Portfolio Question 1.2.1 - Metrics

Daniel Lawson University of Bristol

01.2.1 (v4.0.0)

Portfolio Example Question

- Portfolio Question 1.2: When is an accurate classifier not useful? Can we ensure it is useful?
  - What metrics are better than accuracy? In what sense are they reliable, and do they relate to uncertainty?
  - Hints: take a look at Proper Scoring Rules
- Format
  - Introduction
  - Non-trivial implication
  - Discussion

Context

- We were introduced to ROC (Receiver-Operating Curves) in Lecture 1.2
- Definition:
  - **Accuracy** is the proportion of correctly classified points.
    - But... what is the **baseline**? If there are imbalanced classes?
- If we optimise for a metric with a given test/train setup:
  - Our model will perform well for that setup
- Can we say anything more?
  - Can we reflect **real-world** use?

Why is Accuracy bad?

- Accuracy is a natural metric
- It goes wrong if classes are imbalanced, e.g.
  - A 99-1 split between classes A and B
  - Always guess A
  - Gets 99% accuracy
- It also ignores uncertainty
  - May reward overly-certain claims more than calibrated uncertainty

Metrics for Classification

- Start with the ROC curve:
  - TPR **at a given FPR**
  - AUC characterises the whole ROC curve
- Area Under Precision-Recall Curve (AUPRC) is also a thing people advocate for
- None are "right", we have to define the inference task
- Any of these and more are often optimized
  - If we optimise a parameter or perform model comparison based on test data, we need additional test data to test the meta-algorithm!

PR curve

- Nice things[1] about ROC and PR curves:
- Dominance:
  - If one curve dominates (is always above) another in ROC, it dominates in PR
  - and vice-versa
- ROC curves can be linearly interpolated
  - This is "flipping a coin" to access classifiers in-between
- PR curves have a slightly more complex relationship but the same principle can be applied
- Integrating both scores leads to performance metric that can be optimized

[1] Davis and Goadrich, "The Relationship Between Precision-Recall and ROC Curves", ICML 2006.

Non-trivial point: towards Proper Scoring Rules

- Use $P(Y|X)$ as the loss:
- 'A scoring rule is proper if the forecaster maximizes the expected score for an observation drawn from the distribution F if he or she issues the probabilistic forecast F, rather than $G \neq F$. ... proper scoring rules encourage the forecaster to make careful assessments and to be honest.'
  - Gneiting and Raftery 2007.
- Easy wikipedia article
- Examples include: Log-likehihood, Brier score $\sum_{i=1}^{n}(y_i - \hat{p}_i(\mathbf{x_i}))^2$
- Takehome: We need to keep probabilities in the model

Discussion

- What do we want from a loss function?
- What is the relationship between a loss function and a model?
- Does finding parameters to minimise the loss mean that the model is right?
  - How does a proper score rule work on the logistic regression example in class?
- Is this the same as calibration?
- When should we care?
  - Is this advise still up to date?
  - Do e.g. large language models care about this?
  - How could it work for imaging?