COMP30027 Machine Learning
Evaluation II

Semester 1, 2018

Jeremy Nicholson & Tim Baldwin & Karin Verspoor

THE UNIVERSITY OF MELBOURNE

POSTERA CRESCAM LAUDE

Lecture Outline

1. Recap
2. Overfitting
3. Model Bias and Variance
4. Evaluation Bias and variance
5. Summary

Evaluation in Supervised ML

- We start with a dataset of instances comprised of attributes and labels
- We use a **learner** and the dataset to build a **classifier**
- We attempt to assess the *effectiveness* of the classifier
    - Generally, by comparing its predictions with the actual labels on some (different) instances

Exploring the Inductive Learning Hypothesis

Inductive Learning Hypothesis: Any hypothesis found to approximate the target function well over (a sufficiently large) training data set will also approximate the target function well over held-out *test examples*.

- Why do we need to test our hypothesis on *held-out test examples*?
- What do we mean by "sufficiently large" training set?
- What impact does the size of the test set have?

Tensions in Classification

- **Overfitting:** has the classifier tuned itself to the idiosyncracies of the training data rather than learning its generalisable properties?
- **Consistency:** is the classifier able to flawlessly predict the class of all training instances?
- **Generalisation**: how well does the classifier generalise from the specifics of the training examples to predict the target function?

Our evaluations must take these ideas into consideration.

Lecture Outline

Learning curves I

- Holdout (and cross-validation, to a lesser extent), is based on dividing the data into two (three?) parts:
  - Training set, which we use to build a model
  - Evaluation set ("dev data", "test data"), which we use to assess the effectiveness of that model
- More training instances $\rightarrow$ (usually) better model
- More evaluation instances $\rightarrow$ more reliable estimate of effectiveness

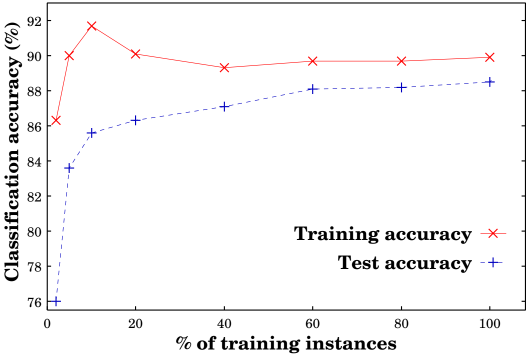
Learning curves II

Learning curve:

- Choose various split *sizes*, and calculate effectiveness
  - For example: 90-10, 80-20, 70-30, 46-40, 50-50, 40-60, 30-70, 20-80, 10-90 (9 points)
  - Might need to average multiple runs per split size
- Plot % of training data vs training/test Accuracy (or other metric)
- This allows us to visualise the data trade-off

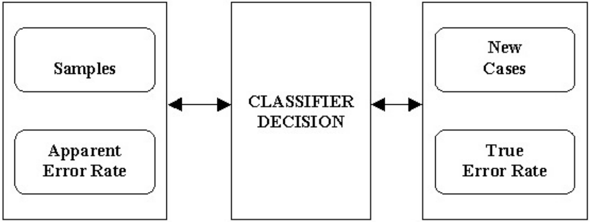
Learning curves III

Learning curve:



What is the Accuracy?

Estimating true performance



- We extrapolate performance from a finite sample of cases.
- Training error is one starting point in estimating the performance of a classifier on new cases.
- With unlimited samples used for learning, apparent error rate will become the true error rate eventually.

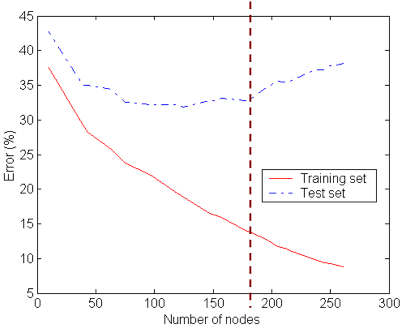
From http://dms.irb.hr/tutorial/tut_mod_eval_3.php

Generalisation

- A good model should fit the training data well, **and** generalise well to unseen data.
- The expectation is that training and test data are randomly selected from the same population, but neither are the entire population.
- True error rate is almost always much higher than training error, due to overfitting to the training data.
- A model that fits the training data *too* well can have poorer generalisation than a model with higher training error.

