Evaluating Classifier Performance

Ryan Miller

Grinnell College
Statistics

Motivating Example

Consider 2 classifiers that output the following predicted probabilities (of cancer) on a small test set of 4 individuals:

Person	True Outcome	Classifier A	Classifier B
1	Cancer	0.55	0.98
2	No Cancer	0.45	0.01
3	Cancer	0.53	0.95
4	No Cancer	0.47	0.05

- Both classifiers achieve 100% accuracy
  - But is one classifier better than the other?

Prediction Scores (probabilities)

- Most machine learning algorithms do not directly produce predicted classes, instead they produce *prediction scores* (often probabilities)
- Class labels derived from these scores, with a natural choice being the label with the highest score
  - For binary classification, this means comparing the prediction score to a threshold of $t = 0.5$

Confusion Matrices

Applying a threshold to map scores to predicted categories allows us to create a **confusion matrix**:

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Note: in this setup the "positive class" is determined by the analyst.

Confusion Matrices (cont.)

Below is the confusion matrix for $t = 0.5$ using LOOCV to evaluate a KNN model ($k = 8$) in a study (Sobar, 2016) that used non-invasive social and behavioral assessments to predict cervical cancer:

	Predicted Cancer	Predicted Healthy
Has Cervical Cancer	13	8
Doesn't Have Cancer	1	50

- The *true positive rate* (TPR) is given by $\frac{\text{TP}}{\text{Total positives}} = 13/21 = 61.9\%$
- The *false positive rate* (FPR) is given by $\frac{\text{FP}}{\text{Total negatives}} = 1/50 = 2\%$

Confusion Matrices (cont.)

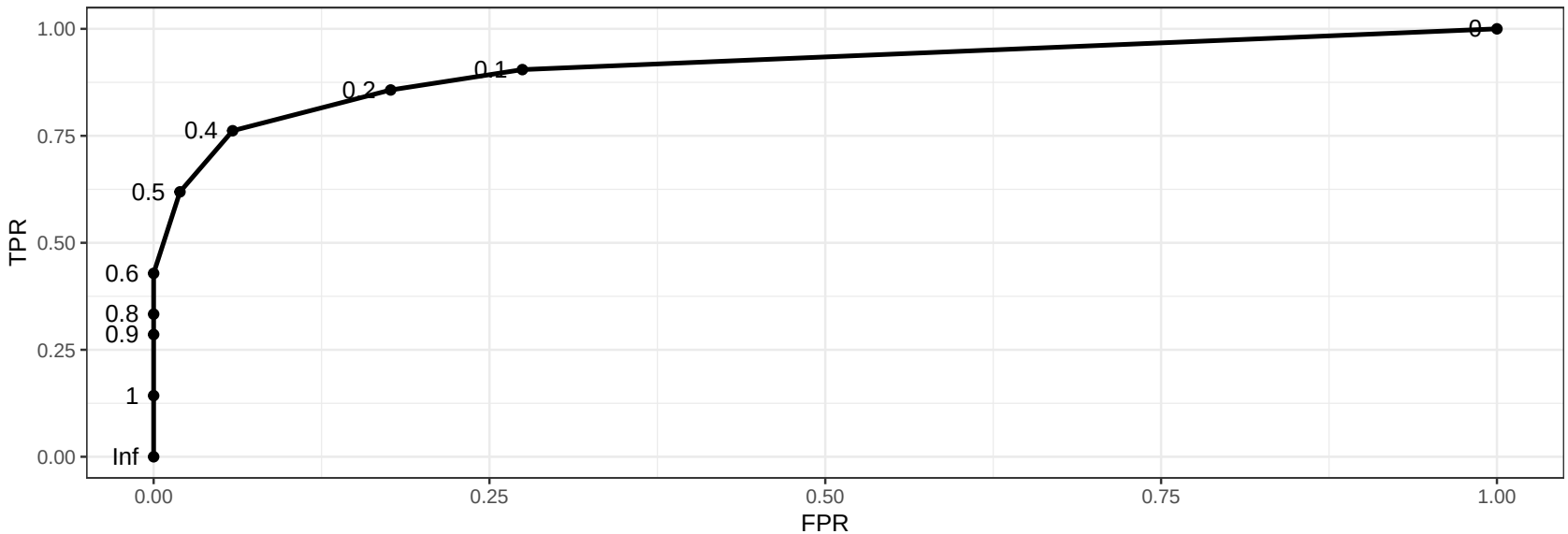
We might consider the confusion matrix after reducing t from 0.5 to 0.375, or lowering the threshold needed to a "cancer" prediction:

