

# Evaluation II

Semester 1, 2021

Ling Luo

# Outline

- Evaluation
- Overfitting
- Model Bias and Variance
- Evaluation Bias and Variance

# Evaluation of Supervised ML

- Start with a dataset of instances comprised of attributes and labels
- Build a classifier using a learner and the dataset
- Assess the effectiveness of the classifier
  - Comparing the predictions with the actual labels on unseen instances
  - Metrics: accuracy, precision, recall, F1-score

# Inductive Learning Hypothesis

→ general and universal rule from particular example

map function

- 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.

large & representative of population

- What does it mean by “large training data set”?
- Why do we need to test our hypothesis on “held-out test examples”? generalize to new data, avoid over estimate on training data
- What impact does the size of the test set have?

too small → bias

# Tensions in Classification

Our evaluations must take these ideas into consideration

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

# Overfitting

# Learning Curves

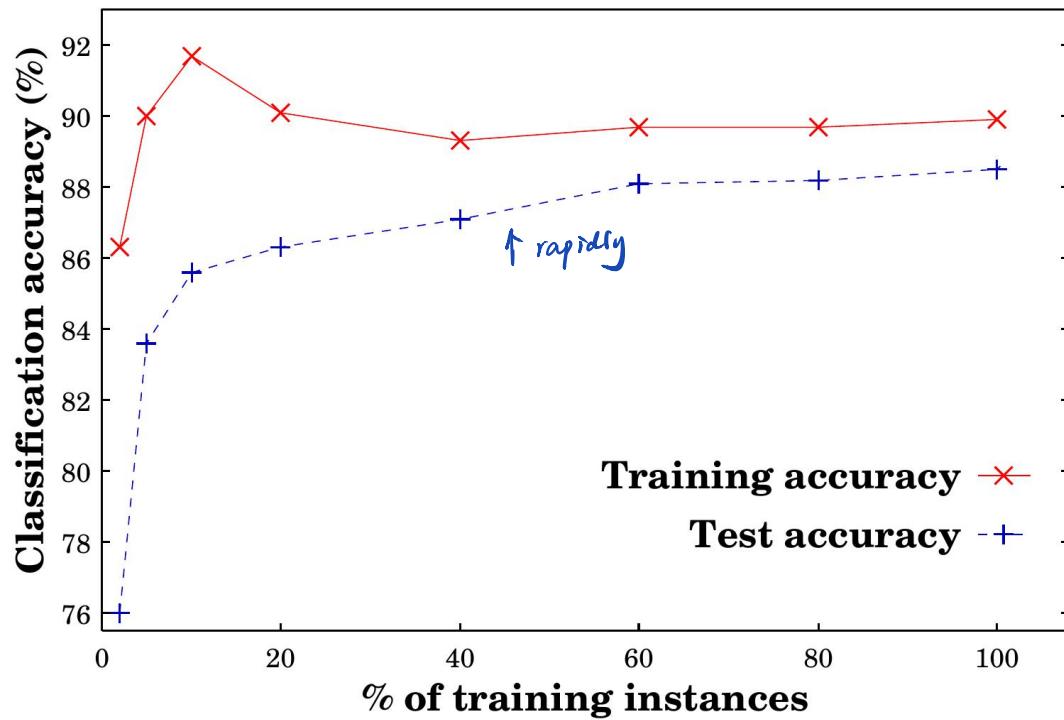
- Learning curve is a plot of learning performance over experience or time
- For machine learning models, we can plot:
  - **y-axis**: performance measured by accuracy, error or other metrics
  - **x-axis**: conditions, e.g. sizes of training sets, model complexity, iterations...
  - **Training learning curve**: calculated from the *training* set that shows how well the model is *learning*.
  - **Test learning curve**: calculated from a *holdout* set that shows how well the model is *generalising*.

