University of
South Australia

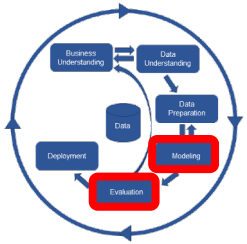
COMP 2019

Week 7
ML Validation

Learning Objectives

- Explain how datasets are used during training (CO3)
- Distinguish bias and variance issues (CO3)
- Explain cross-validation (CO3)
- Understand the large data rationale (CO3)

Supervised Learning from Data

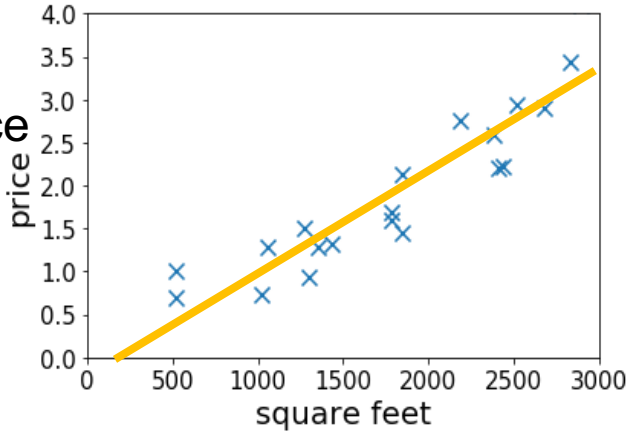


x1	xn	y

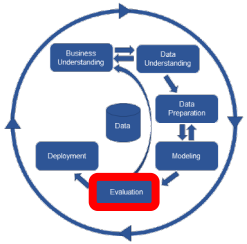
Training

Model
y = f(X)

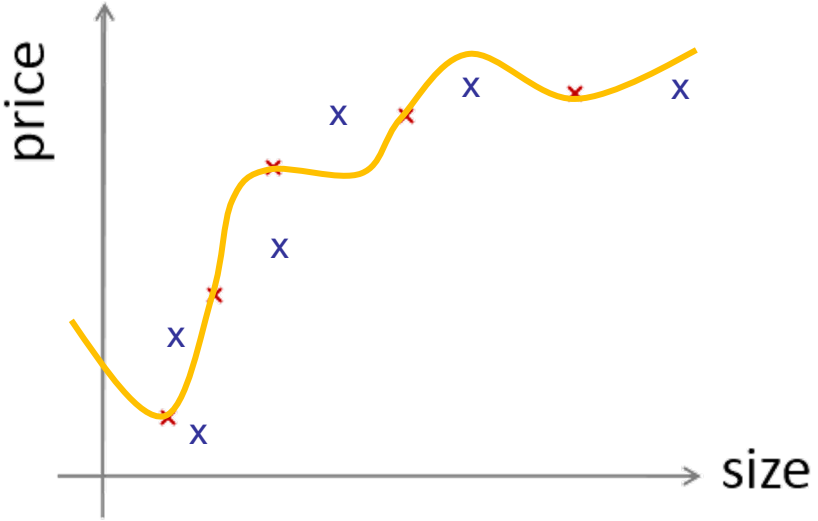
Inference

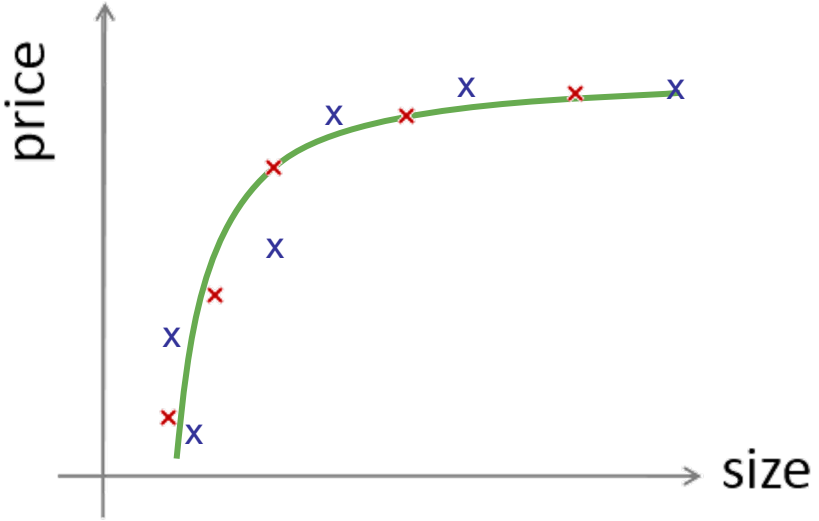


Learning = Generalisation



- How well does the model perform on **UNSEEN** data?





