Lecture 13: Evaluation Part 2

COMP90049

Semester 1, 2021

Lea Frermann, CIS

Acknowledgement: Jeremy Nicholson, Tim Baldwin & Karin Verspoor

Roadmap

So far ...

- Intuition, maths, and application of different classification models of varying complexity
- Feature selection
- Evaluation: How well are we doing?

Today... Evaluation part II

- How do we know whether the model performs ‘good enough’? When to stop/continue model training, parameter tuning or model selection?
- Types of poor model performance
- Diagnosing poor model performance

Evaluation

Evaluation I

Given a dataset of instances comprising of attributes and labels:

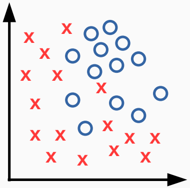
- We use a learner and the dataset to build a classifier
- We assess the effectiveness of the classifier
  - Generally, by comparing its predictions with the actual labels on some unseen instances
  - Metrics: accuracy, precision, recall, Error rate, F-score, etc.

Tensions in Classification

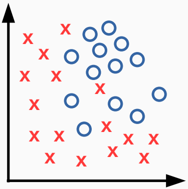
- **Generalisation**: how well does the classifier generalise from the specifics of the training examples to predict the target function?
- **Overfitting:** has the classifier tuned itself to the idiosyncrasies of the training data rather than learning its generalisable properties?
- **Consistency:**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

Generalisation Problem in Classification

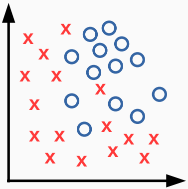
- **Under-fitting:** model not expressive enough to capture patterns in the data.
- **Over-fitting:** model too complicated; capture noise in the data.
- **Appropriate-fitting** model captures essential patterns in the data.



Under-fitting



Appropriate-fitting



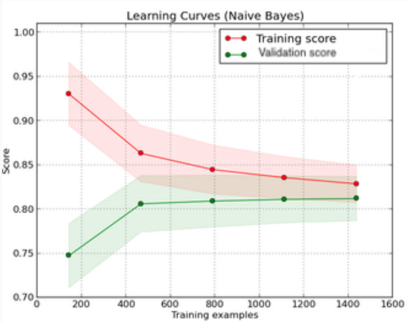
Over-fitting

Learning Curve

Learning Curve I

Learning curve is a plot of learning performance over experience or time

- y-axis: performance measured by an evaluation metric (F-score, precision, ...)
- x-axis: different conditions, e.g. sizes of training dataset, model complexity, number of iterations etc.



Plot on the left

- Learner: Naive Bayes
- What can we say about the difficulty of the problem?

Learning Curve II

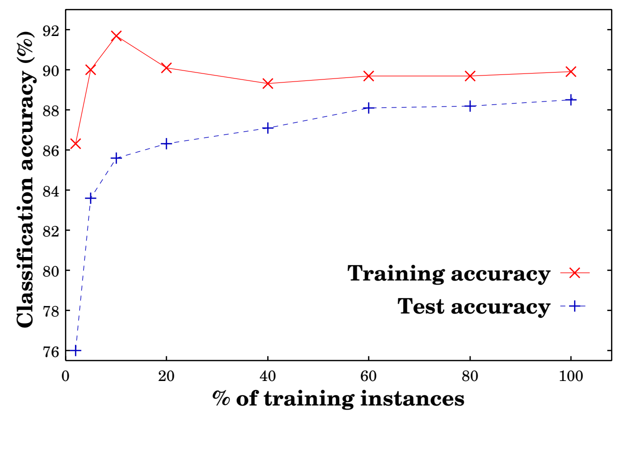
- Holdout (and cross-validation, to a lesser extent), is based on dividing the data into two (three?) parts:
  - Training set, which we use to build a model
  - Evaluation set (“validation data”, “test data”), which we use to assess the effectiveness of that model
- More training instances $\rightarrow$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

Learning Curve III

Learning curve:

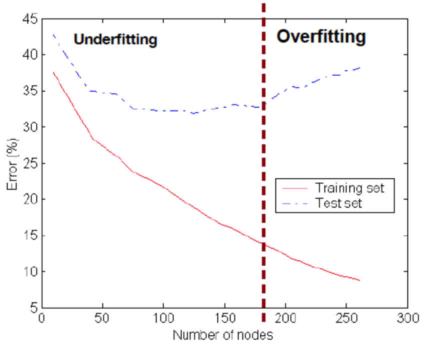
- Choose various split sizes, and calculate effectiveness
  - For example: 90-10, 80-20, 70-30, 46-40, 50-50, 40-60, 30-70, 20-80, 10-90 (9 points)
  - Might need to average multiple runs per split size
- Plot % of training data vs training/test Accuracy (or other metric)
- This allows us to visualise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

