Smooth Receiver Operating Characteristics Curves (smROC)

William Klement[1], Peter Flach[2], Nathalie Japkowicz[1], and Stan Matwin[1,3]

[1] School of Electrical Engineering and Computer science
University of Ottawa, Canada
[2] Dept. of computer Science, Bristol University, UK
[3] Institute of Computer Science, Polish Academy of Science, Poland.

Acknowledgement:
Natural Sciences and Engineering Research Council of Canada
Ontario Centres of Excellence.

Contribution

We develop an evaluation method which:

- extends the ROC to include membership scores
- allows the visualization of individual scores
- depicts the combined performance of classification, ranking and scoring

Consider what information can be obtained from testing a given learning method.

Learning Tasks

Prediction Outcomes

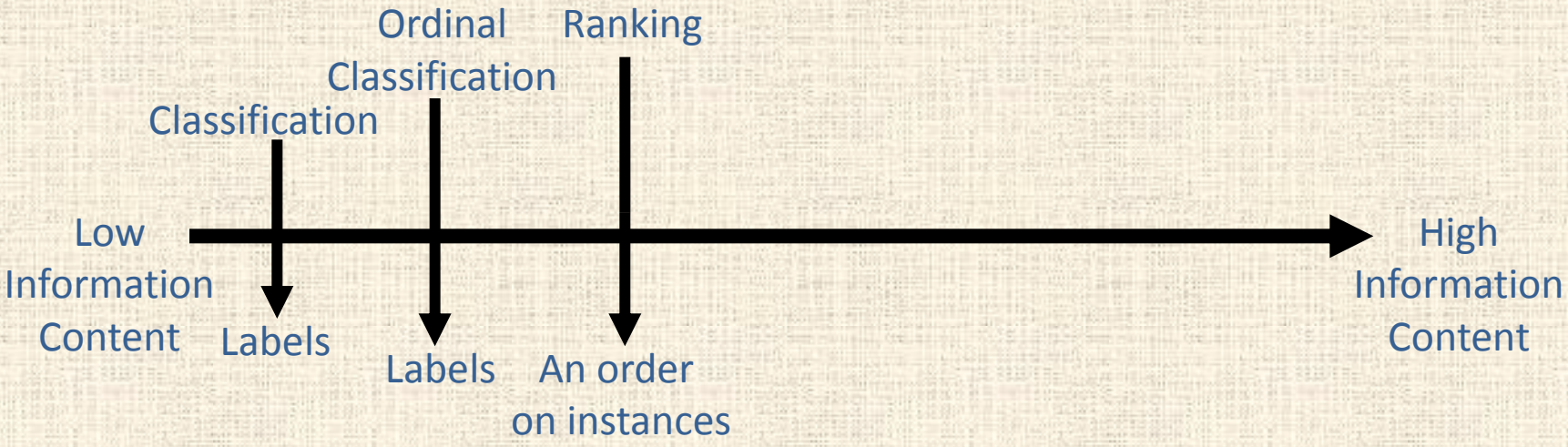
Learning Tasks

Prediction Outcomes

Learning Tasks

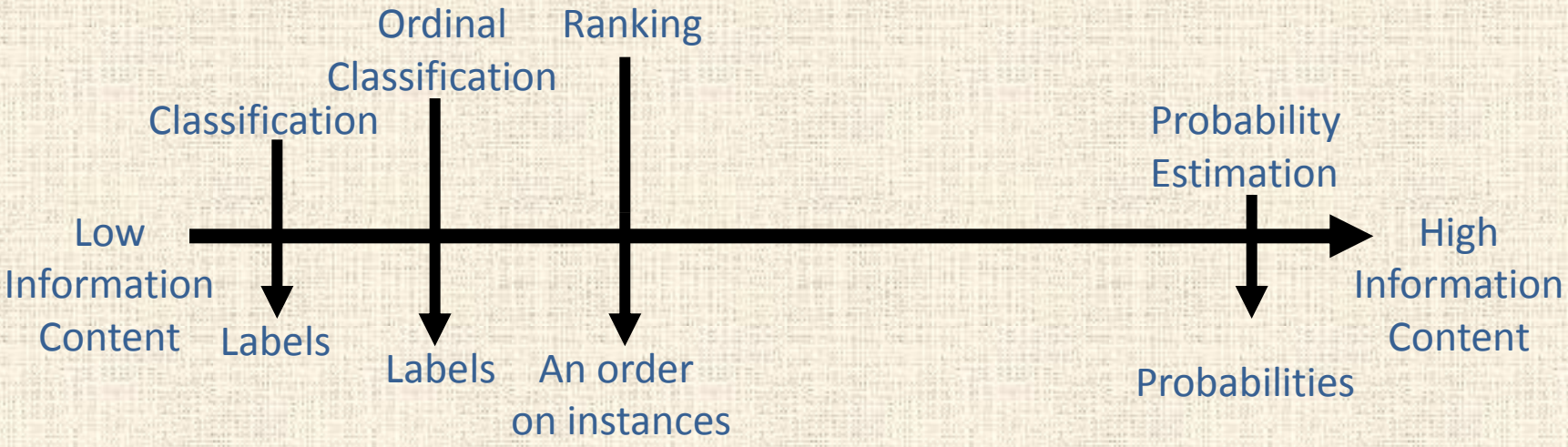
Prediction Outcomes

Learning Tasks



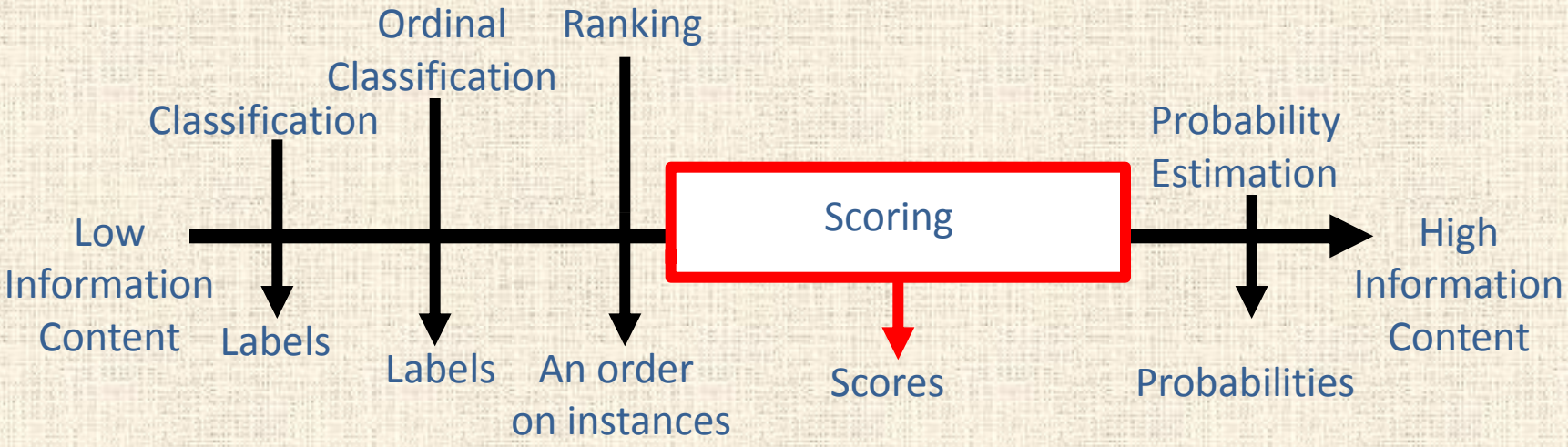
Prediction Outcomes

Learning Tasks



Prediction Outcomes

Learning Tasks



Prediction Outcomes

- Imposing a threshold (on the scores then ignoring them) reduces the task into a classification.
- Sorting the data points (by scores then ignoring them) reduces the task into a ranking.

Motivation

- With scores, one can:
  - compare classifications in terms of decisions, ranking, and scores (confidence)
  - visualize the margins of scores
  - find gaps in scores
- Of course, probabilities tell us all this plus more (theoretical), but not all scores are good estimates of probabilities!