Diagnosing a Learning Algorithm

- Suppose you have implemented learner to predict housing prices based on features $x_1,\ldots,x_n$.
- When you test your learner on a new set of houses, you find that it makes unacceptably large errors in its predictions.

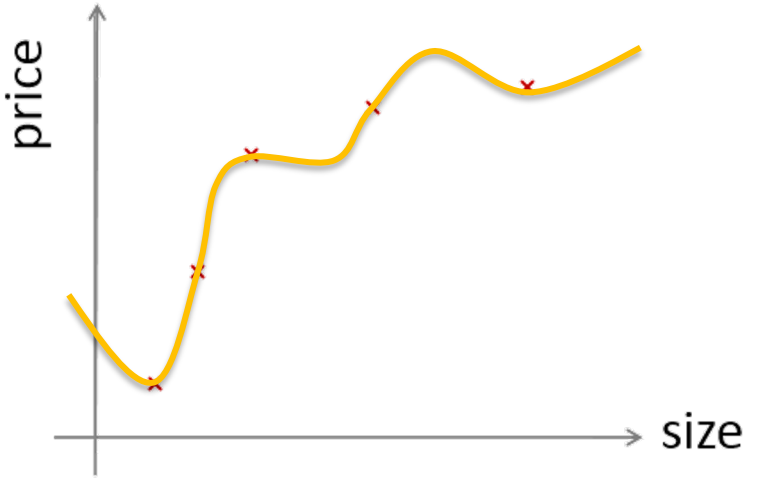
- What should you do?

Possible actions?

- Get more training samples
- Try (adding | removing) features
- Try adding derived features ($\log xi$, x_i^2, $x_i x_j$,…)
- (Impose | Decrease) a penalty on large parameter values
- Hire someone else to do the job.

Overfitting

- Parameters $\theta_0, \ldots, \theta_n$ were fit to the training data to minimize the error as measured **on the training data**.
- This may result in a hypothesis that is tailored too much to the training data.
- The hypothesis fails to generalize from training data.
- The hypothesis “overfits”.



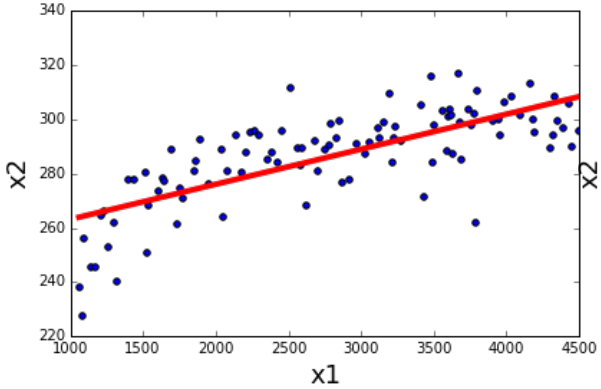
Model Complexity

- Model fit training data well
  - requires a more complex model (with more parameters)
- Behaviour of model on test data should match that on training data
  - requires a less complex (more stable) model

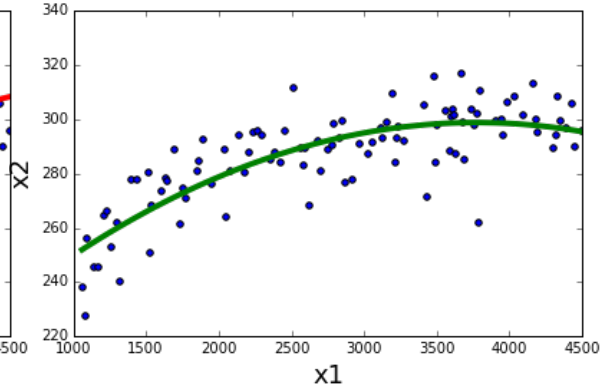
- More complex model
  - smaller training error but larger difference between test and training error
- Less complex model
  - larger training error but smaller difference between test and training error