Learning Curve IV



Overfitting and Underfitting

- **Underfitting**: when model is too simple $\rightarrow$ both training and test errors are large
- **Overfitting**: when model is too complex $\rightarrow$ training error is small and test error is large

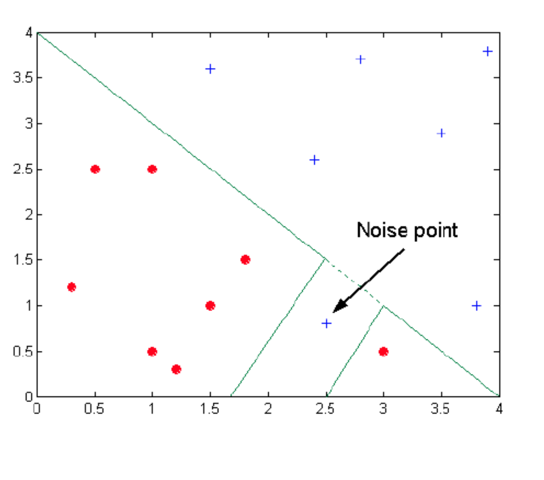


- What would be a good model complexity?

Overfitting

Overfitting due to noise:

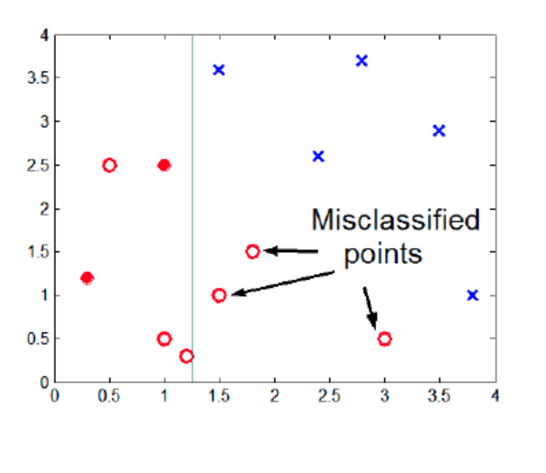
- The decision boundary is distorted by noise



Overfitting

Overfitting due to insufficient training instances

- The data points do not fully represent the patterns in the dataset

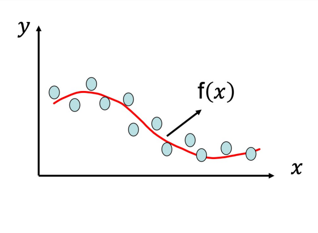


Generalization

Generalization

- A good model generalizes well to unseen data!
- How do we measure the generalizability of a model ?
- Given a training dataset $D = \{x_i, y_i\}$, $i = 1 \dots n$ and $y \in \mathbb{R}$:
    - Assume the data points are generated with a function $f(.)$ plus a noise $\epsilon \in \mathcal{N}(0, \sigma)$. This noise comes from an unknown and unmeasurable source, e.g., annotation error, measure error:

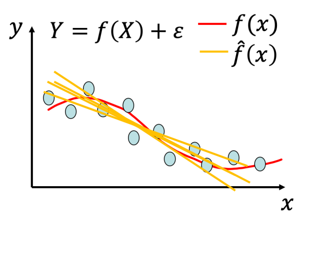
$$Y = f(X) + \epsilon$$



Generalization Error I

- We may estimate a model $\hat{f}(X)$ of $f(X)$ using linear regression or another modelling technique
- **But** different training sets $\rightarrow$ different model weights and outputs
- To remove the dependency $\rightarrow$ repeat modelling many times (on different training sets)
- In this case, the expected squared prediction error at a point x is:

$$Err(x) = E\left[(Y - \hat{f}(x))^2\right]$$



Generalization Error II

- The generalization error can be decomposed to:

$$Err(x) = \left(E[\hat{f}(x)] - f(x)\right)^2 + E\left[\left(\hat{f}(x) - E[\hat{f}(x)]\right)^2\right] + \sigma^2$$

