- ROC curves are insensitive to true label distribution
  - Both TPR and FPR are insensitive to the true label distribution
  - If the proportion of positive to negative examples changes in a test set, the ROC curves will not change
  - Unlike some other performance metrics such as accuracy

Properties of ROC curves (cont'd)

- Each point of a ROC curve corresponds to a unique confusion matrix
- Recall that every confusion matrix has three degrees of freedom
- Each point of a ROC curve is a projection of a 3-dimensional point to a 2-dimensional point
- A ROC curve can be interpreted as a collection of projections of confusion matrices for different decision thresholds

Pseudo-code for ROC curve computation

```
Input: (y_1, y_2, ..., y_n) and (p̂_1, p̂_2, ..., p̂_n)
Output: R: list of ROC points in increasing FPR

N ← # negative examples
P ← # positive examples
FP ← 0, TP ← 0
π ← list of examples in decreasing p̂_i
R ← {}, pprev ← −∞, i ← 1

while i ≤ n
        if p̂_{π_i} ≠ pprev then
                R.append((FP/N, TP/P))
                pprev ← p̂_{π_i}
        end if
        if y_{π_i} = 1, TP ← TP + 1
        else FP ← FP + 1 end if
        i ← i + 1
end while

R.append((1,1))
```

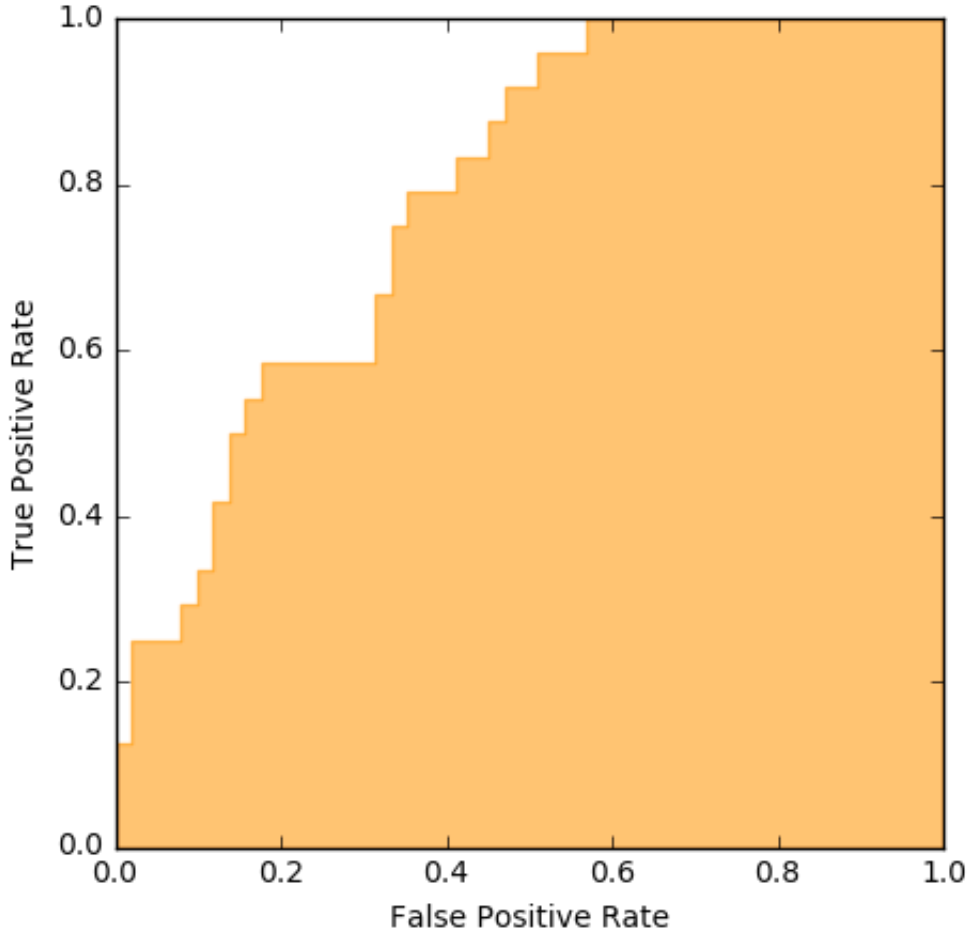
- The computation complexity is dominated by sorting n elements
- Hence, the computation complexity is $O(n \log n)$

Some observations

- ROC curves do not always provide a conclusive answer which classifier has the best performance
- This occurs when no single classifier dominates all other classifiers under consideration over the full operating range
- Having the information over the full operating range simplifies to discover the regions of optimality

Area under the curve (AUC)

- AUC is defined as the area under the ROC curve
- AUC represents “average” performance of a classifier over the full operating range
- AUC represents the ability of a classifier to rank higher a randomly chosen positive example than a randomly chosen negative example



Special AUC values

- $\mathrm{AUC} = 0$ when each positive example has a smaller or equal prediction probability of being a positive example than each negative example

- $\mathrm{AUC} = ½$ random classifier

- $\mathrm{AUC} = 1$ when each positive example has a higher prediction probability of being a positive example than each negative example

