Part III: Multi-class ROC

- The general problem
  - multi-objective optimisation
  - Pareto front
  - convex hull
- Searching and approximating the ROC hyper-surface
  - multi-class AUC
  - multi-class calibration

The general problem

- Two-class ROC analysis is a special case of multi-objective optimisation
  - don't commit to trade-off between objectives
- Pareto front is the set of points for which no other point improves all objectives
  - points not on the Pareto front are dominated
  - assumes monotonic trade-off between objectives
- Convex hull is subset of Pareto front
  - assumes linear trade-off between objectives
    - e.g. accuracy, but not precision

How many dimensions?

- Depends on the cost model
  - 1-vs-rest: fixed misclassification cost $C(\neg c|c)$ for each class $c \in C$ —> $|C|$ dimensions
    - ROC space spanned by either tpr for each class or fpr for each class
  - 1-vs-1: different misclassification costs $C(c_i|c_j)$ for each pair of classes $c_i \neq c_j$ —> $|C|(|C|-1)$ dimensions
    - ROC space spanned by fpr for each (ordered) pair of classes
- Results about convex hull, optimal point given linear cost function etc. generalise
  - (Srinivasan, 1999)

Multi-class AUC

- In the most general case, we want to calculate Volume Under ROC Surface (VUS)
  - See (Mossman, 1999) for VUS in the 1-vs-rest three-class case

- Can be approximated by projecting down to set of two-dimensional curves and averaging
  - MAUC (Hand & Till, 2001): 1-vs-1, unweighted average
  - (Provost & Domingos, 2001): 1-vs-rest, AUC for class c weighted by P(c)

Multi-class calibration

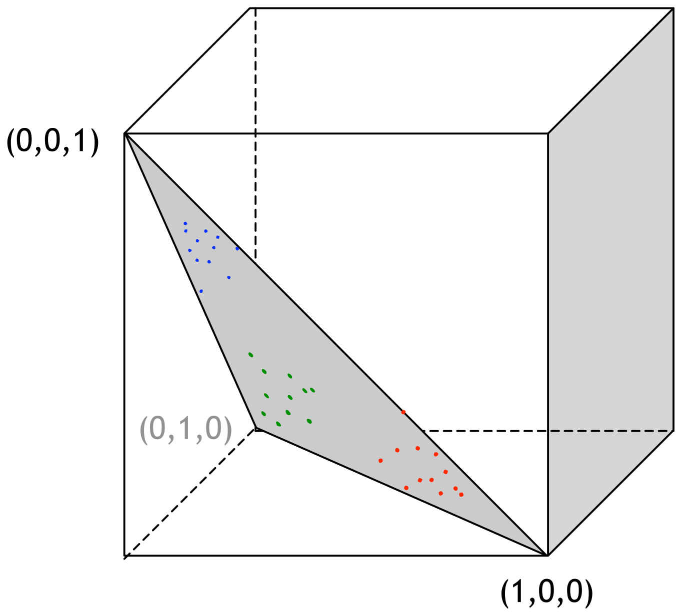
1. How to manipulate scores f(x,c) in order to obtain different ROC points?
   - depends on the cost model