	Predicted Cancer	Predicted Healthy
Has Cervical Cancer	16	5
Doesn't Have Cancer	3	48

The TPR is now 76.2% (compared to 61.9%), but the FPR is 5.9% (compared to 2.0%)

Receiver Operating Characteristics (ROC)

The **receiver operating characteristics (ROC) curve** displays what happens to the TPR and FPR as the classification threshold, t, moves from 0 to 1.



Receiver Operating Characteristics (ROC)

- The ROC curve must include (0,0) at $t = \infty$ (since no observations are predicted as the positive class) and (1,1) at $t = 0$ (since all observations as predicted as the positive class)
  - Thus, a useless classifier would simply connect these two endpoints with a 45-degree line
  - Thresholds producing a combination of TPR and FPR above that line are potentially interesting

AUC

- The trade off between TPR and FPR across a range of classification thresholds that is encapsulated by the ROC curve is often summarized using the *area under the curve* (AUC)
  - For the reasons explained on the previous slide, $AUC = 0.5$ reflects a useless classier that does no better than randomly guessing class labels
  - $AUC = 1$ reflects perfect performance
- Using ROC-AUC to evaluate a classifier often provides a more nuanced performance assessment than classification accuracy

Class Imbalances

- Another strength of AUC is that it is not influenced by imbalances in the class distribution
  - In our example, 70.8% of the data set did not have cancer, so we could achieve 70.8% classification accuracy *without the model adding any predictive value*
- As a more extreme example, research by the St Louis federal reserve found the delinquency rate across all US loans to be 1.2%
  - In this application, would you be impressed by an algorithm that correctly classifies loan delinquency with 99% accuracy?

Balanced Accuracy

- **Balanced accuracy** addresses class imbalance by averaging the accuracy within each class (ie: averaging the accuracy achieved in each row of the confusion matrix).
- For binary classification, this amounts to:

$$\text{Balanced Accuracy} = \frac{\text{TPR}+(1-\text{FPR})}{2}$$

- Note that this is the average of the TPR and the *true negative rate*.

Precision and Recall

- TPR and FPR both address the question "how many of each observed class are classified as positive?"
  - **Precision** answers the question "how many of the samples that were classified as positive are actually positive?"
  - **Recall** is often analyzed in conjunction with precision, but it's just another term for TPR

$$\text{Precision} = \frac{\text{True Positives}}{\text{Total Predicted Positives}}$$

$$\text{Recall} = \text{TPR} = \frac{\text{True Positives}}{\text{Total Positives}}$$

F1 Score

The **F1 Score** is a popular metric that combines a classifier's precision and recall by taking their harmonic mean:

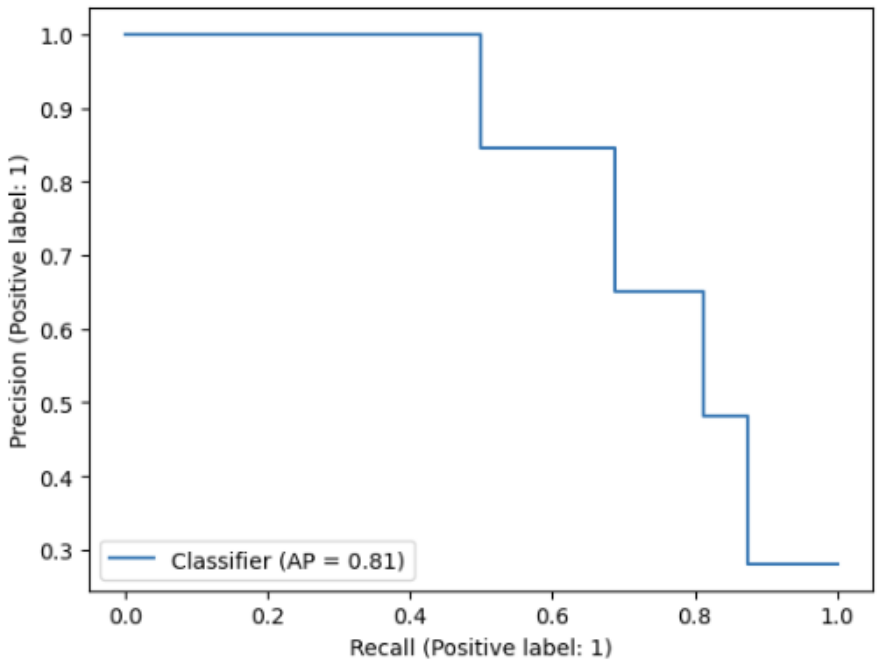
$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Since F1 focuses on the prediction of the positive class, it tends to be popular in applications that seek to identify relatively uncommon events.

- Examples include: predicting loan default, classifying spam emails, etc.

PR-AUC

Precision and recall can also be explored at a variety of classification thresholds in a PR curve:



The area under this curve, known by *PR-AUC* or *AP*, is an alternative single number summary of a classifier's performance.

Multiple Classes

Extensions to multi-label classification require adopting *one-vs-rest scheme* (combining many classes to form the "negative" class) or a *one-vs-one scheme* (pairwise comparisons)

	Pr Setosa	Pr Versicolor	Pr Virginica
Setosa	50	0	0
Versicolor	0	48	4
Virginica	0	2	46

- Under a one-vs-many scheme:
  - For Versicolor flowers, the TPR is 48/52 and the FPR is 2/98
- Under a one-vs-one scheme:
  - For Versicolor flowers compared to Virginica flowers, the TPR is 48/52 but the FPR is 2/48

Micro vs. Macro Averaging with Multiple Classes

For multi-label applications there are two popular for calculating single number metrics like AUC or the F1-score:

1. *Micro-averaging* - aggregate the contributions from each class when calculating the metric (only applicable in the one-vs-many scheme)
2. *Macro-averaging* - calculate the metric independently for each class, then take the average (applicable for both one-vs-many and one-vs-one schemes)

For the Iris example:

1. The *micro-averaged TPR* is $\frac{\sum_{j=1}^{k} \text{True Pos.}}{\sum_{j=1}^{k} \text{Total Pos.}} = \frac{50+48+46}{50+52+48} = 0.9600$
2. The *macro-averaged TPR* is $\frac{1}{k}\sum_{j=1}^{k} \text{TPR}_j = \frac{(50/50)+(48/52)+(46/48)}{3} = 0.9605$

Which Metric(s)?

Consider the following confusion matrix:

	Pred_Pos	Pred_Neg
Has HIV	10	10
Doesn't Have HIV	5	995

- Classification accuracy is 98.5%
- Balanced accuracy is 74.75%
- The F1-score is 0.571

Which metric provides the most useful assessment of this classifier?

Recommendations

- When to use Accuracy?
  - Balanced classes
  - Equal costs for TP and FP
- When to use Balanced Accuracy?
  - Imbalanced classes
  - Equal costs for TP and FP
- When use ROC-AUC?
  - Interested in performance across a range of classification thresholds
- When to use PR-AUC?
  - Rare positive class
  - Interested in accurate detection of the positive class across a range of classification thresholds
- When to use F1?
  - Interested in applying a single threshold
  - Precision and recall both matter