AUC formula

- AUC is equivalent to

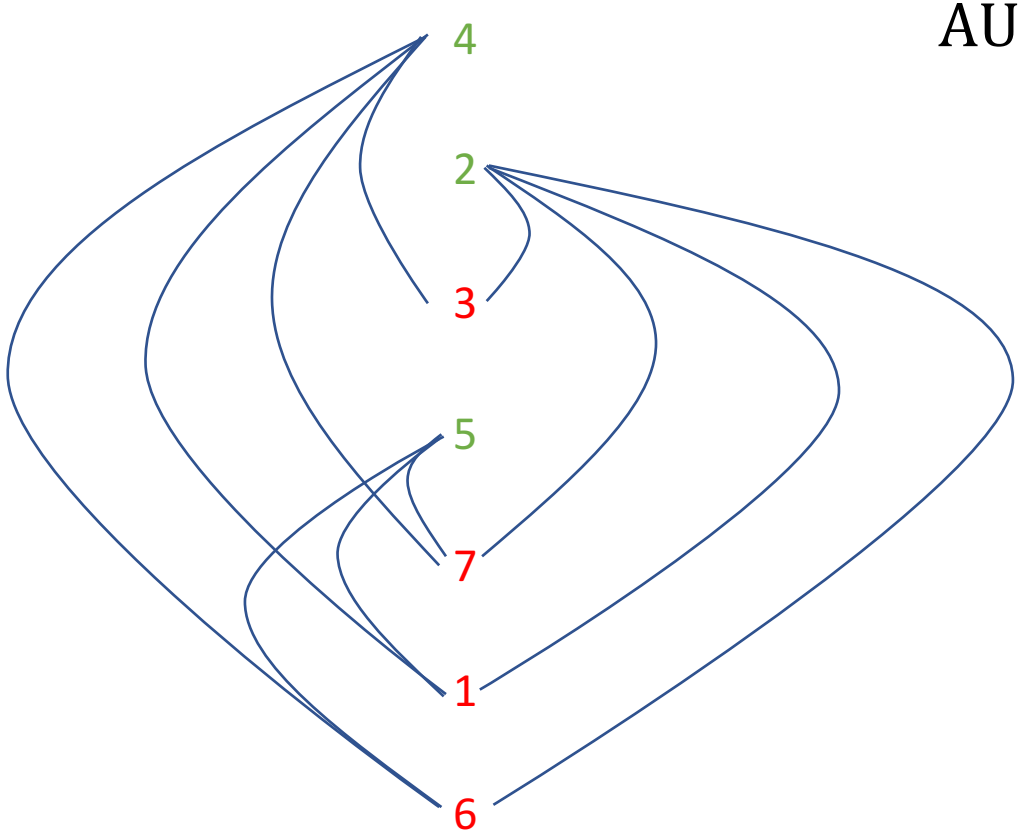
$$\text{AUC} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} y_i (1 - y_j) \mathbf{1}_{\hat{p}_i > \hat{p}_j}}{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} (1 - y_i))}$$

- AUC can be interpreted as the probability that a randomly picked pair of a positive and a negative example is correctly ranked

- AUC is equivalent to the proportion of the concordant pairs comparing the true and predicted rankings of examples

Example

i	y_i	$\hat{p}_i$
1	0	0.2
2	1	0.7
3	0	0.6
4	1	0.8
5	1	0.5
6	0	0.1
7	0	0.3



$$\text{AUC} = \frac{11}{4 \times 3} \approx 0.92$$

Proof sketch

- Let $\hat{p}_{(n)} \leq \hat{p}_{(n-1)} \leq \cdots \leq \hat{p}_{(1)}$ be sorted values of $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_n$
- Let $\hat{p}_{(n+1)} := 0$ and $\hat{p}_{(0)} := 1$
- Let $\mathrm{FPR}_i := \mathrm{FPR}(\hat{p}_{(i)})$ and $\mathrm{TPR}_i := \mathrm{TPR}(\hat{p}_{(i)})$
- Let P := # of positive examples and N := # of negative examples

- Note

$$\mathrm{AUC} = \sum_{i=0}^{n} (\mathrm{FPR}_{i+1} - \mathrm{FPR}_i)\mathrm{TPR}_i$$

$$\mathrm{FPR}_{i+1} - \mathrm{FPR}_i = \frac{1}{N} |\{\text{ negative examples with } \hat{p} \text{ values in } (\hat{p}_{(i+1)}, \hat{p}_{(i)}]\}|$$

$$\mathrm{TPR}_i = \frac{1}{P} |\{\text{positive examples with } \hat{p} \text{ values in } (\hat{p}_{(i)}, 1]\}|$$

- Hence

$$\mathrm{AUC} = \frac{1}{PN} |\{(i,j): i \text{ is positive and } j \text{ is negative and } \hat{p}_i > \hat{p}_j\}|$$

An equivalent representation

- AUC can also be represented as follows:

$$\mathrm{AUC}(f) = \frac{\sum_{i=1}^{P}(R_i - i)}{PN}$$

  where R_i is the rank of the i-th positive example in increasing order of scores (with the smallest score example assigned rank 1)

- Proof left for exercise

Example

i	y_i	$\hat{p}_i$
1	0	0.2
2	1	0.7
3	0	0.6
4	1	0.8
5	1	0.5
6	0	0.1
7	0	0.3

4 $R_3 = 7$

2 $R_2 = 6$

3

5 $R_1 = 4$

7

1

6

$$\text{AUC} = \frac{4 + 6 + 7 - 3\times 4/2}{4\ x\ 3} = \frac{11}{12}$$

Relation to the Wilcoxon's Sum of Rank test

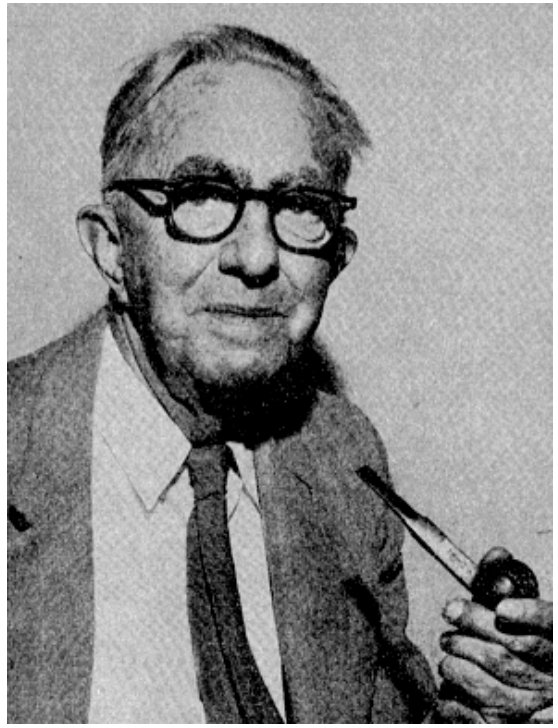
- Estimates the probability that a randomly chosen positive example is ranked before a randomly chosen negative example