# Learning Curves

## Example 1: for different sizes of training set

- Choose various split sizes, and calculate effectiveness
  - For example: 90-10, 80-20, 70-30, 60-40, 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)
- Benefit: allows us to visualise the data trade-off
  - More training instances?      better model.  $\Rightarrow$  but result may be random
  - More evaluation instances?      evaluation result is more reliable

# Learning Curves



# Learning Curves

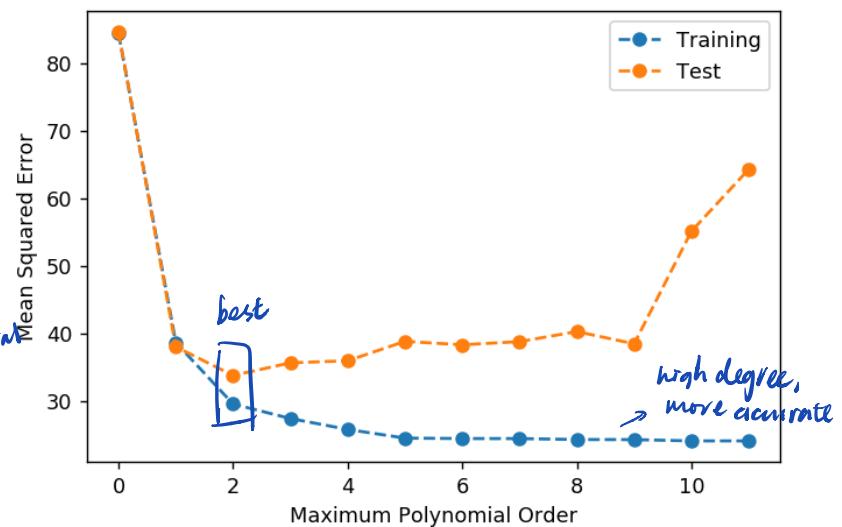
## Example 2: compare models

- Using the polynomial of  $x$  to increase the flexibility of linear regression

$$y = \mathbf{w} \cdot \phi(x)$$

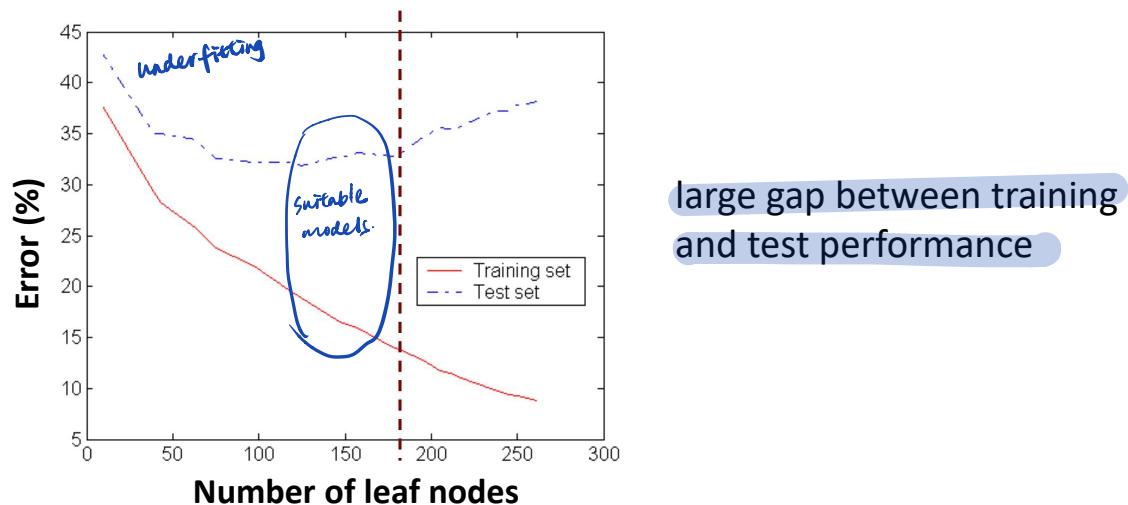
$$\phi(x) = [1, x, x^2, \dots, x^D]$$

$\uparrow$   
highest  
order  
of polynomial



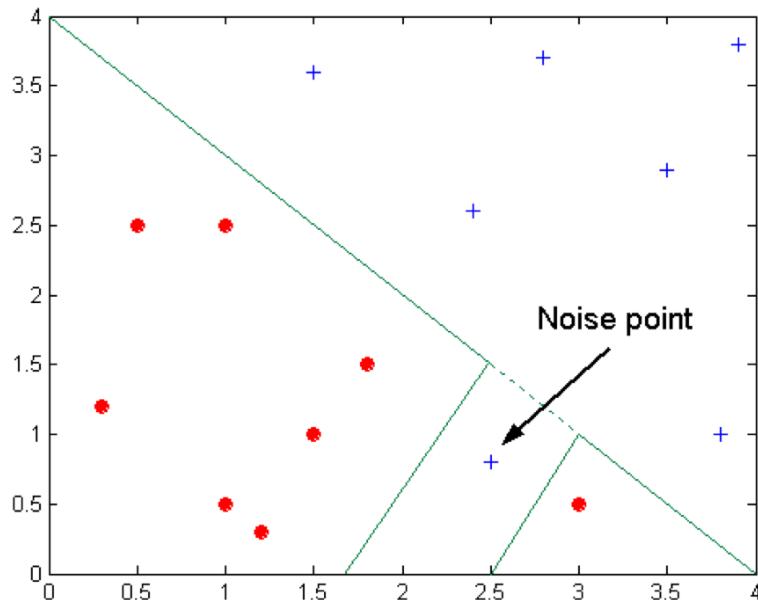
# Overfitting

- An **overly complex model** is selected that **captures specific patterns** in the training data but **fails to learn the true nature** of relationships between attributes and class labels
- Possible evidence of overfitting