Overfitting I

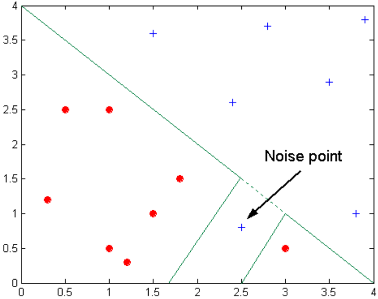
- Possible evidence of overfitting: large gap between training and test performance



From Tan et al. (2006)

Overfitting II

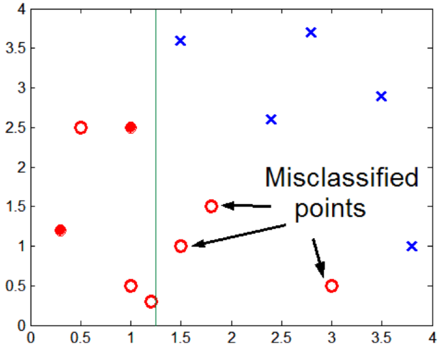
- Decision “boundary” distorted by noise.
- Adding new instance of (sunny, hot, normal, strong $\rightarrow$ no).



A simpler decision boundary would generalise better for this data.

Overfitting III

- Lack of coverage of population sample could lead to poor model.
    - could be due to small numbers of examples
    - could be due to non-randomness in training sample ("sampling bias")



Lecture Outline

Statistical definition of bias and variance

(Statistical) bias:

$$\mathrm{Bias}(\hat{\theta}; \theta) = \mathrm{E}_x[\hat{\theta}(x) - \theta(x)]$$

(Statistical) variance:

$$\mathrm{Var}(\hat{\theta}; \theta) = \mathrm{E}_x[\hat{\theta}(x)^2] - \mathrm{E}_x[\hat{\theta}(x)]^2$$

... How does this relate to ML?

Bias and Variance in ML

In the ML world, **bias** is used to refer to a number of things:

- "model bias" — the propensity of our classifier to make systematically wrong predictions
- "evaluation bias" — the propensity of our evaluation strategy to over- or under-estimate the effectiveness of our classifier
- "sampling bias" — if our training or evaluation dataset isn't representative of the population (effectively, breaking the Inductive Learning Hypothesis)
- occasionally "prejudicial", like in casual speech (more later)

Mercifully, **variance** only refers to "model variance" and "evaluation variance" (and these are difficult to distinguish).

"Model" bias and variance I

In an ML context, model bias is easiest to understand relative to **regression**:

- The regressor is an "estimator": for every evaluation instance, the (signed) error can be calculated
- Assuming every instance is independent, bias is the average of these (signed) errors

So, we can infer:

- A model is biased if the predictions are systematically higher than the true value, or systematically lower than the true value
- A model is unbiased if (i) the predictions are systematically correct, or (ii) some of the predictions are too high, and some of the predictions are too low

"Model" bias and variance II

Model variance, relative to **regression**, can follow logically:

- The expected value function is the mean in this context
- So, we can compare the average of the squared predictions with the square of the average of the predictions
- It isn't immediately clear how to interpret this, however

Model bias and variance in classification I

What if instead of a regression problem, we have a classification problem? (i.e. no obvious definition of expected value function)

- Model bias relates to Accuracy, relative to different training sets (sampled from the same population)
- Model variance relates to the propensity of different training sets to produce different models/predictions (with the same learner)
    - A model has high variance if a different randomly sampled training set leads to very different predictions on the evaluation set
    - A model has low variance if a different randomly sampled training set leads to similar predictions *independent of whether the predictions are correct*

Model bias and variance in classification II

One typical (conversational) definition of model bias in a classification context:

- Label predictions can't be "too high" or "too low"
- Rather, we typically compare the **class distribution**:
  - An unbiased classifier produces labels with the same distribution as the actual class distribution
  - An biased classifier produces labels with a different distribution
- A biased classifier is guaranteed to be making errors (why?); an unbiased classifier might be making errors, or might not
- "*...biased towards the majority class...*": our model predicts too many instances as the majority class

Model bias and variance in classification III

These are *informal* definitions, and can't be measured quantitatively:

- Bias is generally binary: a classifier *is* biased, or it is isn't
  - Polynomial/RBF kernal SVM tends to have low bias
- (or sometimes relative: one classifier is *more* biased than a second classifier)
- Variance is generally relative: one classifier has *more* variance than another classifier
  - Naive Bayes tends to have lower variance than other classifiers

Model bias and variance in classification IV

Remember:

- High bias and high variance are often “bad”, but low bias and low variance are no guarantee of “good”!
    - The weighted random classifier is low bias
    - 0-R is low variance (zero variance)
- Lower bias and lower variance is no guarantee of “better”!
    - But generally desirable, all else equal

Lecture Outline

1. Recap
2. Overfitting
3. Model Bias and Variance
4. Evaluation Bias and variance
5. Summary

Evaluation, again I

Perhaps obvious, but worth re-stating:

- In supervised ML, the way we evaluate a model is typically independent of the way we build the model
  - Some counter-examples, like RSS in Linear Regression
- This is made explicit in Holdout/Cross-Validation, compared to testing on the training data

Evaluation, again II

Bias (of an estimator), again:

$$\mathrm{Bias}(\hat{\theta}; \theta) = \mathrm{E}_x[\hat{\theta}(x) - \theta(x)]$$

Our evaluation metric is also an estimator...

Evaluation, again III

- We want to know the "true" error rate of a classifier, but we only have an estimate of the error rate, subject to some particular set of evaluation instances
- (It's confusing, but remember: this is *independent* of the trained model itself)
- Why do we wish to know the "true" error rate? Generalisation.
- What's the risk with our estimated error rate? Overfitting.
  - i.e. We have good Accuracy with respect to some specific evaluation set, but poor Accuracy with respect to other unseen evaluation sets
  - It's also possible to overfit the *development* data, with respect to our evaluation function!

Bias and variance in Evaluation I

Evaluation bias:

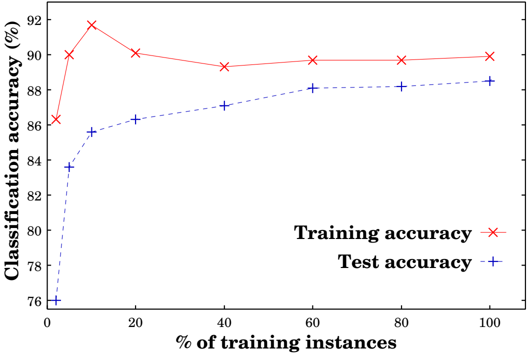
- Similar logic to model bias
- Our estimate of the effectiveness of a model is systematically too high/low

Evaluation variance:

- Our estimate of the effectiveness of a model changes a lot, as we alter the instances in the evaluation set
- Again, this can be hard to distinguish from model variance

Bias and variance in Evaluation II

Learning curve, again:



What is the "true" Accuracy?

Bias and variance in Evaluation III

How do we control bias and variance in evaluation?

- Holdout partition size
  - More training data: less model variance, more evaluation variance
  - Less training (more test) data: more model variance, less evaluation variance
- Repeated random subsampling and *M*-fold Cross-Validation
  - Less variance than Holdout
- Stratification: less model bias
- Leave-one-out Cross-Validation
  - No possibility of sampling bias, lowest bias/variance in general

Lecture Outline

1. Recap
2. Overfitting
3. Model Bias and Variance
4. Evaluation Bias and variance
5. Summary

Summary

- What is generalisation and overfitting?
- What is a learning curve, and why is it useful?
- How are bias and variance different?
- How is model bias different to evaluation bias?
- How do we try to control for bias and variance in evaluation?

References I

Daniel Jurafsky and James Martin. *Speech and Language Processing*. Prentice Hall, 2nd edition, 2008.

Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.