$$\frac{S_+ - P(P+1)/2}{PN}$$

where S_+ is the sum of ranks of positive examples

- AUC = U/(NP)
- U is the Mann-Whitney U statistic

Historical Remarks



- F. Wilcoxon
- 1892-1965
- A chemist and statistician

- Wilcoxon signed-rank test:
  - F. Wilcoxon, *Individual Comparisons by Ranking Methods.* Biometrics Bulletin 1: 80–83, 1945

AUC and Gini coefficient

- Gini coefficient G is a popular measure of statistical dispersion used in economics as a measure of income inequality

- $\mathrm{G} = 2\mathrm{AUC} - 1$

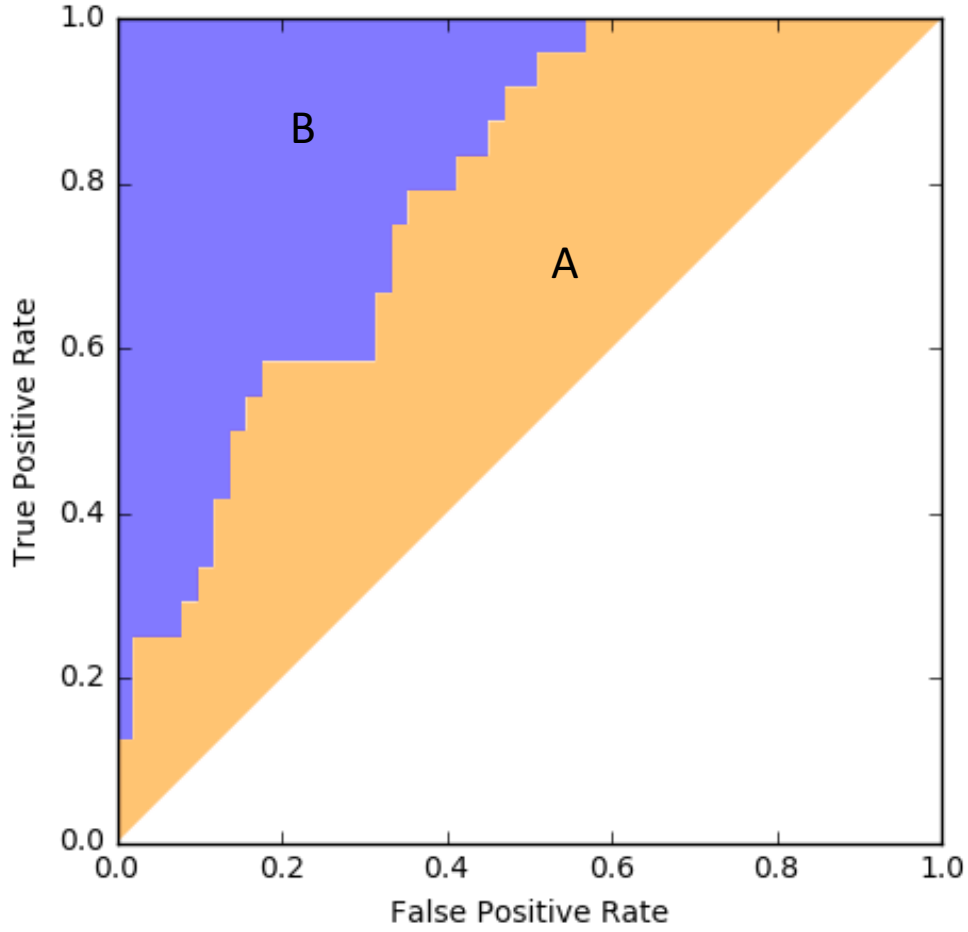
- It can be seen as a “chance standardized version”
  - Subtracts the AUC expected from a random classifier, for which $\mathrm{AUC} = 1/2$

- Not the same as the Gini index

AUC and Gini coefficient (cont'd)

$$\mathrm{G} = \frac{A}{A+B}$$

$$\mathrm{AUC} = A + \frac{1}{2} = (A+B)\mathrm{G} + \frac{1}{2} = \frac{1}{2}\mathrm{G} + \frac{1}{2}$$



