Cross-Validation

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- In this "lecture", we will talk about cross-validation.
- First, we need to motivate the problem.
- A good motivation point is Assignment 5...
- In particular, the meaning of training/test as opposed to labelled/unlabelled.

Lesson From Assignment 5

- You were asked to do labelled/unlabelled splits of the data.
- Some of you took the labelled points to be the training set and the unlabelled points to be the test set.
- This is not ridiculous; however, observations in the test set must be labelled — in other words, points treated as truly unlabelled cannot be used for model selection.
- In some cases, e.g., when we looked at kNN, we considered a sort of training/validation/test framework.
- In general, however, we have used a training and test set in our examples.

Training/Test

- In several of the examples we have looked at, we used a training set to build models and a test set to select the best model.
- One downside to this approach is that the performance on the test set cannot be taken as reflective of the performance of the model on unlabelled or "new" points.
- Let's think back to our neural network examples from last week.

Example

- Last week, we did some examples in R, where a "vanilla" neural network was used for classification.
- In essence, we used the training set to build lots of networks and then used the test set to select the best model.
- While we did choose the "best" model (in terms of misclassification rate), can we say how well it would do on "new" data?

Example contd.

- In this context, what does it mean if the chosen model has a 3% misclassification rate on the test set?
- Does that mean that we might reasonably expect a 3% misclassification rate on new data?
- It does not; it will generally be a (perhaps substantial) underestimate.
- This is why some people like to use a validation set in addition to the training and test sets.

Training/Validation/Test

- This approach can be used in a so-called "data rich" situation.
- Here is one view of the training/validation/test approach:
  - The **training set** is used to build (lots of) models.
  - The **validation set** is used to select the model.
  - The **test set** is used to compute the prediction error, e.g. to estimate the misclassification rate.
- The split might be 50%/30%/20% but there is no hard and fast rule.
- In practice, we may not have enough data to "afford" a separate validation set — we may not be "data rich".

K-Fold Cross-Validation

```
split the training data into K (roughly) equally sized parts

for k in 1:K
     x = training data with kth part removed
     build model using x
     use the kth part to compute the prediction error
end for

return a combination of the K prediction errors
```

- On the kth iteration of the `for` loop, the kth part plays the role of a validation set.
- A test set can still be used.

K-Fold Cross-Validation contd.

- Common choices are $K = 5$ and $K = 10$.
- Taking $K = n$ gives "leave one out" cross-validation.
- The choice of K can be thought of as a variance-bias trade-off.
- Hastie et al. (2009) gives very good coverage of this material, and it is also discussed by James et al. (2013) — see course website for bibliographic details.

Comments

- Where does the OOB error in random forests fit in here?
- Note that notation can vary greatly across different sources — I recommend focusing on what is actually being done as opposed to semantics.
- As usual, further reading is strongly encouraged; James et al. (2013) and Hastie et al. (2009) are good starting points.
- Now, let's look at some examples in R.