Occam's Razor

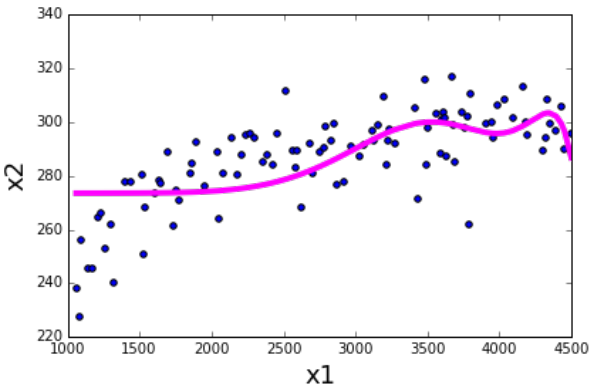
- Among all suitable hypotheses, select the simplest (the one with fewest assumptions).



High bias
("underfit")



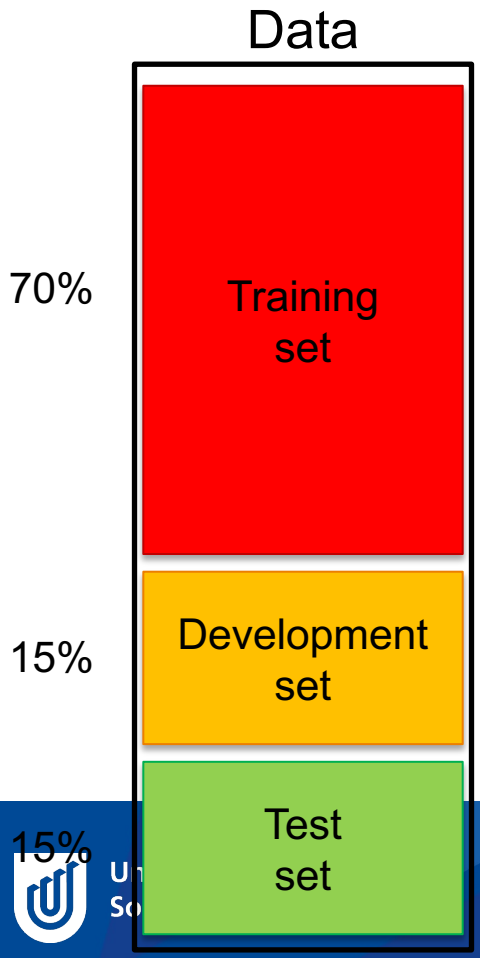
Just right



High variance
("overfit")

Model Selection

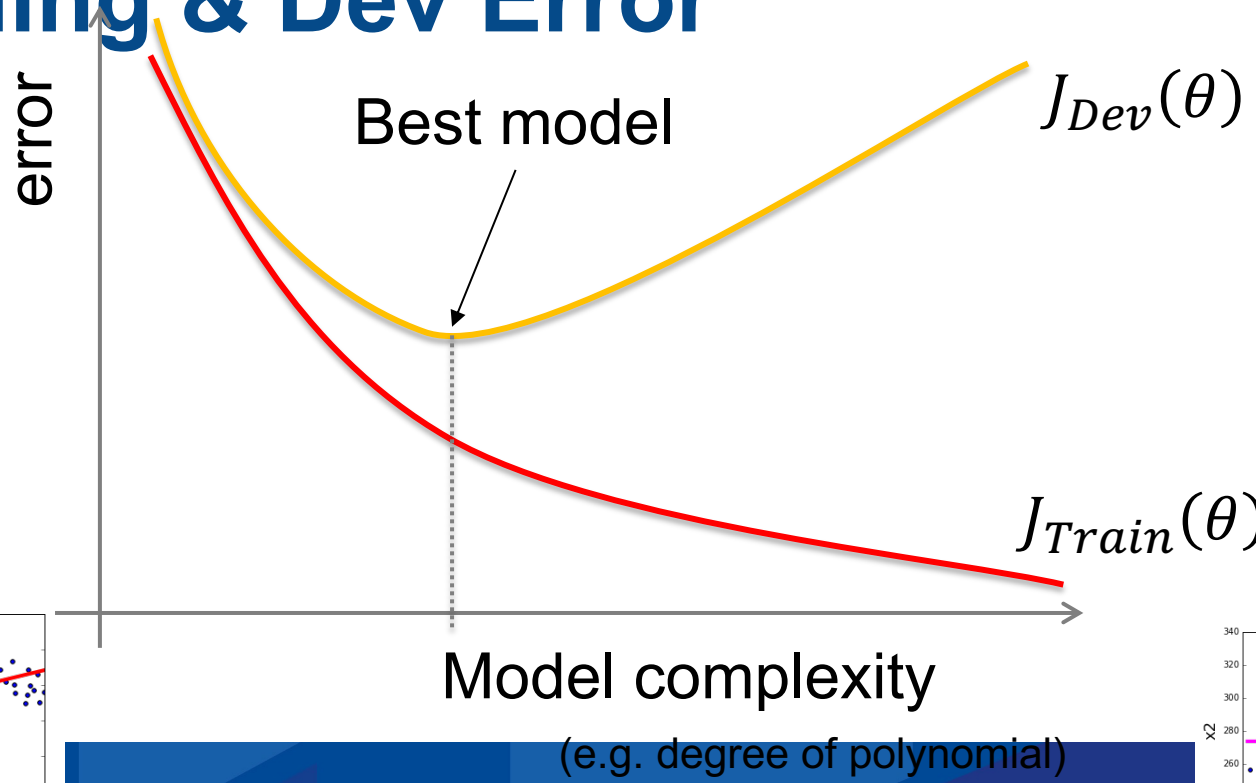
- Q: How do we find a good model?
  - We don’t know which model (among a set of alternatives) we should select
  - Let’s learn multiple models and take the one that performs best
- How do we know which one performs best?
  - We **cannot** use the training and test sets to perform this evaluation