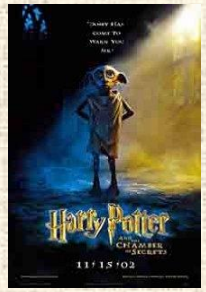
Applications

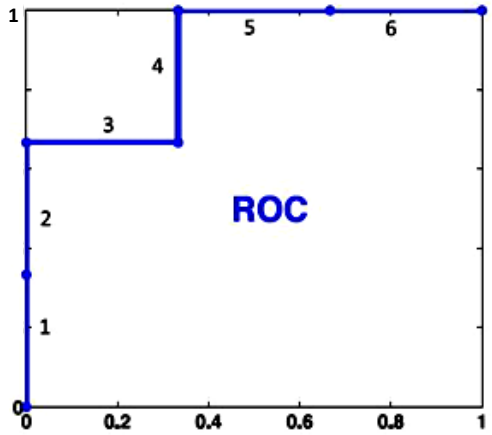
- Comparing user preferences
- Assessing relevance scores in search engines
- Magnitude-preserving ranking (Cortes et. al ICML'07)
- Research Tool (PET / DT / Naïve Bayes)
- Bioinformatics (gene expression)

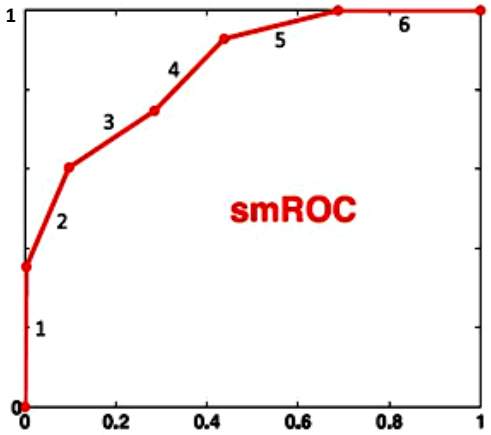
An Example: Movie Recommendation

Anna 

 Jan



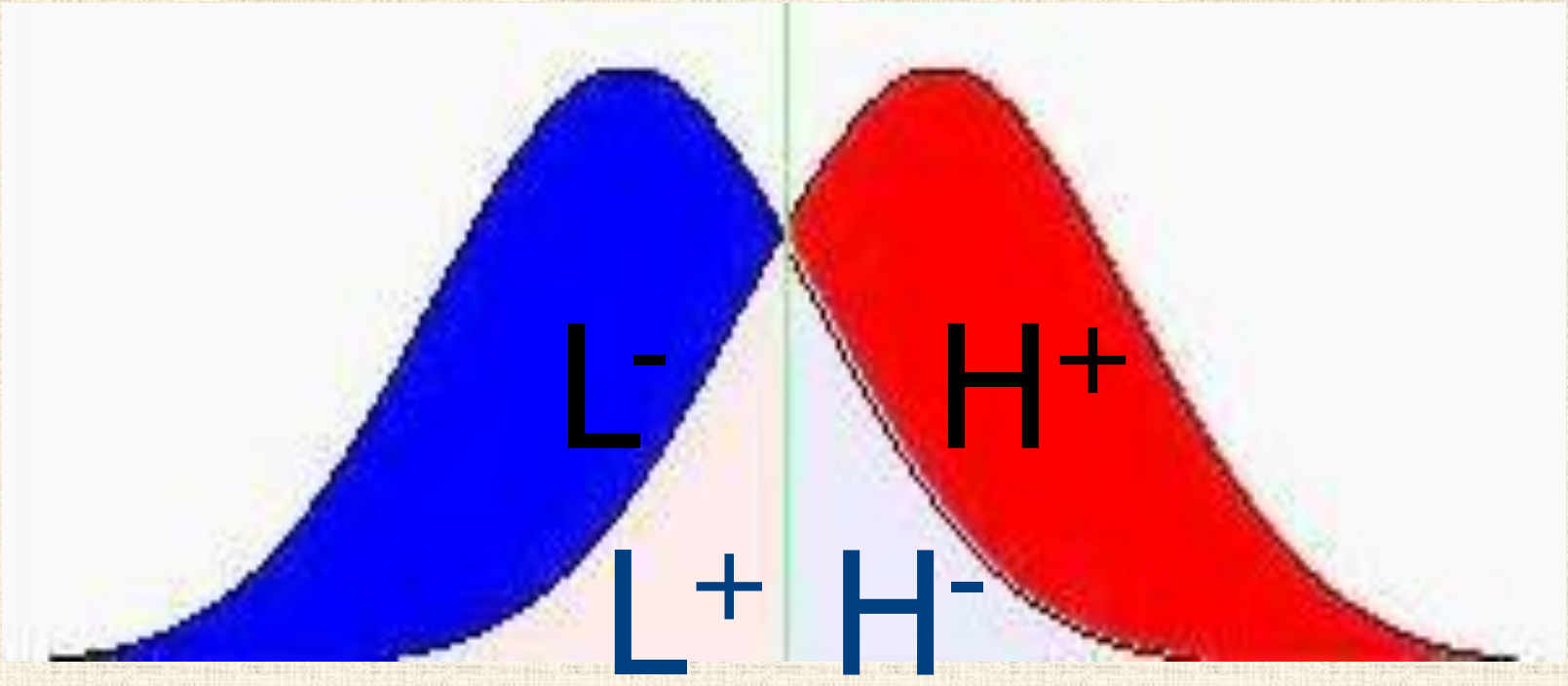




Anna's Assessment

i	Decision	Score
1.	+	0.99
2.	+	0.70
3.	-	0.60
4.	+	0.51
5.	-	0.20
6.	-	0.00

Methodology



$$\Theta(x_i) = \begin{cases} s_i & \text{if } x_i \in \{H^+ \cup L^-\} \text{ (Appropriate Scores)} \\ 1 - s_i & \text{if } x_i \in \{H^- \cup L^+\} \text{ (Inappropriate Scores)} \end{cases}$$

