Evaluating Classifier Performance

Ryan Miller

Grinnell College
Statistics

Motivating Example

- Sobar (2016) aimed to predict cases of cervical cancer using non-invasive social and behavioral assessments
  - They recorded 18 different assessment scores for 72 individuals, 21 with cervical cancer and 51 without cancer
- A k-nearest neighbors model (using $k = 8$) evaluated using LOOCV results in 87.5% classification accuracy
  - Is this a worthwhile result?

Confusion Matrices

The table below displays the actual status and the cross-validated prediction of all 72 individuals:

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

Despite being $< 30\%$ of the data, individuals with cancer made up 8 of 9 incorrect classifications.

Confusion Matrices

The table from the previous slide is known as 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 that the *positive class* is determined by the data analyst

Receiver Operating Characteristics (ROC) Analysis

Receiver Operating Characteristics (ROC) reflect the *row proportions* of the confusion matrix:

- **True positive rate** (TPR), or $\frac{\text{True Positives}}{\text{Total Positives}}$, also known as *sensitivity*, *hit rate*, and *recall*
- **False positive rate** (FPR), or $\frac{\text{False Positives}}{\text{Total Negatives}}$, also known as 1−*specificity*

A perfect classifier has a TPR of 1 and a FPR of 0, but there's an inherent trade off between the two quantities.

Grinnell College
Statistics

Receiver Operating Characteristics (ROC) Analysis

- Classification algorithms return a *score*, or a predicted probability of the positive class
  - These scores must be mapped to a class label, with the typical threshold for binary classification being 0.5
  - This threshold can be changed to manipulate the trade off between TPR and FPR

Receiver Operating Characteristics (ROC) Analysis

- Classification algorithms return a *score*, or a predicted probability of the positive class
  - These scores must be mapped to a class label, with the typical threshold for binary classification being 0.5
  - This threshold can be changed to manipulate the trade off between TPR and FPR
- For example, a "positive" classification for any observation with at least 3 of 8 neighbors having cancer leads to the following confusion matrix:

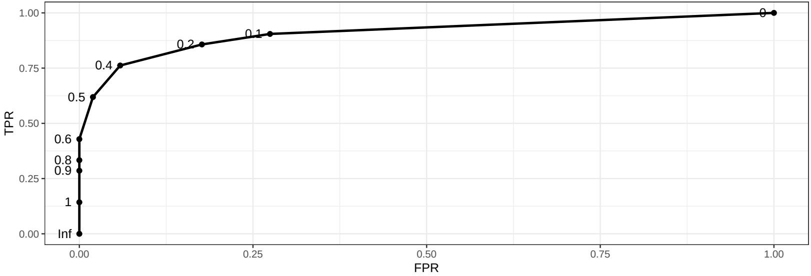
	Predicted Cancer	Predicted Healthy
Has Cervical Cancer	16	5
Doesn't Have Cancer	3	48

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

Grinnell College
Statistics

ROC Curves

The trade-off between TPR and FPR of a classifier can be summarized by an ROC curve:



The *area under the ROC curve* (AUC) provides a single-number summary of how well the classifier performs. An AUC value of 0.5 reflects no predictive value, while 1.0 indicates perfect classification.

Grinnell College
Statistics

Class Imbalances

- A strength of AUC is that it is not influenced by imbalances in the class distribution
  - In our example, 70.8% of the sample did not have cancer, so we could achieve 70.8% classification accuracy *without the model adding any predictive value*

Class Imbalances

- A strength of AUC is that it is not influenced by imbalances in the class distribution
  - In our example, 70.8% of the sample 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:

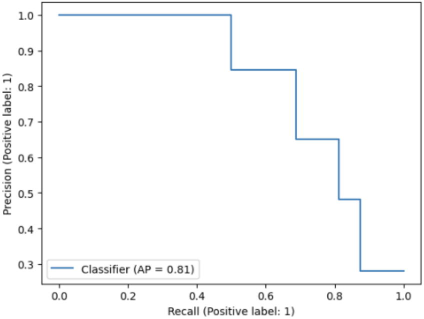
$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

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 requires 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:

Pos	Neg
10	10
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?

Grinnell College
Statistics

Recommendations

- Classification accuracy is acceptable when classes are roughly balanced, and false positives/negatives are equally problematic
  - It is also the most easily interpreted metric, so non-technical clients may prefer it
- Balanced accuracy is useful when classes are imbalanced, and false positives/negatives are equally problematic
- ROC analysis and AUC are useful when classes are imbalanced, and false positives/negatives have differential impact
- The F1-score and PR analysis are useful when classes are imbalanced, and you mostly care about predicting the positive class