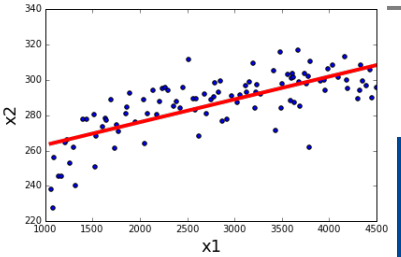


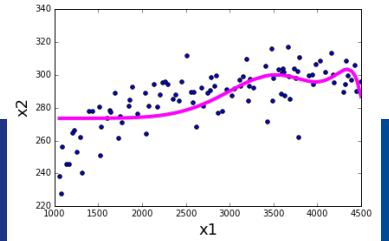
do
1. learn a model m from Training Set
2. evaluate m on Dev set and adjust learning parameters

Repeat until Dev error increases
Evaluate m using Test Set

Training & Dev Error







Underfitting: High Bias

error

Best model

$J_{Dev}(\theta)$

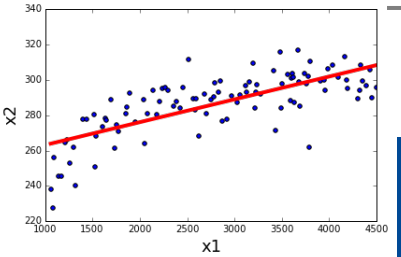
$J_{Train}(\theta)$ is high

$J_{Dev}(\theta) \approx J_{Train}(\theta)$

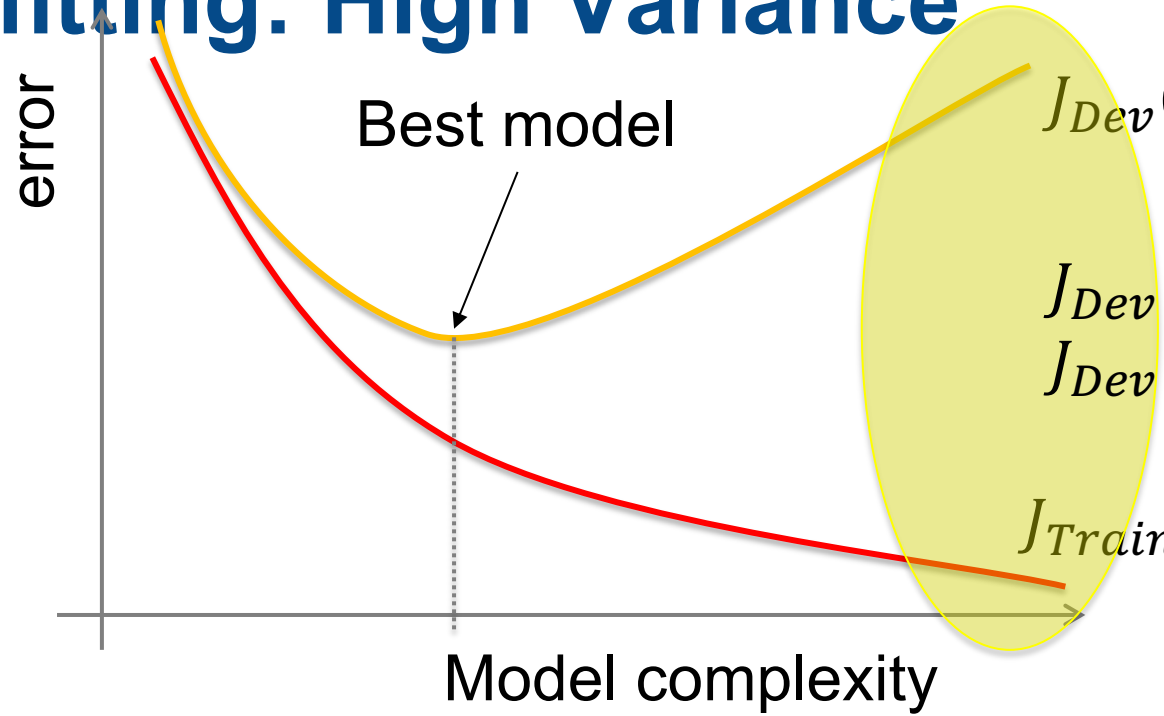
$J_{Train}(\theta)$

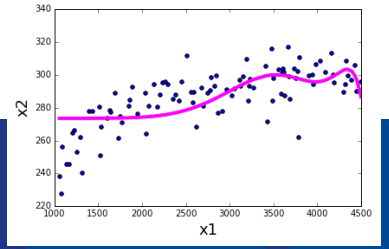
Model complexity

(e.g. degree of polynomial)