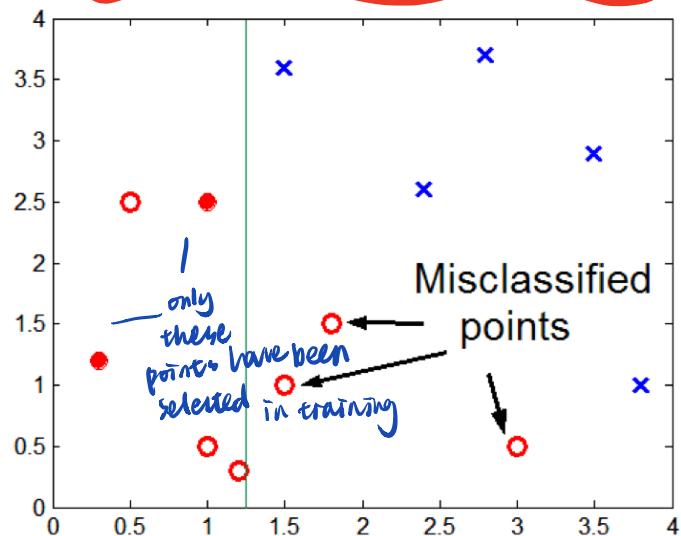
# Reasons of Overfitting

- Decision boundary distorted by noise
- A simpler decision boundary would generalise better for this data



# Reasons of Overfitting

- Limited training set: not fully represent the patterns in the population
  - could be due to small number of examples
  - could be due to non-randomness in training samples (sampling bias)



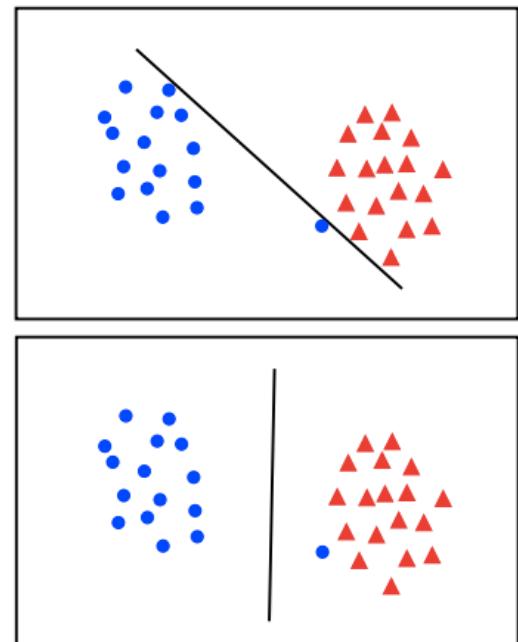
# Solution: Regularisation

- decision tree: limit leaf nodes & depth*
- In soft-margin SVM, the slack variables provide regularisation

$$\min_w \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i$$

subject to  $y_i(w^T x_i + b) + \xi_i - 1 \geq 0,$

$$\xi_i \geq 0, \forall i \in \{1, 2, \dots, N\}$$



# Solution: Regularisation

- In linear regression, the norm of the parameters can be used as regulariser
- Mathematically, a norm is a total length of a vector. The  $L_p$  norm of  $\beta = [\beta_1, \dots, \beta_k, \dots, \beta_D]$  is defined as:

$$\|\beta\|_p = \sqrt[p]{\sum_k |\beta_k|^p}$$

eg : euclidean distance

$L_2$  norm

$L_1$  norm: manhattan distance.