Precision-recall curve

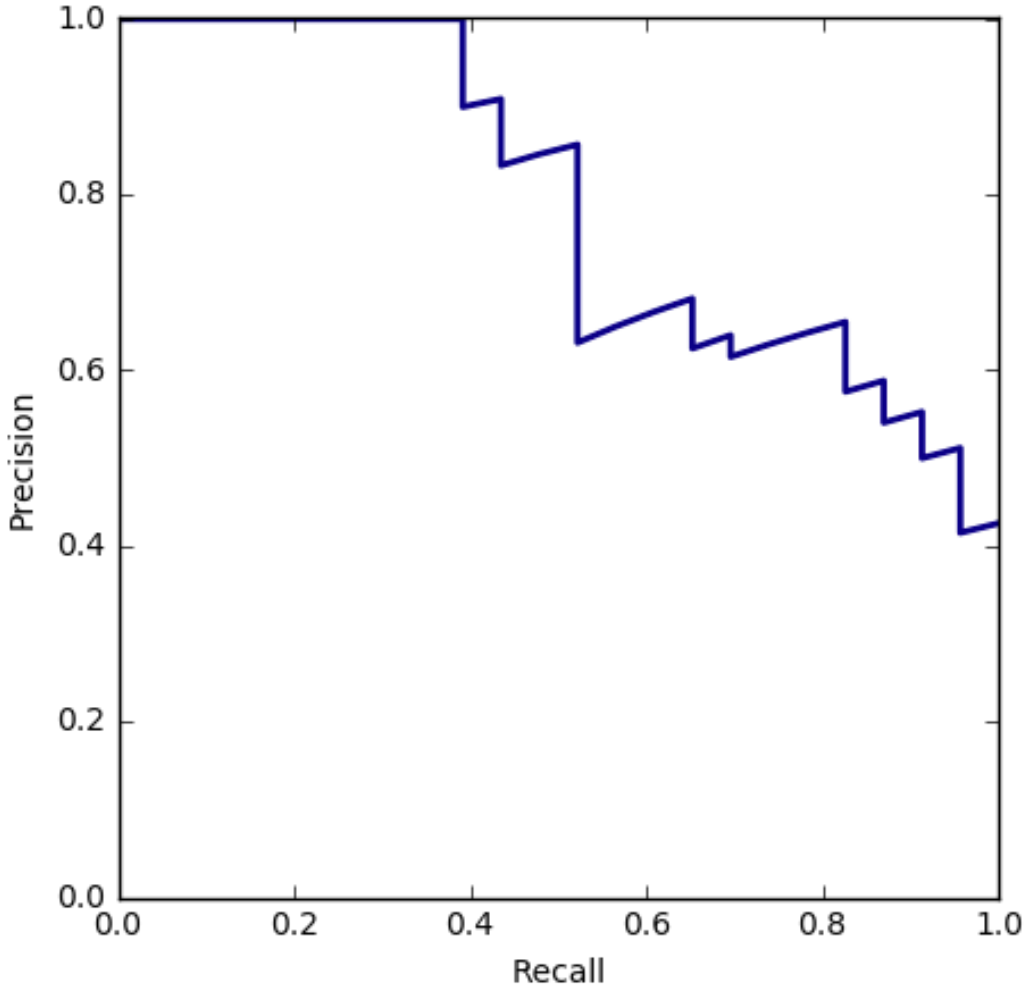
- Popular visualization of a classifier performance used in information-retrieval

- Interpretation:

  precision: portion of relevant documents in the set of documents displayed to the user

  vs.

  recall: portion of all relevant documents displayed to the user



Precision

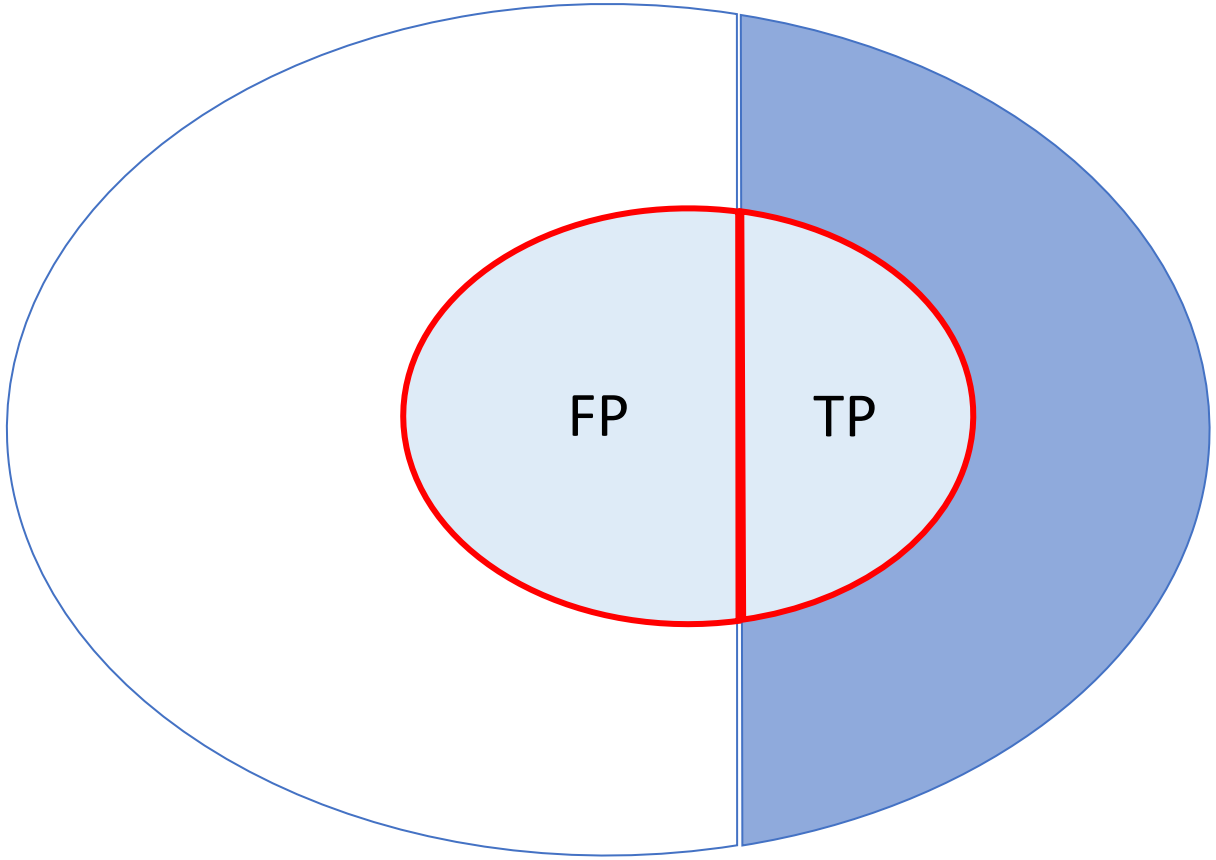
- Precision is defined as the portion of the true positive examples in positively classified examples

$$\text{Precision}(y, \hat{y}) = \frac{\text{TP}(y, \hat{y})}{\text{TP}(y, \hat{y}) + \text{FP}(y, \hat{y})} = \frac{\sum_{i=1}^{n} y_i \hat{y}_i}{\sum_{i=1}^{n} \hat{y}_i}$$