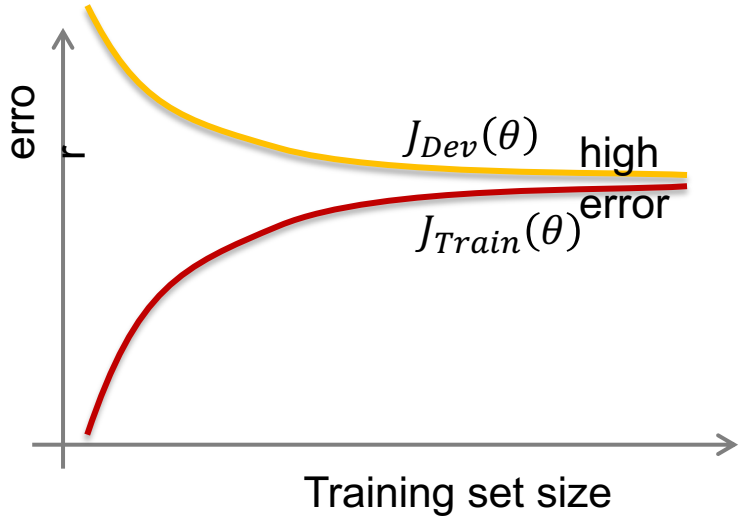


Overfitting: High Variance

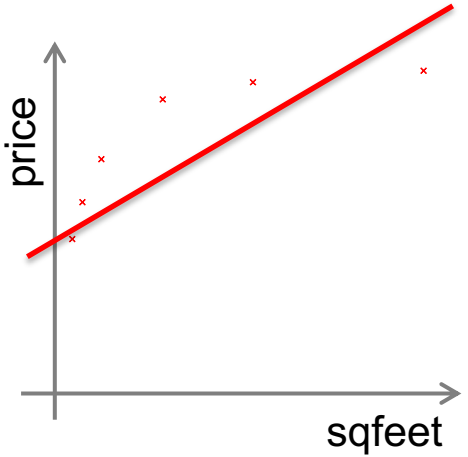


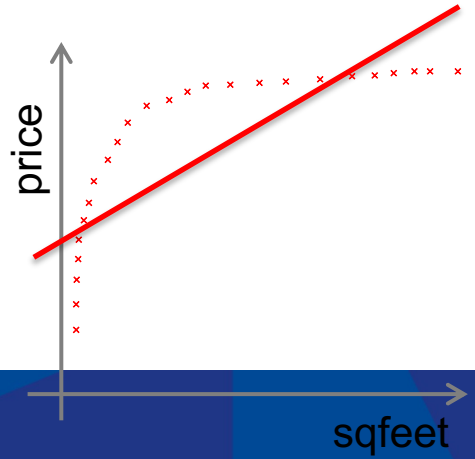


Underfitting Remedies

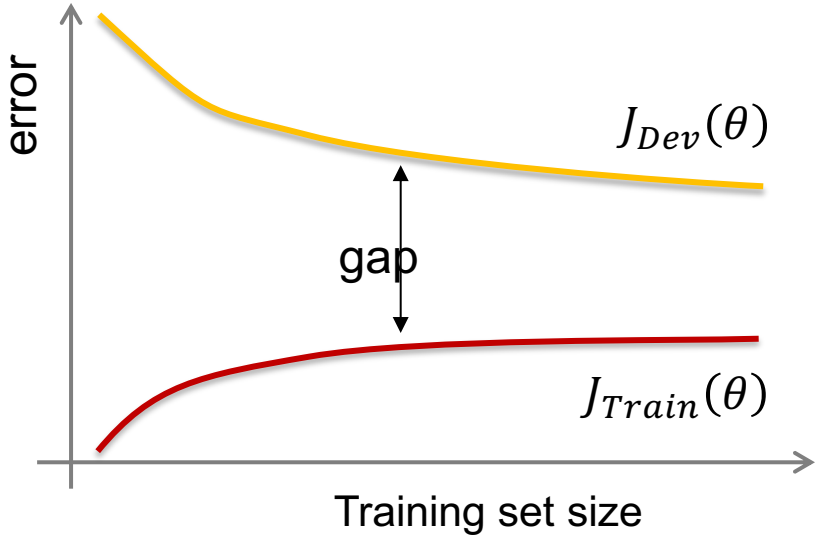


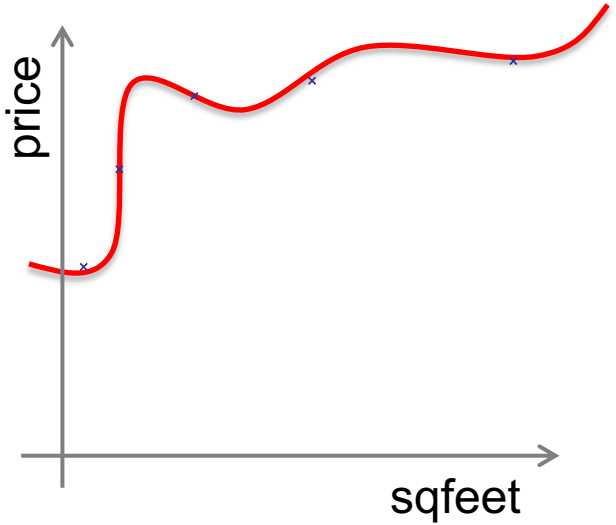
- Change the model
- Adding data won't help