- Optimisation objective function:

$$\hat{\beta} = \arg \min_{\beta} [Error(\beta; \{X, Y\}) + \lambda \psi(\beta)]$$

- $\lambda (> 0)$  is a hyperparameter, tuned empirically through trial-and-error
- $\psi(\beta)$  is regulariser

# Solution: Regularisation

- Linear regression with L2-norm regularisation (ridge regression)

$$\psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_2^2 = \sum_k |\beta_k|^2$$

- penalises parameter values by adding the sum of their *squared* values to the error term
- encourages solutions where most parameter values are *small*.  
*avoid only min error*
- Linear regression with L1-norm regularisation (LASSO)

$$\psi(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1 = \sum_k |\beta_k|$$

- penalises parameter values by adding the sum of their *absolute* values to the error term
- encourages solutions where few parameters are non-zero

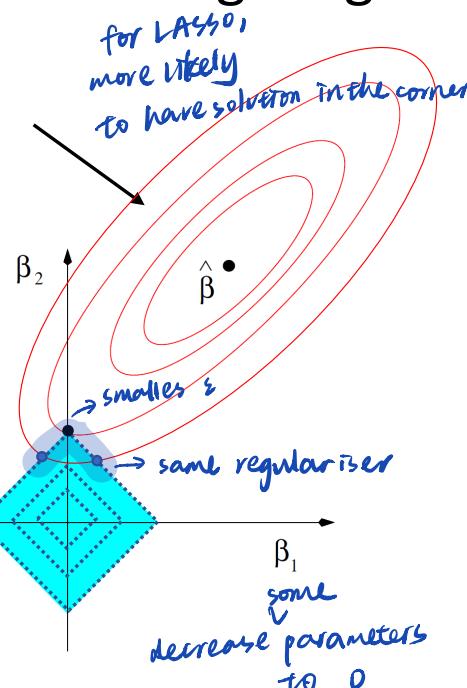
! embedded feature selection

# Solution: Regularisation

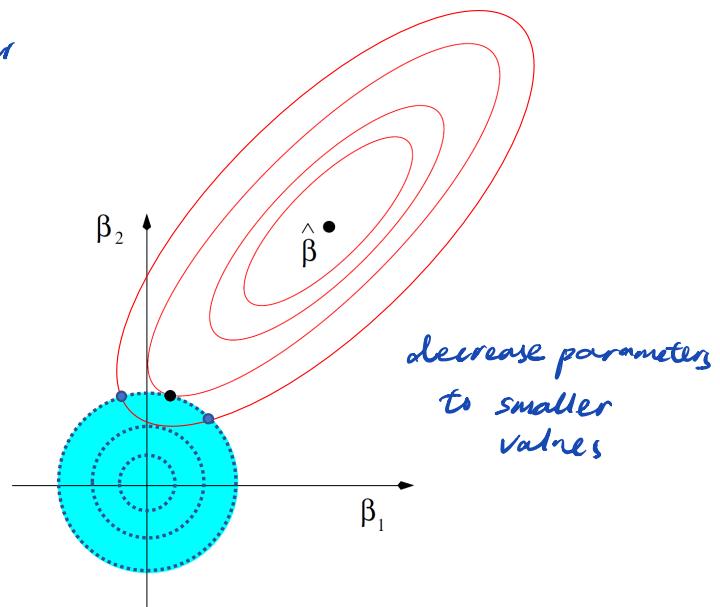
- Comparison of ridge regression and LASSO

contour of error, min value in the centre  
value in the centre

contour of regulariser



LASSO



Ridge

# Model Bias and Variance

# Bias and Variance

- Statistical definition: the bias and variance of estimation  $\hat{\theta}$  for true  $\theta$  are

$$Bias(\hat{\theta}, \theta) = E[\hat{\theta}(x) - \theta(x)]$$

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

only related to estimation rather than true  $\theta$

# Bias and Variance

## In machine learning

*true vs estimation*

- Bias is used to refer to a number of things:
  - ① Model bias: the tendency of our model to make systematically wrong predictions
  - ② Evaluation bias: the tendency of our evaluation strategy to over- or under estimate the effectiveness of our model  
*→ assume true evaluation result for our model = true accuracy is 80%*
  - ③ Sampling bias: if our training or evaluation dataset isn't representative of the population, which breaks the Inductive Learning Hypothesis
- Variance refers to model variance and evaluation variance