Methodology: Score Appropriateness

(Appropriateness of Scores)

Label	Scores: High	Low
+	yes	no
−	no	yes

(Accuracy of Appropriate Scores)

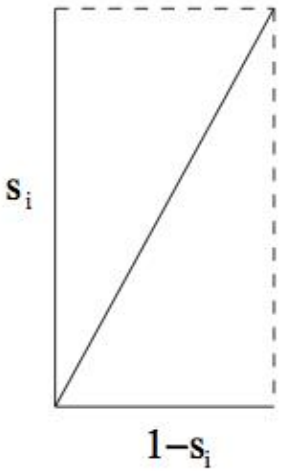
Score Label		Predicted: Y	N
High	+	correct	incorrect
Low	−	incorrect	correct

(Accuracy of Inappropriate Scores)

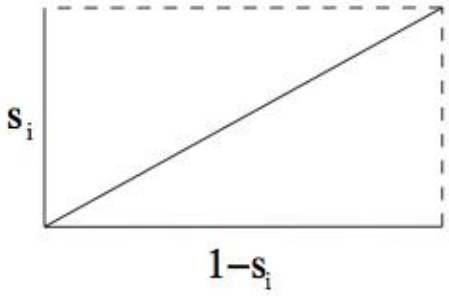
Score Label		Predicted: Y	N
High	−	incorrect	correct
Low	+	correct	incorrect

Appropriate

Positive Instance

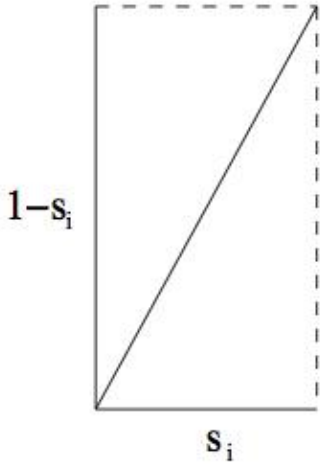


Negative Instance

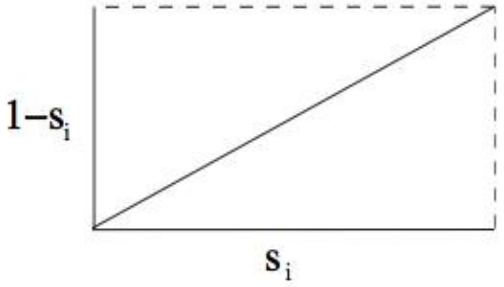


Inappropriate

Positive Instance



Negative Instance



Constructing the smROC Curve

$$Mid = \frac{1}{2}(m^{+} + \frac{m^{-}}{c})$$

$$smTPR = \frac{\Theta(x_i)}{\propto_v} \qquad smFPR = \frac{\Theta(x_i)}{\propto_h}$$