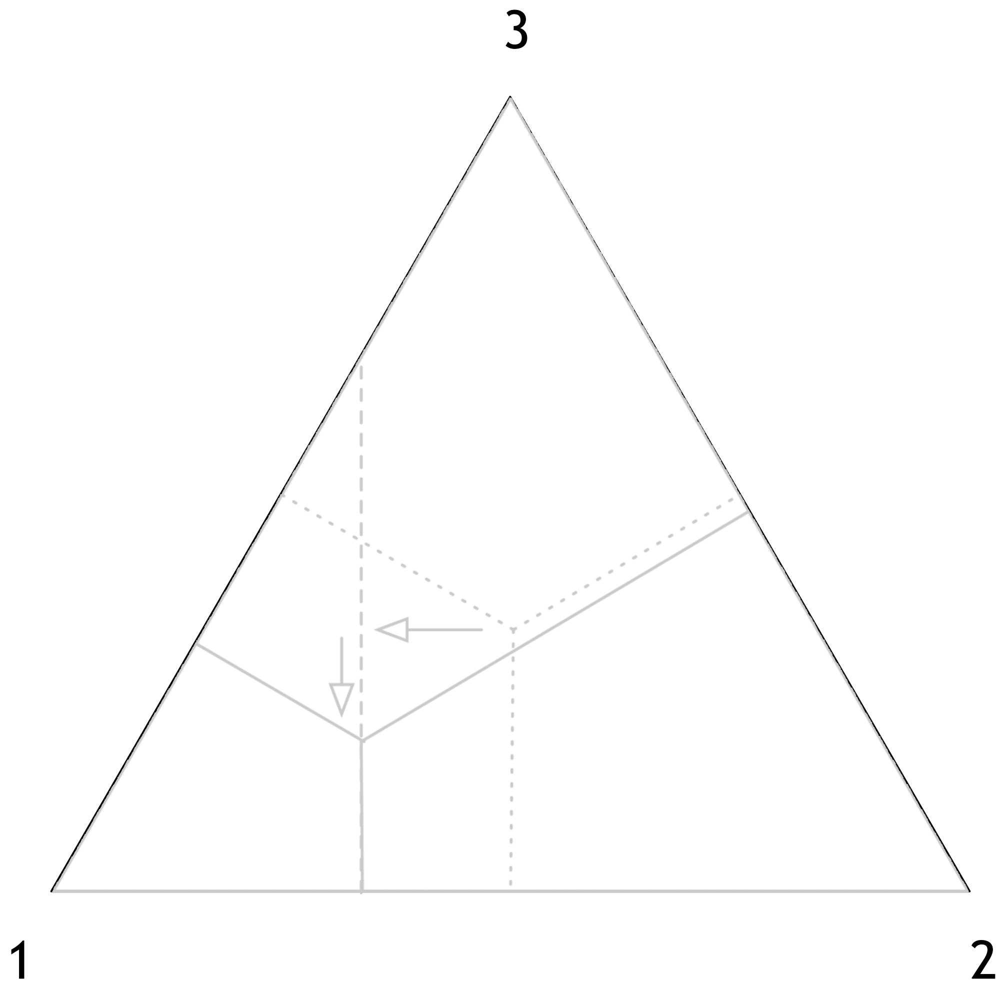
2. How to search these ROC points to find optimum?
   - exhaustive search probably infeasible, so needs to be approximated

A simple 1-vs-rest approach

1. From thresholds to weights:
   - predict $\text{argmax}_c\ w_c\ f(x,c)$
   - NB. two-class thresholds are a special case:
     - $w_+ f(x,+) > w_- f(x,-) \Leftrightarrow f(x,+)/f(x,-) > w_-/w_+$
2. Setting the weights (Lachiche & Flach, 2003)
   - Assume an ordering on classes and set the weights in a greedy fashion
     - Set $w_1 = 1$
     - For classes c=2 to n
       - look for the best weight w_c according to the weights fixed so far for classes c'<c, using the two-class algorithm

Example: 3 classes





Discussion

- Strong experimental results
  - 13 significant wins (95%), 22 draws, 2 losses on UCI data
- Sensitive to the ordering of classes
  - largest classes first is best
- No guarantee to find a global (or even a local) optimum
  - lots of scope for improvement, e.g. stochastic search

The many faces of ROC analysis

- ROC analysis for model evaluation and selection
  - key idea: separate performance on classes
  - think rankers, not classifiers!
  - information in ROC curves not easily captured by statistics
- ROC visualisation for understanding ML metrics
  - towards a theory of ML metrics
    - types of metrics, equivalences, skew-sensitivity
- ROC metrics for use within ML algorithms
  - one classifier can be many classifiers!
  - separate skew-insensitive parts of learning…
    - probabilistic model, unlabelled tree
  - …from skew-sensitive parts
    - selecting thresholds or class weights, labelling and pruning

Outlook

- Several issues not covered in this tutorial
  - optimising AUC rather than accuracy when training (several papers at ICML’03 and ICML’04)
    - e.g. RankBoost optimises AUC (Cortes & Mohri, 2003)

- Many open problems remain
  - ROC analysis in rule learning
    - overlapping rules
  - relation between training skew and testing skew
  - multi-class ROC analysis

References

C. Cortes and M. Mohri (2003). AUC optimization vs. error rate minimization. In Advances in Neural Information Processing Systems (NIPS'03). MIT Press.