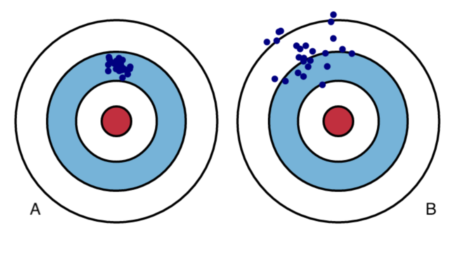
- Or simply written as:

$$Err(x) = \mathrm{Bias}^2 + \mathrm{Variance} + \mathrm{Irreducible\ Error}$$

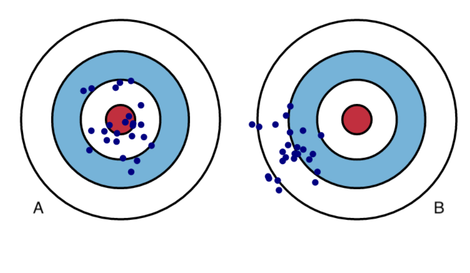
- **Variance**: Captures how much your model changes if you train on a different training set. How ”over-specialized” is your classifier to a particular training set?
- **Bias**: What is the inherent error that you obtain from your model even with infinite training data? This is due to your model being ”biased” to a particular kind of solution. In other words, bias is inherent to your model.
- **Noise**: This error measures ambiguity due to your data distribution and feature representation. You can never beat this, it is an aspect of the data.

Generalization Error III

- Which one has lower variance:

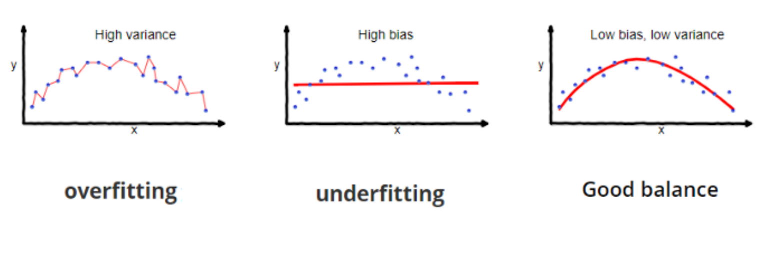


- Which one has lower bias:



Generalization Error VI

- Causes of Poor Generalization:
  - Underfitting: Variance is zero and bias is large
  - Overfitting: bias is zero and variance is substantial
- A Good model
  - Lower bias and lower variance $\rightarrow$ better generalisation



Quiz (via Zoom)

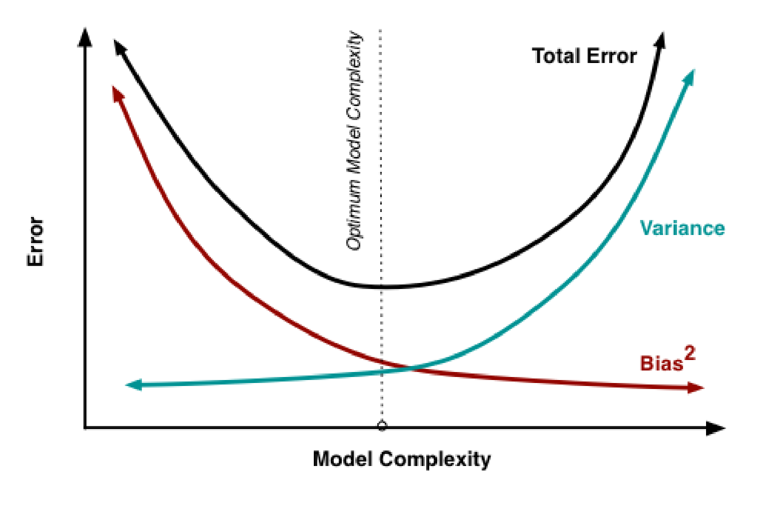
which baseline has lower variance?

1. weighted random classifier
2. 0-R (majority voting)

Diagnosing High Bias and Variance

Bias-Variance Tradeoff

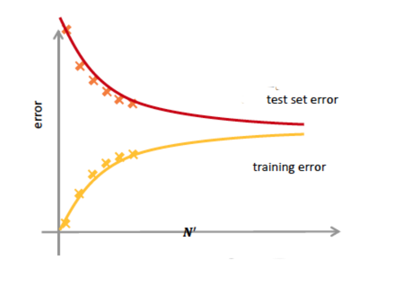
At its root, dealing with bias and variance is really about dealing with overfitting and underfitting. Bias is reduced and variance is increased in relation to model complexity.

