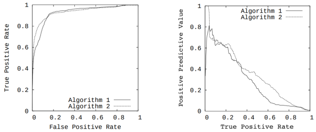
Advanced Machine Learning

Imbalanced Learning: Ranking Metrics



(a) Comparison in ROC space   (b) Comparison in PR space

Learning goals

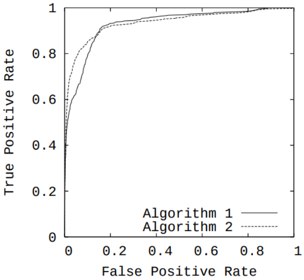
- Know the limits of ROC curves for imbalanced data
- Understand PR curves
- Understand cost curves for cost-sensitive learning settings

ROC CURVES FOR SCORING CLASSIFIERS

- Many binary classification methods use a score (function) $s : \mathcal{X} \to \mathbb{R}$ and a threshold value c to make the prediction:

$$f(\mathbf{x}) = 2 \cdot \mathbb{1}_{[s(\mathbf{x}) \geq c]} - 1.$$

- The choice of threshold affects the TPR and FPR, so it is interesting to examine the effects of different thresholds on these.
- A ROC curve is a visual tool to help in finding good threshold values.



Davis and Goadrich (2006): The Relationship Between Precision-Recall and ROC Curves (URL).

FPR VS. PPV

- For imbalanced data sets, i.e., if $n_- \gg n_+$, if the number of negative instances is large, then we are typically less interested in a high TNR or equivalently a low FPR.
- Instead PPV (precision) is more interesting, i.e., the fraction of positives we correctly classified among all positive classifications.
- The reason is that a large (absolute) change in the number of FP yields only a small change in FPR, but the PPV might be affected significantly and consequently is more informative.

FP=10:

	True +1	True -1
Pred. Pos	100	10
Pred. Neg	10	9990
Total	110	10000

$$\text{TPR} = 10/11$$
$$\text{FPR} = 0.001$$
$$\text{PPV} = 10/11$$

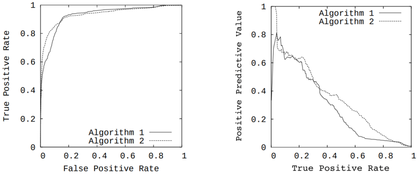
FP=100:

	True +1	True -1
Pred. +1	100	100
Pred. -1	10	9900
Total	110	10000

$$\text{TPR} = 10/11$$
$$\text{FPR} = 0.01$$
$$\text{PPV} = 1/2$$

PRECISION-RECALL CURVES

- Instead of the TPR-FPR curve in ROC plots, we could consider TPR-PPV curves. These are also called Precision-Recall (PR) curves, since precision = $\rho_{PPV} = \frac{TP}{TP+FP}$ and recall = $\rho_{TPR} = \frac{TP}{TP+FN}$.
- Note that performance measures such as the F_1 score or the G score also consider exactly the trade-off between precision and recall rather than the TPR-FPR tradeoff.
- Considering the figures below: both learning algorithms seem to perform well w.r.t. the ROC curves, but the PR-curves reveal that:
  1. both algorithms have room for improvement, as the best we can get is in the top-right corner,
  2. the 2nd algorithm is better than the 1st in terms of recall-precision combinations.

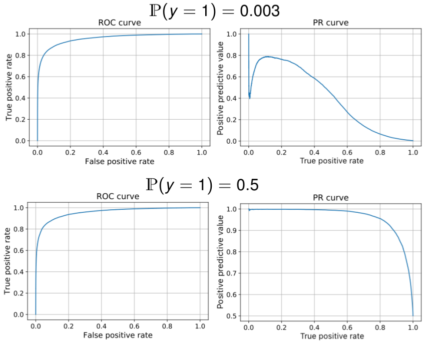


(a) Comparison in ROC space (b) Comparison in PR space

Davis and Goadrich (2006): The Relationship Between Precision-Recall and ROC Curves (URL).

