Overview

Data

Models

* HOP as preferred for now
* Measuring the accuracy of predictions
  + Sensitivity/specificity, PPV/NPV
  + Predicted probabilities, w/estimation error
* Problems:
  + Year predicts score

Lots of parameters to tweak:

* Kappa weights?
* Category thresholds instead of max?
* Classification “accuracy” vs. prediction + error

Student behaviors may change if they know audition will be used for placement

## Evaluating Prediction Quality

Overall accuracy – the proportion of students whose predicted scores match their actual scores – may be an intuitive statistic for evaluating the quality of predictions, but this statistic is often misleading and uninformative. In Fall 2018, for example, about 40% of students were assigned an ensemble score of 2; a model that simply predicted a score of 2 for *all* students would therefore have an overall accuracy of 40%, despite using no student information at all to produce its predictions.

### Recall and Precision

Instead, it is recommended to examine at least the following two, category-specific statistics when assessing the quality of predictions:

*Recall* (a.k.a. *Sensitivity*): The proportion of students with an *actual* score of *k* whose *predicted* score was also *k*.

*Precision* (a.k.a. *Positive Predicted Value*): The proportion of students with a *predicted* score of *k* whose *actual* score was also *k*.

In a confusion matrix, students whose predicted scores match their actual scores appear in the main diagonal cells (shaded gray in Table 1). In Table 1, the recall for each score is the number of correct predictions in the score’s row divided by the total number of students in the row, while the precision for each score is the number of correct predictions in the score’s column divided by the total number of students in the column. This results in the following values:

Because all students were assigned a predicted score of 2, the recall for category 2 was 1.0; in other words, among students with an actual score of 2, 100% were accurately “predicted” to have a score of 2. However, the recall for the other categories is 0, revealing that the model is not accurate for anyone whose actual score is not 2.

Similarly, the precision for category 2 was 0.50, meaning that 50% of students who were assigned a predicted score of 2 actually obtained a score of 2; however, for the other categories, precision was effectively 0 (technically undefined).

An overall goal of prediction is to maximize both of these values for each category. In practice, however, an increase in precision or recall for some categories often results in a decrease for other categories.

Cohen’s kappa

Cohen’s kappa statistic