- Information retrieval interpretation:
  - Proportion of relevant documents in a set of documents classified as relevant

Precision (cont'd)

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$



Recall

- Recall is defined as the portion of correctly classified positive examples

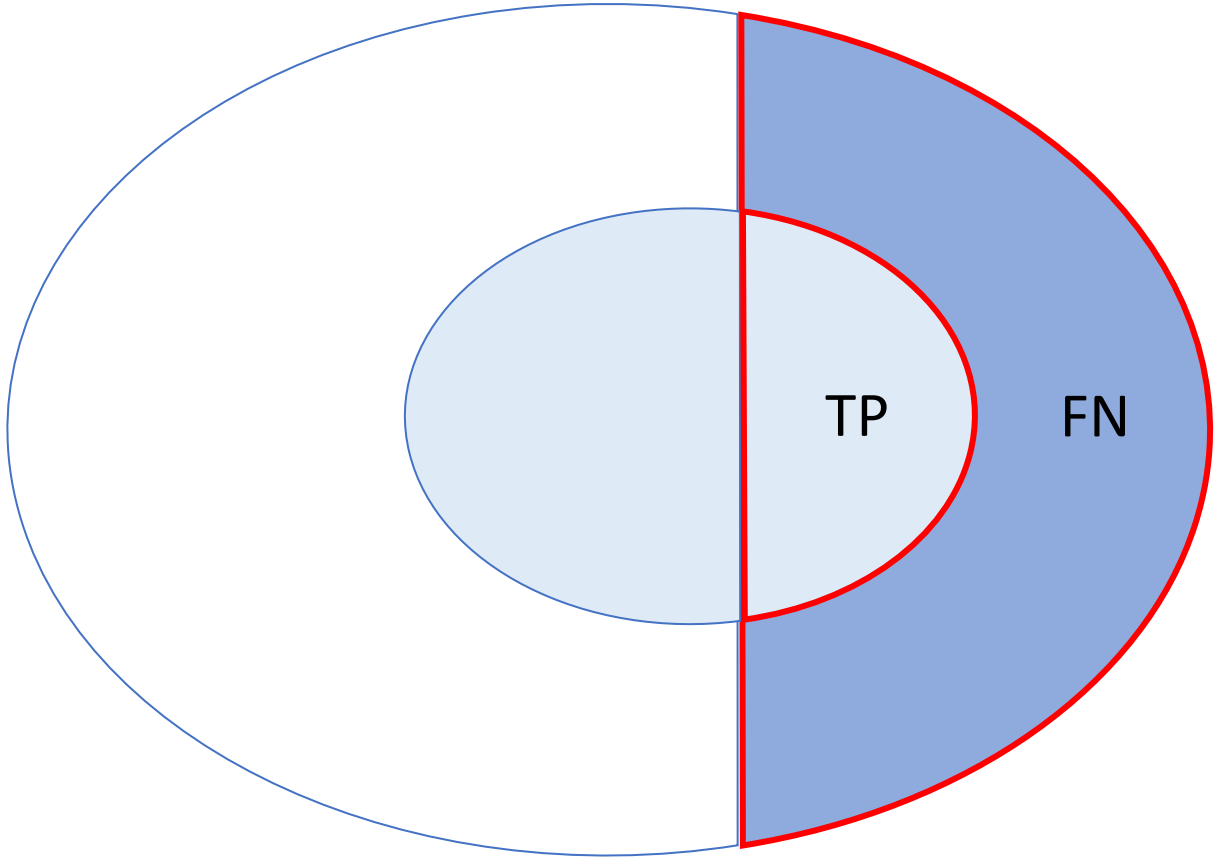
$$\text{Recall}(y, \hat{y}) = \frac{\text{TP}(y, \hat{y})}{\text{TP}(y, \hat{y}) + \text{FN}(y, \hat{y})} = \frac{\sum_{i=1}^{n} y_i \hat{y_i}}{\sum_{i=1}^{n} y_i}$$

- Same as the true positive rate (TPR)

- Information retrieval interpretation:
  - Portion of relevant documents classified as relevant

Recall (cont'd)

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$



Precision-recall curve (PR)

- PR curve is defined as the parametric function:

$$(\text{Precision}(y, \hat{y}(\theta)), \text{Recall}(y, \hat{y}(\theta))), \text{ for } \theta \in [0,1]$$