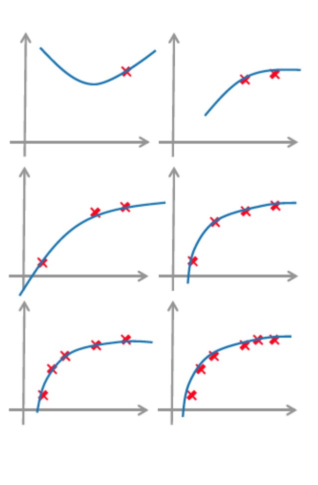


Diagnose Overfitting and Underfitting I

- Plot Training and Test Error as function of data size
- The following situations may occur:

- Fitting a quadratic regression function to data: $h(x:\theta) = \theta_0 + \theta_1 x + \theta_2 x^2$
- Plot training and test errors vs. training set size $N' = 1, 2, 3 \dots n$

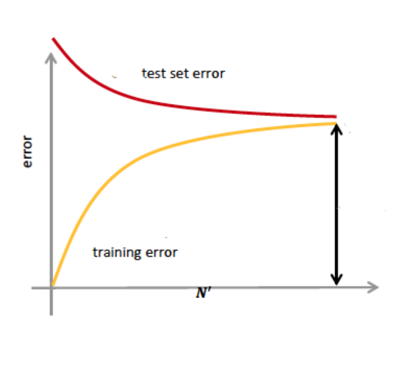


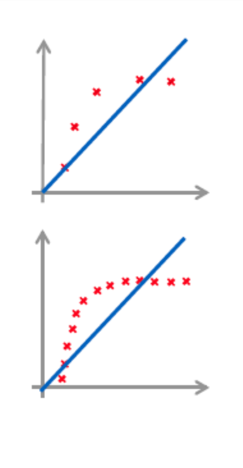


Diagnose Overfitting and Underfitting III

High Bias

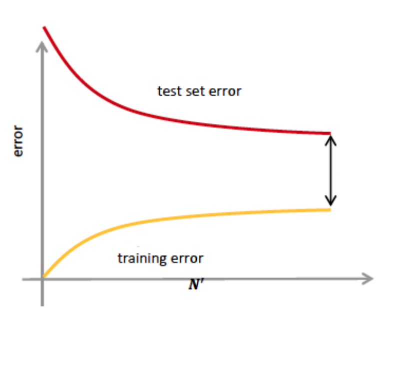
- Getting more training data will not (by itself) help much
- Learning curve is characterized by high training and test errors

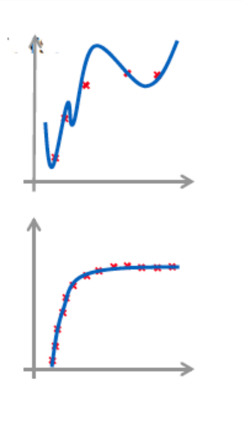




High Variance

- Getting more training data is likely to help
- Learning curve is characterized by gap between the two errors





THE UNIVERSITY OF MELBOURNE

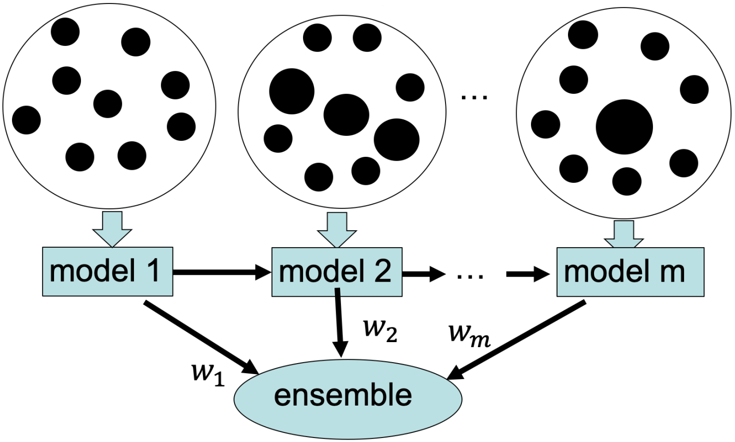
Remedy for High Bias and Variance

High Bias Remedy

- Use more complex model (e.g. use nonlinear models)
- Add features
- Boosting

Boosting

- training data: different weights (probabilities to be selected)
- Use multiple weak models → a stronger model; reduces bias (improves performance)