T.G. Dietterich, M. Kearns, and Y. Mansour (1996). Applying the weak learning framework to understand and improve C4.5. In L. Saitta, editor, Proc. 13th International Conference on Machine Learning (ICML'96), pp. 96-103. Morgan Kaufmann.

C. Drummond and R.C. Holte (2000). Exploiting the cost (in)sensitivity of decision tree splitting criteria. In P. Langley, editor, Proc. 17th International Conference on Machine Learning (ICML'00), pp. 239-246.

T. Fawcett (2004). ROC graphs: Notes and practical considerations for data mining researchers. Technical report HPL-2003-4, HP Laboratories, Palo Alto, CA, USA. Revised March 16, 2004. Available at http://www.purl.org/NET/tfawcett/papers/ROC101.pdf.

C. Ferri, P.A. Flach, and J. Hernández-Orallo (2002). Learning Decision Trees Using the Area Under the ROC Curve. In C. Sammut and A. Hoffmann, editors, Proc. 19th International Conference on Machine Learning (ICML'02), pp. 139-146. Morgan Kaufmann.

P.A. Flach (2003). The geometry of ROC space: understanding machine learning metrics through ROC isometrics. In T. Fawcett and N. Mishra, editors, Proc. 20th International Conference on Machine Learning (ICML'03), pp. 194-201. AAAI Press.

P.A. Flach and S. Wu (2003). Reparing concavities in ROC curves. In J.M. Rossiter and T.P. Martin, editors, Proc. 2003 UK workshop on Computational Intelligence (UKCI'03), pp. 38–44. University of Bristol.

J. Fürnkranz and P.A. Flach (2003). An analysis of rule evaluation metrics. In T. Fawcett and N. Mishra, editors, Proc. 20th International Conference on Machine Learning (ICML'03), pp. 202-209. AAAI Press.

J. Fürnkranz and P.A. Flach (forthcoming). ROC 'n' rule learning — towards a better understanding of covering algorithms. Machine Learning, accepted for publication.

D. Gamberger and N. Lavrac (2002). Expert-guided subgroup discovery: methodology and application. Journal of Artificial Intelligence Research, 17, 501-527.

D.J. Hand and R.J. Till (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems, Machine Learning, 45, 171-186.

N. Lachiche and P.A. Flach (2003). Improving accuracy and cost of two-class and multi-class probabilistic classifiers using ROC curves. In T. Fawcett and N. Mishra, editors, Proc. 20th International Conference on Machine Learning (ICML'03), pp. 416-423. AAAI Press.

A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki (1997). The DET curve in assessment of detection task performance. In Proc. 5th European Conference on Speech Communication and Technology, vol. 4, pp. 1895-1898.

D. Mossman (1999). Three-way ROCs. Medical Decision Making 1999(19): 78-89.

F. Provost and T. Fawcett (2001). Robust classification for imprecise environments. Machine Learning, 42, 203-231.

F. Provost and P. Domingos (2003). Tree induction for probability-based rankings. Machine Learning 52:3.

A. Srinivasan (1999). Note on the location of optimal classifiers in n-dimensional ROC space. Technical Report PRG-TR-2-99, Oxford University Computing Laboratory.

B. Zadrozny and C. Elkan (2002). Transforming classifier scores into accurate multiclass probability estimates. In Proc. 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'02), pp. 694-699.

See also http://splweb.bwh.harvard.edu:8000/pages/ppl/zou/roc.html for pointers to ROC literature.

Acknowledgements

- Many thanks to:
  - Johannes Fürnkranz, Cèsar Ferri, José Hernández-Orallo, Nicolas Lachiche & Shaomin Wu for joint work on ROC analysis and for some of the material
  - Jim Farrand & Ronaldo Prati for ROC visualisation software
  - Chris Drummond & Rob Holte for material and discussion on cost curves
  - Tom Fawcett & Rich Roberts for some of the ROC graphs
  - Terri Kamm for the DET slide