5: Overtraining and Cross-validation
Machine Learning and Real-world Data

Simone Teufel and Ann Copestake

Computer Laboratory
University of Cambridge

Lent 2017

- You looked at various possible system improvements, e.g. concerning the Laplace smoothing parameter.
- You can now decide whether a manipulation leads to a statistically significant difference.
- Let us now think about what our NB classifier has learned.
    - We hope is has learned that "excellent" is an indicator for Positive
    - We hope it hasn't learned that certain people are bad actors.

Ability to Generalise

- We want a classifier that performs well on new, never-before seen data.
- That is equivalent to saying we want our classifier to generalise well.
- In detail, we want it to
  - recognise only those characteristics of the data that are general enough to also apply to some unseen data
  - while ignoring the characteristics of the training data that are overly specific to the training data
- Because of this, we never test on trained, but use separate test data.
- But overtraining can still happen even if we use separate test data.

Overtraining with repeated use of test data

- You could make repeated improvements to your classifier, choose the one that performs best on the training data, and declare that as your final result.
- Overtraining is when you think you are making improvements (because your performance on the test data goes up) . . .
- . . . but in reality you are making your classifier worse because it generalises less well to data other than your test data.
- It has now indirectly also picked up accidental properties of the (small) test data.

Overtraining, the hidden danger

- Until deployed to real unseen data, there is a danger that overtraining will go unnoticed.
- One of the biggest dangers in ML
  - because you have to be vigilant to notice that it's happening
  - because performance "increases" are always tempting (even if you know they might be unjustified).
- Other names for this phenomenon:
  - Overfitting
  - Type III errors
  - "Testing hypotheses suggested by the data" errors

Am I overtraining?

- You are absolutely sure from overtraining if you have large amounts of test data, and use new (and large enough) test data each time you make an improvement.
- You can't be sure if you are overtraining if you make incremental improvements to your classifier and repeatedly optimise the system based on its performance on the same small test data.
- You can inspect the most characteristic features for each class (cf. starred tick) and get suspicious when you find features that are unlikely to generalise
  - "theater"

The "Wayne Rooney" effect

- One way to notice overtraining is by time effects.
  - Time changes public opinion on particular people or effects.
  - Vampire movies go out of fashion, superhero movies come into fashion.
  - People who were hailed as superstars in 2003 might later get bad press in 2010
  - Called the "Wayne Rooney" effect
- You will test how well your system (trained on data from up to 2004) performs on reviews from 2015/6
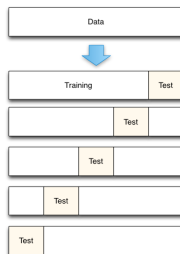
Cross-validation: motivation

- We can't afford getting new test data each time.
- We must never test on trained.
- We also want to use as much training material as possible (because ML systems trained on more data are almost always better).
- We can achieve this by using every little bit of training data for testing – under the right kind of conditions.
- By cleverly iterating the test and training split around

N-Fold Cross-validation

- Split data randomly into *N* folds
- For each fold X – use all other folds for training, test on fold X only
- The final performance is the average of the performances for each fold

- Use your significance test as before, on *all* of the test folds → you have now gained more usable test data and are more likely to pass the test if there is a difference.
- Stratified cross-validation: a special case of cross-validation where each split is done in such a way that it mirrors the distribution of classes observed in the overall data.

- If all splits perform equally well, this is a good sign
- We can calculate variance:

$$var = \frac{1}{n}\sum_{i}^{n}(x_i - \mu)^2$$

  - x_i: the score of the ith fold
  - $\mu : avg_i(x_i)$: the average of the scores

Data splits in our experiment

- Training 80%
- Validation set 10% – used up to now for testing
- Test 10% – new today!
- Use training + validation corpus for cross-validation

First task today

- Use the precious test data for the first time (on the best system you currently have)
- Download the 2015/6 review data and run that system on it too.
- Compare results with the accuracies you are used to from testing on the validation set

- Implement two different cross-validation schemes:
    - Random
    - Random Stratified
- Observe results. Calculate variance between splits. Perform significance tests wherever applicable.

Tick 3 comprises:

- Task 3 – Statistical Laws of Language
- Task 4 – Significance Testing

Literature

- James, Witten, Hastie and Tibshirani (2013). *An introduction to statistical learning*, Springer Texts in Statistics. Section 5.1.3 p. 181-183 (k-fold Cross-Validation)