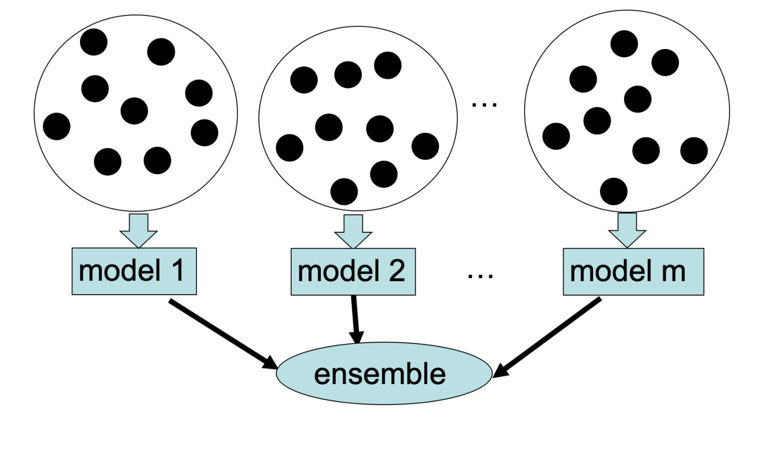


High Variance Remedy

- Add more training data
- Reduce features
- Reduce model complexity – complex models are prone to high variance
- Bagging

Bagging

- Construct new datasets: randomly select the training data with replacement
- Combining multiple models$\rightarrow$ predictions are more stable; reduces variance of individual model.



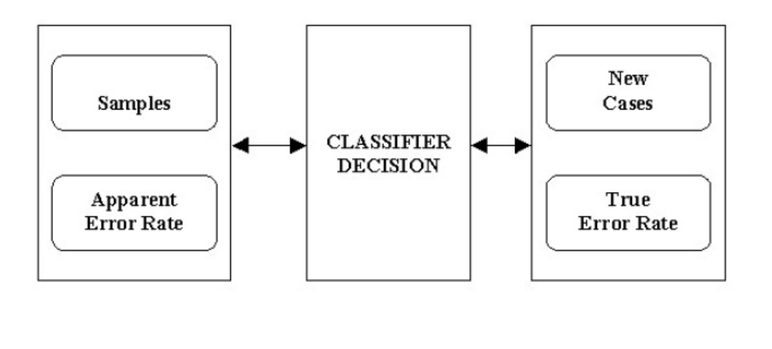
Evaluation Bias and Variance

Evaluation Bias and Variance I

- Our evaluation metric is also an estimator
- Desire to know the ‘true’ error rate of a classifier, but only have an estimate of the error rate, subject to some particular set of evaluation instances
- The quality of the estimation is independent of the trained model

Evaluation Bias and Variance II

- 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, apparent error rate will become the true error rate eventually.



Evaluation Bias and Variance III

- What are the potential problems with our estimated error rate?
  - 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 validation data, with respect to our evaluation function

Evaluation Bias and Variance VI

- 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
  - **Evaluation 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 (very hard to distinguish from model variance)

Evaluation Bias and Variance V

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 K-fold Cross-Validation
  - Less variance than Holdout for model and evaluation
- Stratification
  - less model and evaluation bias
- Leave-one-out Cross-Validation
  - No sampling bias, lowest bias/variance in general

THE UNIVERSITY OF MELBOURNE

Summary

Summary

Today... Evaluation part II

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

Next week

- Guest lecture on academic writing (Assignment 3!)
- Decision Trees & Ensembling

References

- Sammut, Claude; Webb, Geoffrey I., eds. (2011). Bias Variance Decomposition. Encyclopedia of Machine Learning. Springer. pp. 100–101.
- Luxburg, Ulrike V.; Schölkopf, B. (2011). Statistical learning theory: Models, concepts, and results. Handbook of the History of Logic. 10: Section 2.4.
- Vijayakumar, Sethu (2007). The Bias–Variance Tradeoff. University of Edinburgh. Retrieved 19 August 2014.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer. Chapter 2.
- Jeremy Nicholson & Tim Baldwin & Karin Verspoor: Machine Learning

THE UNIVERSITY OF MELBOURNE