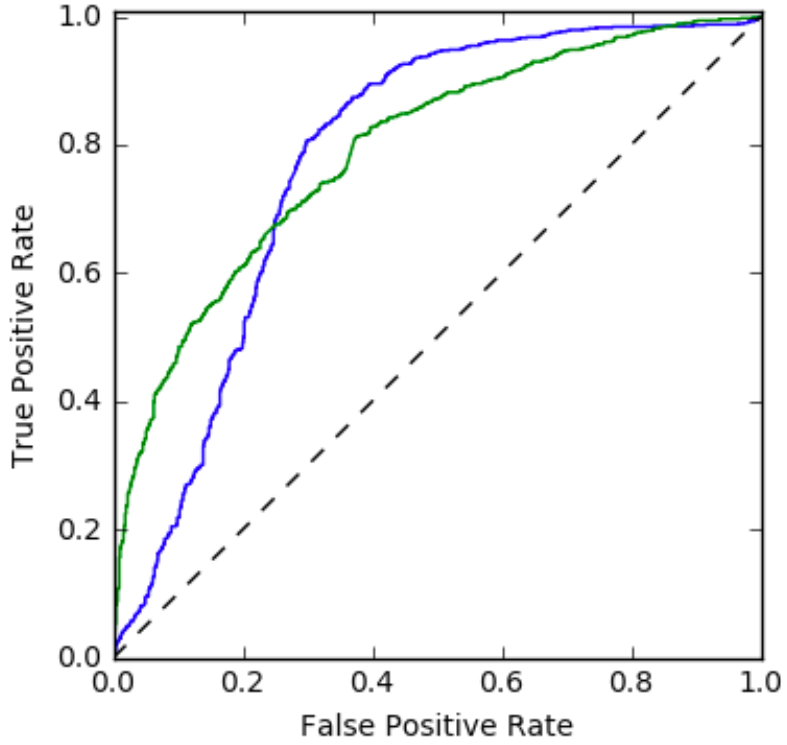
Properties of PR curves

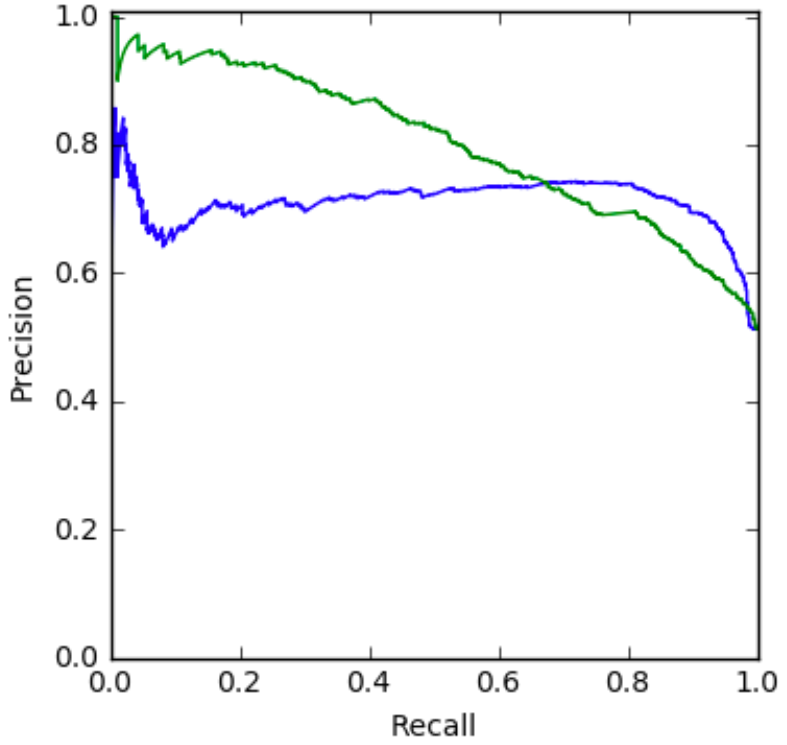
- Decreasing function (because precision decreases as recall increases)
- Explores the trade-off between the correctly classified positive examples and the number of misclassified negative examples
- PR curves can be sometimes more appropriate than ROC curves for highly skewed examples

ROC vs PR curves

- ROC curves can present an overly optimistic view of a classifier's performance if there is a large skew in the true label distribution
- PR curves considered as an alternative to ROC curves for tasks with a large skew in the true label distribution
- PR curves are sensitive to the skew of the true label distribution
- Looking into PR curves can expose differences between classifiers that are not apparent in ROC space

