CIS 419/519 Recitation

07 October

Evaluation

If we have two models H1 and H2, can we say that H1 is more accurate than H2? How much confidence should we have in such a conclusion?

Typical method: compute the probability that the accuracies H1 and H2 are farther apart than we saw in our testing (under the assumption that H1 and H2 give the same performance); if this probability is low enough, then we can reject the assumption

Paired t-test

Used to compare the performance of two models on several test sets

Uses the accuracies of the two models on each set

Does not require knowledge of whether each model classified a particular example correctly but does require multiple test sets for meaningful results

We can use the paired t-test to compare two models during cross-validation by pairing the accuracy of H1 on fold i with the accuracy of H2 on fold i

McNemar’s Test

Used to compare the performance of two models on a single test set

Requires knowledge of whether each model made a mistake or not on each input example in the test set

Compares models by looking at the test instances where they disagree and seeing which model is right more often

Bootstrap Hypothesis Testing

Estimates quantity on future test sets using a single test set

For example, the quantity could be the difference between the accuracies of two models

Approximate future test sets by resampling the current test set: choose n examples from the current test set with replacement, where n is the size of the current test set

Cross-Validation

Advantages of cross-validation over train/test split?

- Cross validation gives a more accurate measure (lessens randomness)
  - More relevant for smaller datasets
- Allows you to observe performance consistency
  - If all folds don’t give similar results, it is a good sign that something is wrong
- Allows for more accurate parameter tuning
  - Stops you from choosing parameters that overfit on one validation set

Which metric is best?

- We have seen that accuracy isn’t always the best choice

- However which metric is best out of precision/recall/f1?

- Depends on the task at hand

Choosing a metric

Precision: TP/(TP + FP), best when you really want to minimize FP

- Eg: If you are deciding whether or not a patient will survive an intensive treatment

Recall: TP/(TP+FN), best when you want to minimize FN

- Eg: If you are testing someone for a contagious virus

F1 Score: When you care about FN and FP

- Eg: Trying to predict whether or not someone will commit a crime

Accuracy isn’t always bad, it has its uses!

Which metric is best?

Precision vs Recall vs F1 vs Accuracy:

Trying to predict if a patient is at risk of cancer?

A: Recall (You don’t want any FNs, i.e. telling people they aren’t at risk when they are)

Which metric is best?

Precision vs Recall vs F1 vs Accuracy:

Predicting the outcome of a chess match?

A: Accuracy (no class imbalance, and not necessarily worried about minimizing FNs or FPs)

Which metric is best?

Precision vs Recall vs F1 vs Accuracy:

Whether or not to recommend a youtube video?

A: Precision (You don’t want any FPs, to recommend videos that people don’t like)

Which metric is best?

Precision vs Recall vs F1 vs Accuracy:

Deciding who to promote your product to? (assume that promoting comes at a significant cost)

A: F1 (you don’t want FPs, people who you promote to that won’t buy, as that will incur significant promoting cost. You also don’t want FNs, people who will buy your product that you aren’t promoting to)

Which metric is best?

Precision vs Recall vs F1 vs Accuracy:

Predicting if the weather is right to launch a rocket?

A: Precision (you don’t want FPs, as if you think a day will have good weather and it doesn’t, you incur significant rescheduling costs and possible equipment damage from rain)

Which metric is best?

Precision vs Recall vs F1 vs Accuracy:

Predicting if someone poses a threat to an event and should be searched?

A: Recall (You don’t want any FNs, i.e. not searching people who do pose a threat)

Which metric is best?

Precision vs Recall vs F1 vs Accuracy:

Is this email a spam mail?

A: Precision (You don’t want any FPs, i.e. sending important emails to the junk section)