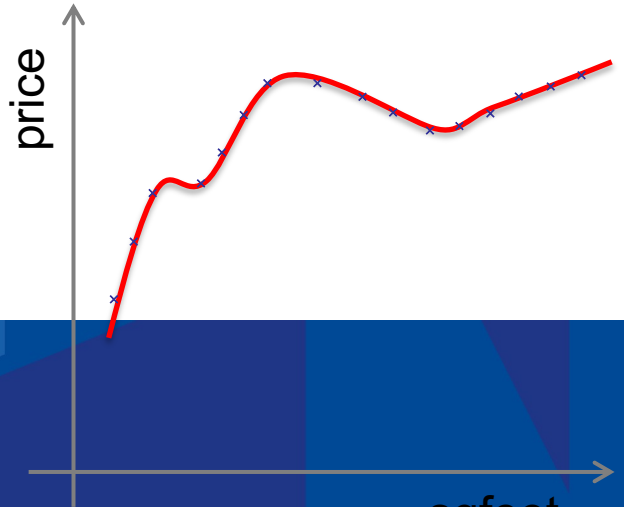




Overfitting Remedies





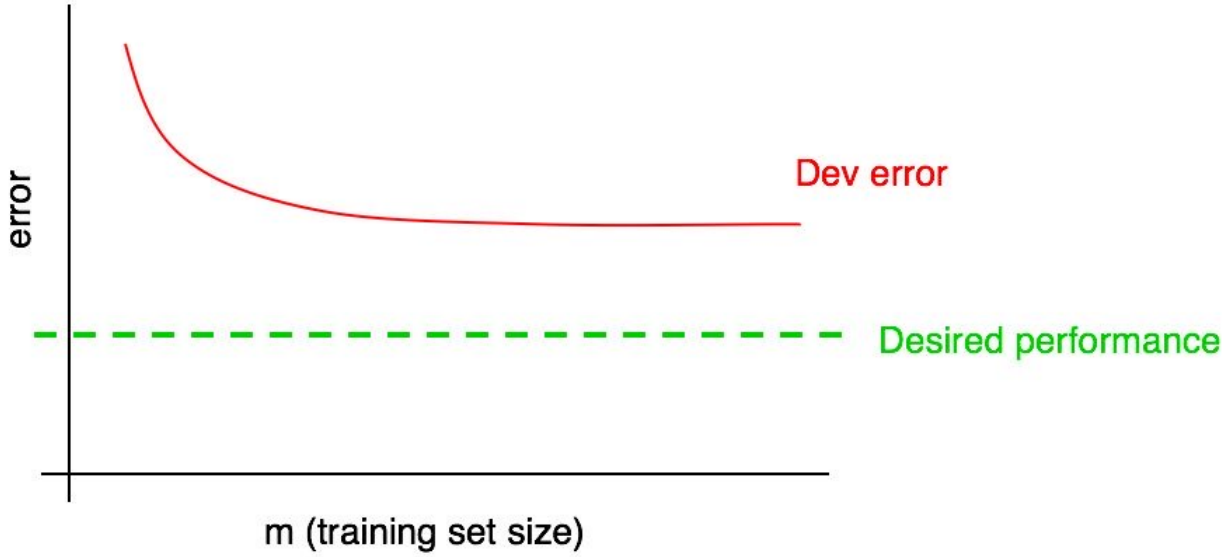


- Adding more data may help to estimate parameters more accurately.

Performance Remedies

- Get more training examples: fixes overfitting
- Remove features: fixes overfitting
- Add features: fixes underfitting
- Impose a penalty on parameter values: fixes overfitting
- Decrease penalty: fixes underfitting
- Change the model architecture: fixes either

Learning Curves



Building Highly Accurate ML Systems

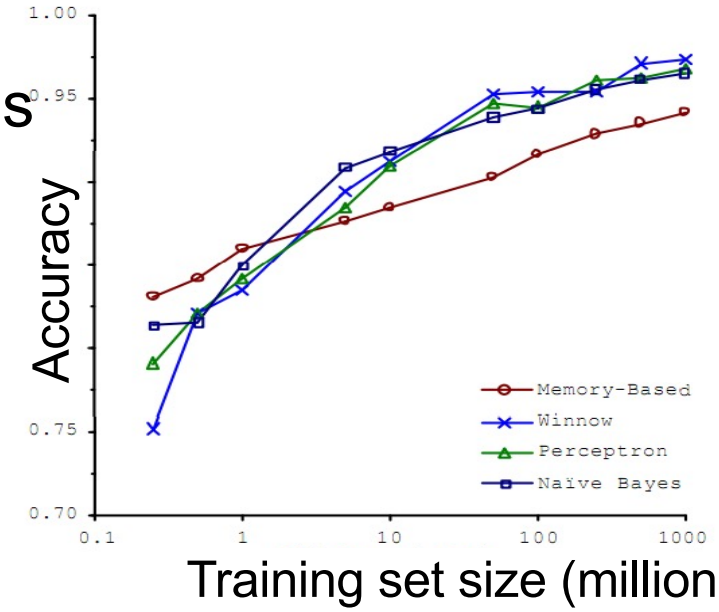
- It is all about data
- Classify between confusable words

  {to, two, too}, {then, than}

  For breakfast I ate _____ eggs.

“It’s not who has the best algorithm that wins.
It’s who has the most data.”