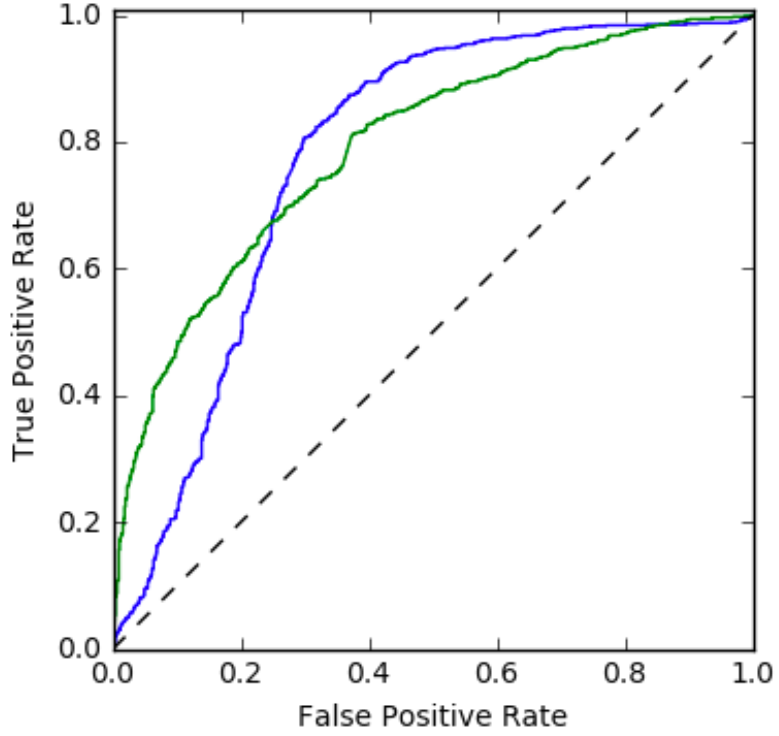
Example ROC vs PR for two classifiers

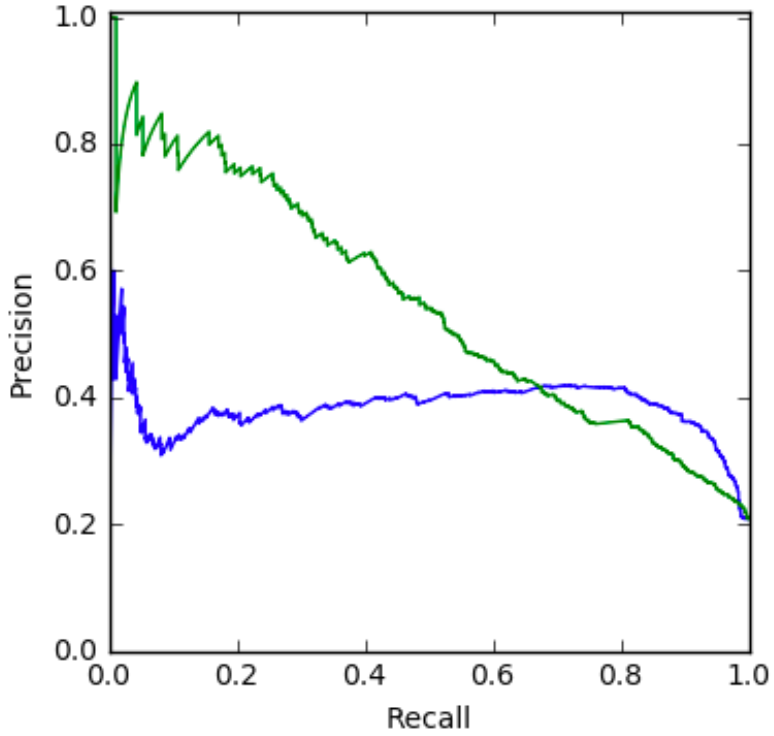




- The curves are for the prediction task “gender recognition by voice” studied in the class, for two classifiers: linear SVM and RBF SVM
- Performances of these two classifiers appear to be comparable in ROC space whilst in the PR space one of hem has a clear advantage over the other

Example ROC vs PR for two classifiers (cont'd)





- Same example as in the previous slide but for skewed input data obtained by adding 3 copies of each negative example
- Note that ROC curves remain the same (insensitive to true distribution of labels)
- Note that PR curves substantially changed

Relation between ROC and PR curves

- For any given dataset of positive and negative examples, there exists a one-to-one correspondence between a curve in ROC space and a curve in PR space, such that the curves contain exactly the same confusion matrices, if recall > 0

- Theorem: For a fixed number of positive and negative examples, one curve dominates another curve in ROC space if and only if the first dominates the second in PR space

Multi-class case

- ROC analysis in the multi-class case is more complex than in the two-class case

- In the two-class case:
  - ROC plots are easier to visualize and interpret
  - The symmetry of the two-class classification

- Here is one definition of AUC for the multi-class case:

$$\mathrm{AUC}_{\mathrm{multiclass}}(f) = \frac{1}{\binom{m}{2}} \sum_{a,b \in L} \mathrm{AUC}_{a,b}(f)$$

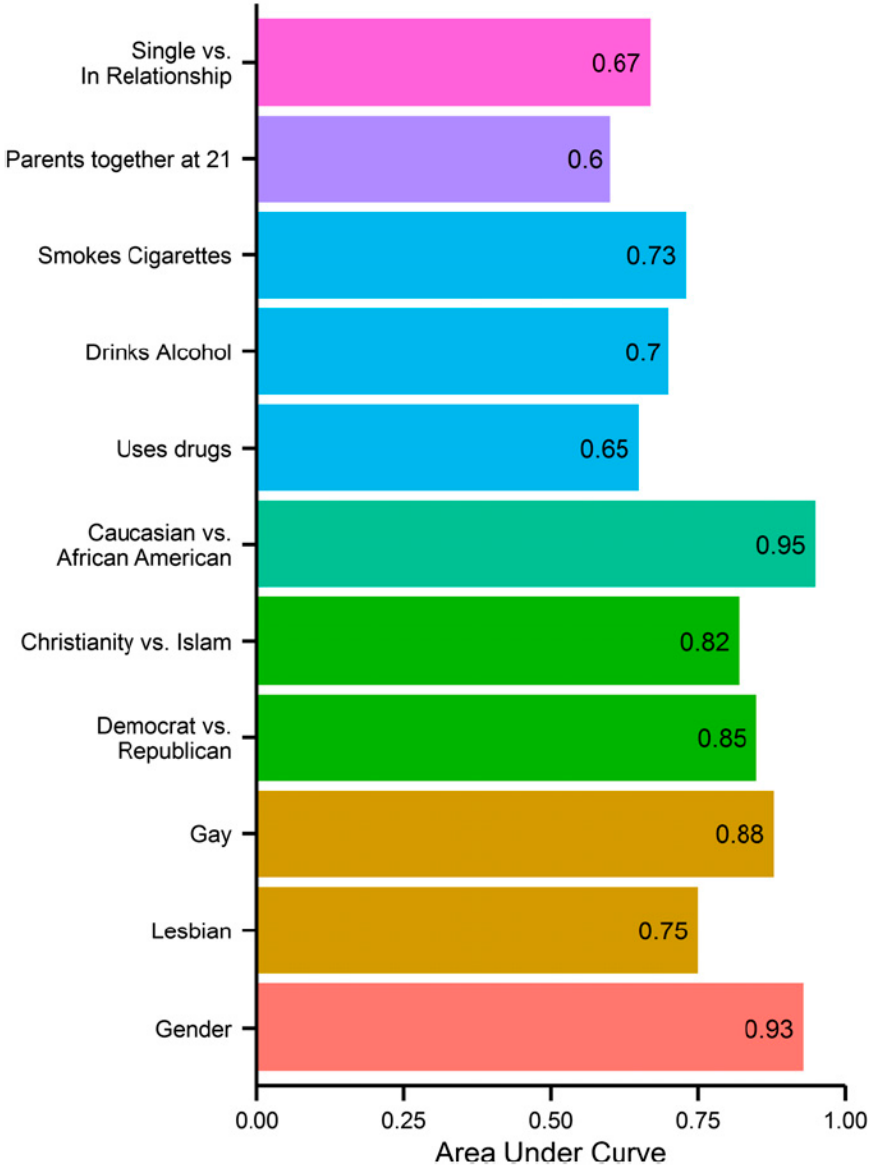
where L is the set of distinct label values