# Model Bias and Variance

- **Model bias** in regression context:
  - For every evaluation instance, the signed error can be calculated
  - Assuming every instance is independent, bias is the average of these signed errors
- A model is **biased** if
  - ① the predictions are systematically *higher* than the true value, or
  - ② the predictions are systematically *lower* than the true value
- A model is **unbiased** if
  - ① the predictions are systematically correct, or
  - ② some of the predictions are too high, and some of the predictions are too low

$$\left| \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i) \right|$$

keep sign    no  $|x|$  or  $(x)^2$

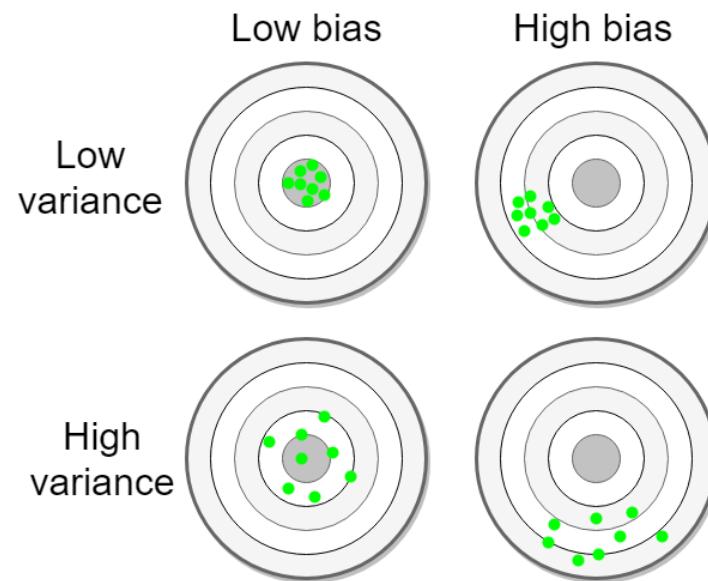
# Model Bias and Variance

- Model bias in classification context:
  - Label predictions can't be "too high" or "too low"
  - "biased towards the majority class" means our model predicts too many instances as the majority class  
*compare distribution*
  - Typically compare the class distribution:
    - An unbiased classifier produces labels with the same distribution as the actual distribution
    - A biased classifier produces labels with a different distribution from the actual distribution

# Model Bias and Variance

- **Model variance** relates to the tendency of different training sets to produce different models or predictions with the same type of learner  
different training set → same learner → different predictions.  
unstable, if changing to new samples, will give us different results
    - 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
- stable
- ✗ low variance & no bias
- ✗ low variance high bias

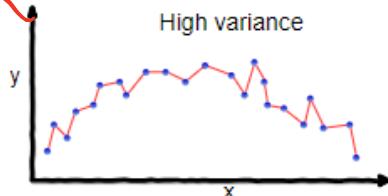
# Model Bias and Variance



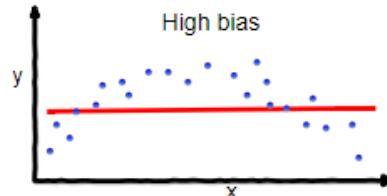
source: [www.machinelearningtutorial.net/2017/01/26/the-bias-variance-tradeoff/](http://www.machinelearningtutorial.net/2017/01/26/the-bias-variance-tradeoff/)