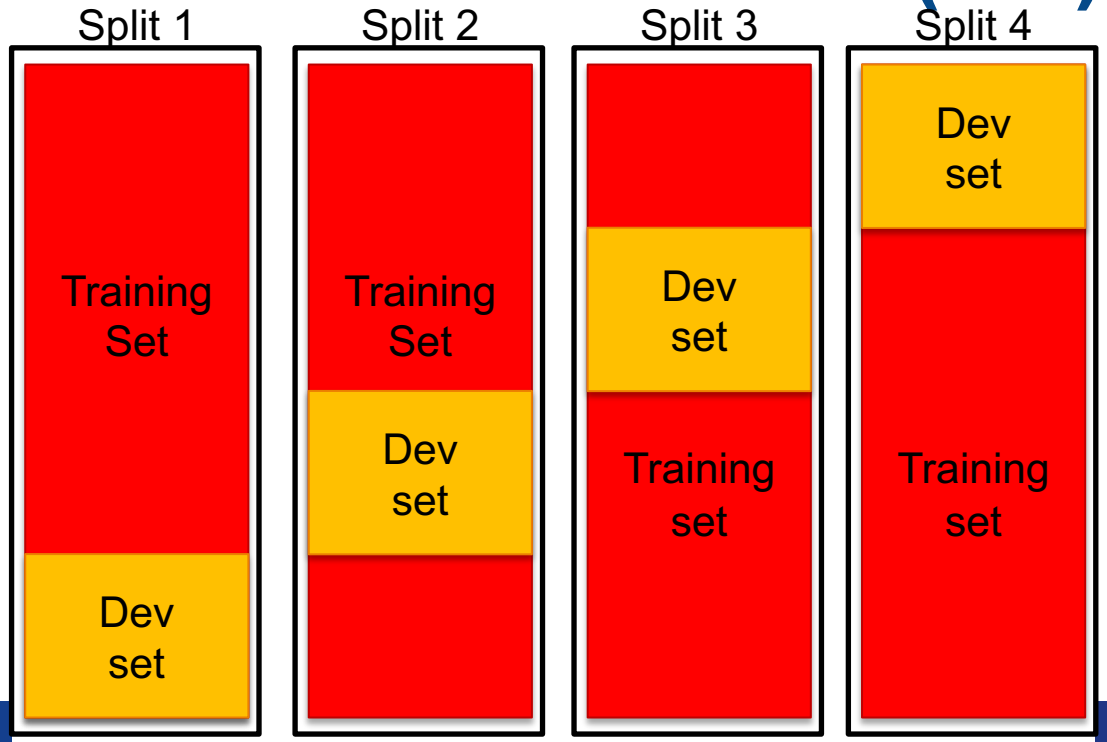
[Banko and Brill, 2001]



Large Data Rationale

- Use a learning algorithm with many parameters
  - e.g. logistic regression/linear regression with many features
  - Low bias algorithm can learn complex concepts
  - $J_{Train}(\theta)$ will be small
- Use a vary large training set
  - unlikely to overfit, low variance
  - $J_{Train}(\theta) \approx J_{Test}(\theta)$

K-Fold Cross-Validation (CV)



4-fold Cross Validation: average the results

The partitions must be formed randomly!

CV Training Process

- Can use CV for model selection
  - Train n models and evaluate each with (k-fold) CV
  - Select the model that exhibits best results
  - Re-train the model on all of the training data
- Estimate performance on unseen data using the Test set

What Could Go Wrong?

- The actual distribution you need to do well on is different from the dev/test sets
- You have overfit to the dev set
- The metric is measuring something other than what the project needs to optimize
- Datasets include lots of mislabelled samples

Error Analysis

- Determine where the model errs and what to do about it
  - Diagnosing underfit/overfit is only one form of error analysis
- Manual inspection of errors on the Dev set
- Guides development effort and bounds potential improvement
  - Error rate and frequency of types of classes determine what can be gained

Optimal Error

- The optimal achievable error rate is not always 0%
  - Unintelligible speech in audio recordings even human's cannot decipher
- Compare to human performance
  - But some problems are hard for humans

Summary

- Use separate datasets to train, tune/evaluate, and test the model
- Error analysis helps distinguish high bias from high variance issues, and select actions for improvement
- Cross-validation is a technique for evaluating models based on repeated splitting of data
- Building an ML system is a highly iterative process
- More data is usually better than improving the algorithm

University of
South Australia

Questions?