Example use of AUC

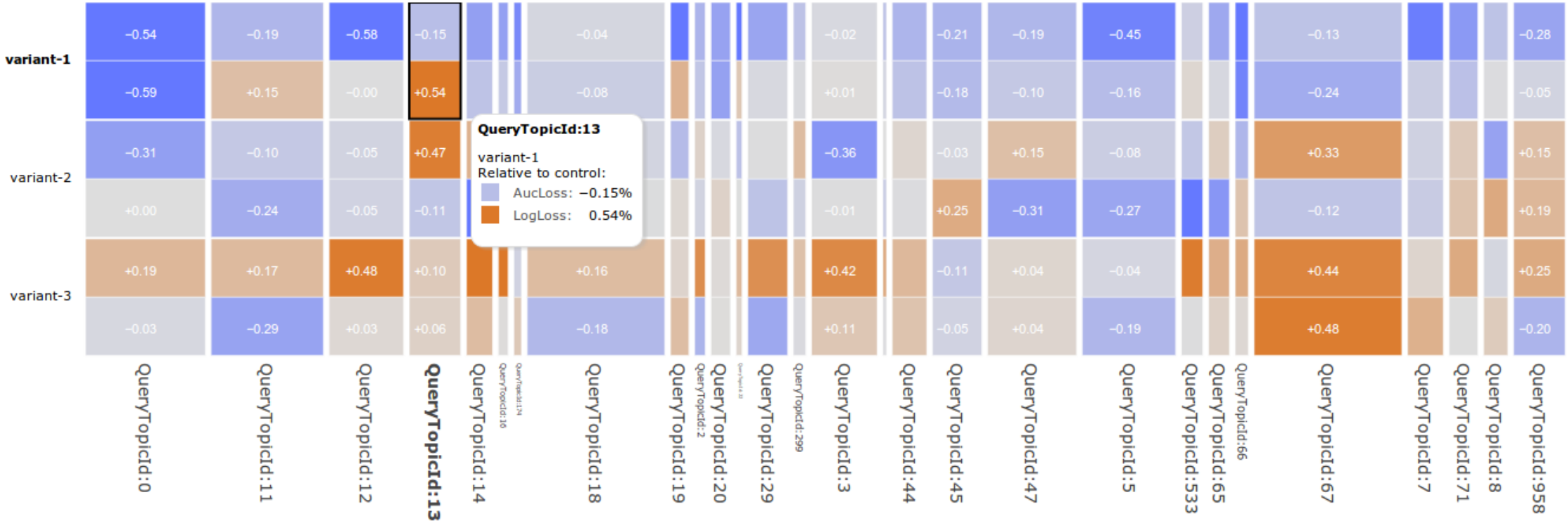
- Prediction classification accuracy for dichotomous / dichotomized attributes

- Kosinski et al, Personal traits and attributes are predictable from digital records of human behavior, PNAS, Vol 110, No 15, 2013

  http://www.pnas.org/content/110/15/5802.full



Another example



Screen shot of the high-dimensional analysis visualization. Here, three variants are compared with a control model, with results for AucLoss and LogLoss computed across a range of query topics. Column width reflects impression count. Detailed information pops up for a specific breakdown on mouse-over. The user interface allows for selection of multiple metrics and several possible breakdowns, including breakdowns by topic, country, match type, and page layouts. This allows fast scanning for anomalies and deep understanding of model performance. Best viewed in color.

McMahan et al, Ad-click prediction: a view from the trenches, ACM KDD 2013
https://www.eecs.tufts.edu/~dsculley/papers/ad-click-prediction.pdf

Summary

- Accuracy as a performance metric of a binary classifier should be taken with care because of its sensitivity to the true label distribution
- ROC analysis is a useful visual method for evaluating performance of classifiers
- ROC curves are insensitive to the true label distribution
- AUC defined as the area under ROC curve is insensitive to the true label distribution
- Precision-recall curve is an alternative method, not insensitive to the true label distribution

References

- J. Davis and M. Goadrich, The relationship between precision-recall and ROC curves, Proc. of ICML 2006
- T. Fawcett, An introduction to ROC analysis, Pattern recognition letters, Vol 27, pp 861-874, 2006
- P. Flach, J. Hernandez-Orallo and C. Ferri, A coherent interpretation of AUC as a measure of aggregated classification performance, Proc. of ICML 2011
- C. Gini, Reprinted: On the measurement of concentration and variability of characters, Metron LXIII(1), 338, 2005
- D. J. Hand and R. J. Till, A simple generalization of the area under the ROC curve to multiple class classification problems, Machine Learning, Vol 45, No 2, pp. 171-186, 2001
- N. Japkowicz and M. Shah, Evaluating Learning Algorithms: A Classification Perspective, Cambridge University Press, 2011
- F. Provost, T. Fawcett and R. Kohavi, The case against accuracy estimation for comparing induction algorithms, Proc. of ICML 1998
- J. A. Sweets, R. M. Dawes and J. Monahan, Betters Decisions through Science, Scientific American, pp 82-87, October 2000