# Model Bias and Variance

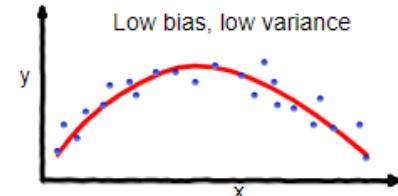
- High bias or high variance are bad
  - For example, O-R has zero variance but high bias.
  - It is important to keep balance.
- Lower bias and lower variance --> better generalisation



overfitting



underfitting



Good balance

source: <https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229/>

# Evaluation Bias and Variance

# Evaluation

evaluation 的过程和 training 过程相似。  
==

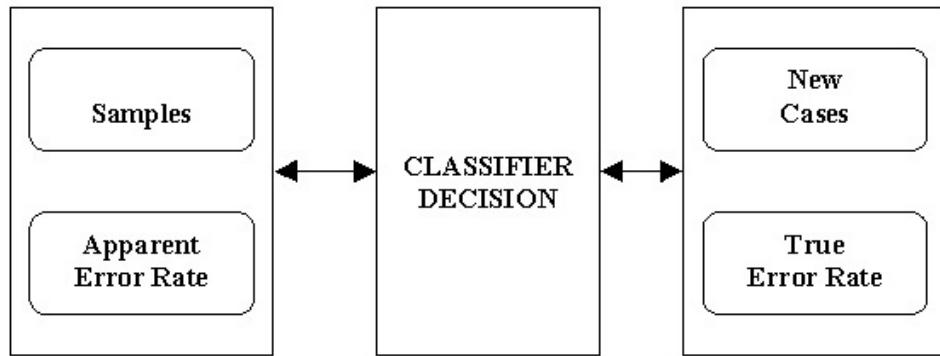
- The evaluation metric is also an estimator
- Desire to know the *true error rate* of a classifier, but only have an estimate of the error rate, subject to some particular set of evaluation instances
- The quality of the estimation is independent of the trained model

# Evaluation

apparent error rate  $\Rightarrow$  error rate measured by training data

true error rate  $\Rightarrow$  performance on population /  
unseen new cases

- For example, **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  
*when sample = population*



source: [http://dms1.irb.hr/tutorial/tut\\_mod\\_eval\\_3.php](http://dms1.irb.hr/tutorial/tut_mod_eval_3.php)

# Evaluation Bias and Variance

- Evaluation bias

$$\text{Bias}(\hat{\theta}, \theta) = E[\hat{\theta}(x) - \theta(x)]$$

*estimated error rate vs true error rate  
problem: can't get  
true error rate*

- 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 test set
- This can be hard to distinguish from model variance

*can't quantify but can compare*

# Evaluation

Holdout:  
→ given a certain dataset, balance between model variance and evaluation variance

How do we control bias and variance in evaluation?

(only one)

- Holdout partition size
  - More training data: less model variance, more evaluation variance
  - Less training but more test data: more model variance, less evaluation variance
- Repeated random subsampling and K-fold Cross-Validation
  - Less variance than Holdout

eg  
 $80:20$

more training data

limited test data

enough training & test data

- Stratification: less model and evaluation bias
  - representative of population
- Leave-one-out Cross-Validation
  - No sampling bias, lowest bias/variance in general

N total N-1 training 1 test → max training instance & make sure every

without stratification:

minor class may be ignored and hard to predict

instance being test once

train on maximum sampling instance we can have

⇒ time consuming

# 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

Pang-Ning Tan, Michael Steinbach, Anuj Karpatne, and Vipin Kumar. Introduction to Data Mining. Pearson, 2018.

the model overfits to training set. In this case, the training error may decrease, but test error will increase.

**Question 2** 0 / 1 pts

During training process, if your model shows significantly different performance across different training sets, which of the following is NOT a valid way to reduce this variance?

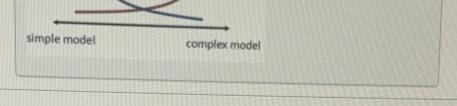
You Answered

Decrease model complexity

Improve optimisation algorithm used for error minimisation

Reduce noise in training

Improving your optimisation algorithm would decrease the bias. To reduce variance, using the other three options would be helpful. To decrease model complexity, you can consider reducing the number of features or using regularisations.



**Question 4** 0.67 / 1 pts

What are the possible solutions to reduce evaluation variance?

Correct!

Correct!

Correct!

You Answered

Stratification

To reduce evaluation variance, one can increase the size of test set, or evaluate multiple times using repeated random subsampling or K-fold cross-validation, and get the average performance across different runs. The stratification generates training and test sets that contain approximately the same distribution of class labels as the overall set. The stratification can help to reduce the bias.