ROC/PR CURVES FOR IMBALANCED DATA

- Top row: Imbalanced classes with $\mathbb{P}(y = 1) = 0.003$.
- Bottom row: Balanced classes, i.e., $\mathbb{P}(y = 1) = 0.5$.
- Note that the ROC curves (LHS) are similar in both cases. However, the PR curve (RHS) changes strongly from the imbalanced to the balanced case.



Wissam Siblini et. al. (2004): Master your Metrics with Calibration (URL).

EXPECTED COSTS

In the case of cost-sensitive learning, where we are provided with a cost matrix **C** (no costs for correct classifications), we can compute the expected costs of a classifier f with the corresponding prediction (random) variable $\hat{y} = f(\mathbf{x})$ as follows:

$$\begin{aligned}
\mathbb{E}[\mathrm{Cost}(f)] &= \mathbb{P}(\hat{y} = 1, y = -1)C(+1,-1) + \mathbb{P}(\hat{y} = -1, y = 1)C(-1,+1) \\
&= \mathbb{P}(\hat{y} = 1 \mid y = -1)\mathbb{P}(y = -1)C(+1,-1) \\
&\quad + \mathbb{P}(\hat{y} = -1 \mid y = 1)\mathbb{P}(y = 1)C(-1,+1) \\
&= \mathrm{FPR} \cdot \mathbb{P}(y = -1)C(+1,-1) \\
&\quad + \mathrm{FNR} \cdot \mathbb{P}(y = 1)C(-1,+1) \\
&= \mathrm{FPR} \cdot (1 - \mathbb{P}(y = 1))C(+1,-1) + \mathrm{FNR} \cdot \mathbb{P}(y = 1)C(-1,+1)
\end{aligned}$$

Since the highest costs are achieved if $\mathrm{FPR} = \mathrm{FNR} = 1$ we can normalize the expected costs:

$$\mathbb{E}[\mathrm{Cost}_{norm}(f)] = \frac{\mathrm{FPR} \cdot (1 - \mathbb{P}(y = 1))C(+1,-1) + \mathrm{FNR} \cdot \mathbb{P}(y = 1)C(-1,+1)}{(1 - \mathbb{P}(y = 1))C(+1,-1) + \mathbb{P}(y = 1)C(-1,+1)}$$

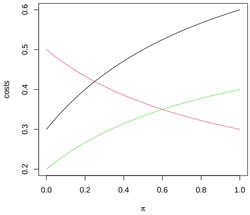
Setting $\pi_+ = \mathbb{P}(y = 1)$, this specifies a function in π_+ :

$$\mathrm{Costs}_{norm}(\pi_+ | f) = \frac{\mathrm{FPR} \cdot (1 - \pi_+)C(+1,-1) + \mathrm{FNR} \cdot \pi_+ C(-1,+1)}{(1 - \pi_+)C(+1,-1) + \pi_+ C(-1,+1)}$$

COST CURVES

- A cost curve of a classifier f is simply the plot of the function $\pi_+ \mapsto \mathrm{Costs}_{norm}(\pi_+|f)$.
- Consider as an example the cost matrix **C** such that $C(+1,-1) = 1$ and $C(-1,+1) = 2$. Moreover, consider three classifiers with:

  - 1st classifier: FNR = 0.6, FPR = 0.3
  - 2nd classifier: FNR = 0.3, FPR = 0.5
  - 3rd classifier: FNR = 0.4, FPR = 0.2



- 3rd classifier is always better than the 1st, while 2nd classifier is better than the 3rd if $\pi_+ \geq 0.6$.

COST CURVES

- In the case of equal costs for false positives and false negatives, the cost curves correspond to straight lines.
- The perfect classifier is in every case the one which has for every π_+ (normalized) costs of zero. The worst classifier is in every case the one having (normalized) costs of 1.
- The (normalized) cost curves of the always-positive classifier and the always-negative classifier meet each other at $c^* = \frac{C(1,-1)}{C(-1,1)+C(1,-1)}$, i.e., the perfect threshold value we derived to minimize the expected costs for predicting the positive class.