$$\propto_v = \sum_{i=1}^{|H^+|} S_i + \sum_{i=1}^{|L^-|} S_i + \sum_{i=1}^{|L^+|} (1 - S_i) + \sum_{i=1}^{|H^-|} (1 - S_i) = \sum_{i=1}^{n} \Theta(x_i)$$

$$\propto_h = \sum_{i=1}^{|H^+|} (1 - S_i) + \sum_{i=1}^{|L^-|} (1 - S_i) + \sum_{i=1}^{|L^+|} S_i + \sum_{i=1}^{|H^-|} S_i = \sum_{i=1}^{n} (1 - \Theta(x_i))$$

smAUC

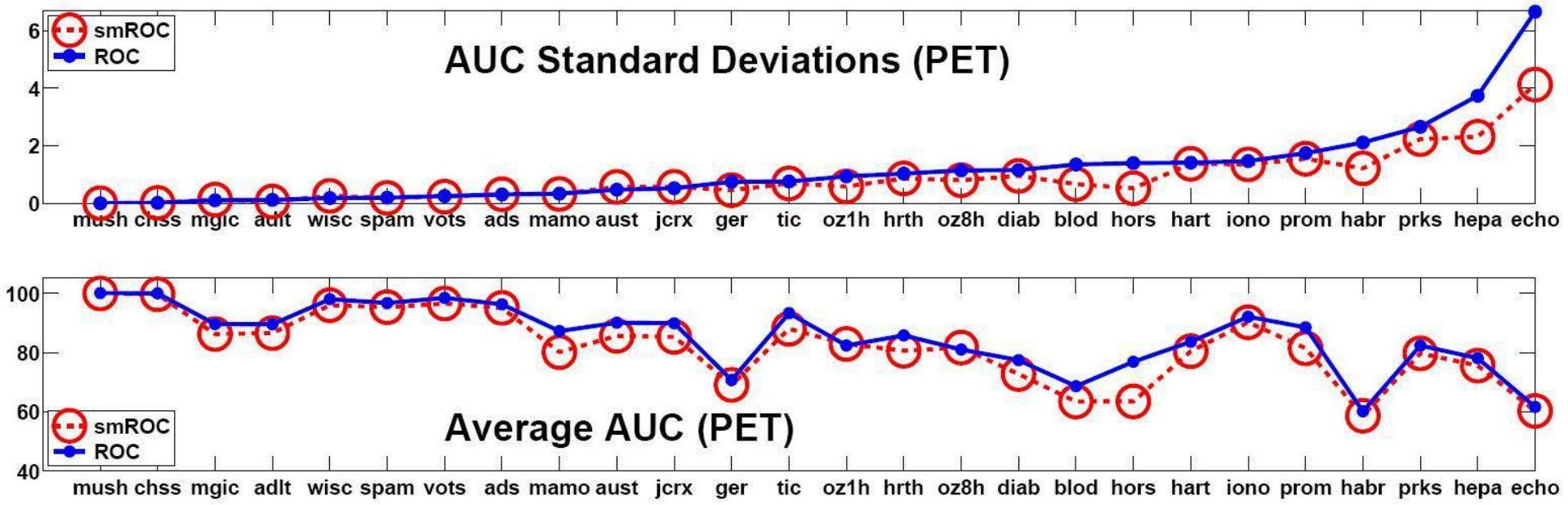
$$smAUC = \frac{1}{\propto_v \propto_h} \sum_{i=1}^{n} \sum_{j=1}^{n} \Theta(x_i) \Psi(x_i, x_j)$$

$$\Psi(x_i, x_j) = \begin{cases} 1 - \Theta(x_i) & \text{for } (S_i > S_j) \text{ and } (i \neq j) \\ \frac{1}{2}(1 - \Theta(x_i)) & \text{for } i = j \\ 0 & \text{otherwise} \end{cases}$$

Experiment

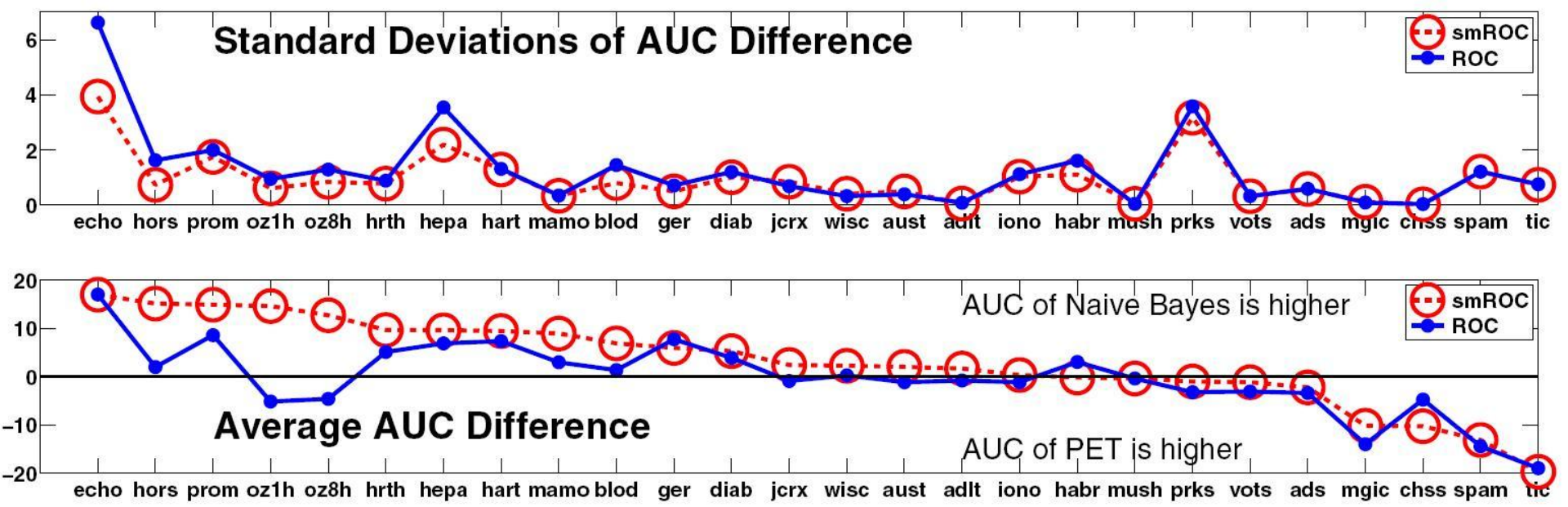
- Use 26 UCI data sets of binary classification problems.
- Classification by PET and Naïve Bayes.
- Test by 10-fold cross-validation repeated 10 times.
- Measure performance similarities among similar models (same learning method on various random splits of the same data).
- Verify well-documented performance differences of PET and NB (different methods on the same data).
- Record the average and standard deviation of smAUC and AUC.

Similar PET Models



- Lower std. dev. for smAUC with increasing variations
- smAUC is lower than AUC

PET & Naive Bayes Differences



- smAUC measures a higher difference in favour of Naive Bayes scores.
- AUC = smAUC in favour of PET.
- Lower std. dev. of smAUC difference.

Conclusions & Future Plans

- smROC is sensitive to scores assigned to data points by the classifier but retains sensitivity to ranking performance.
- smROC is more sensitive to performance similarities and differences between scores.
- For similarities models, smAUC produces lower std. deviations, and for different ones, the difference in the smROC space is higher.
- smROC can be sensitive to changes in the underlying distribution of data and scores (sensitivity to the mid point?).