COMP24112: Machine Learning

Chapter 3: Machine Learning Experiments II

Dr. Tingting Mu

Email: tingting.mu@manchester.ac.uk

Content

- Sample error and true error.
- Issues of limited data.
- How to estimate the error of your model by splitting data?
- How to set a complete machine learning experiment?
- Bias-variance decomposition.

Sample Error and True Error

- **Hypothesis**: Prediction made by a trained machine learning model.
- **Sample error** of a hypothesis ($error_S$):

  Error computed by a performance metric using a set of data samples.

- **True error** of a hypothesis ($error_D$):
  - For classification, it is the **probability that a single sample is misclassified**, where the sample is randomly drawn from a distribution.
  - For regression case, it is the **expectation of the error**. See example:

sample error $\frac{1}{n}\sum_{i=1}^{n}\left(y_i - f\left(\mathbf{x}_i\right)\right)^2$ $(\mathbf{x}_i, y_i) \sim p(\mathbf{x}, y)$

true error $E_{p(\mathbf{x},y)}\left[\left(y - f(\mathbf{x})\right)^2\right]$

Sample Error and True Error

- What we wish to know is the true error.
- In most cases, it is very hard or not possible to compute the true error.

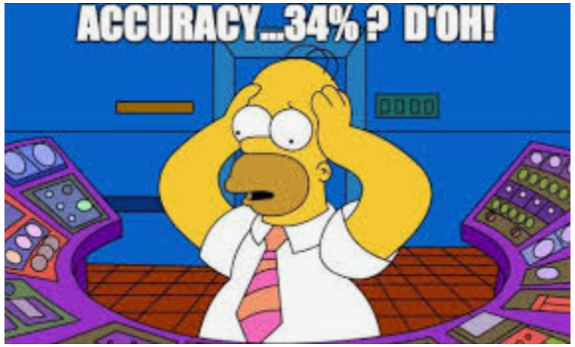
- We can always compute the sample error.
- Infinite samples: Sample error converges to true error.
- Insufficient samples: Sample error may not approximate true error well.

Limited Data: Bias and Variance Issues

Given **limited data**, you will probably encounter the following issues, caused by the gap between the true and sample errors.

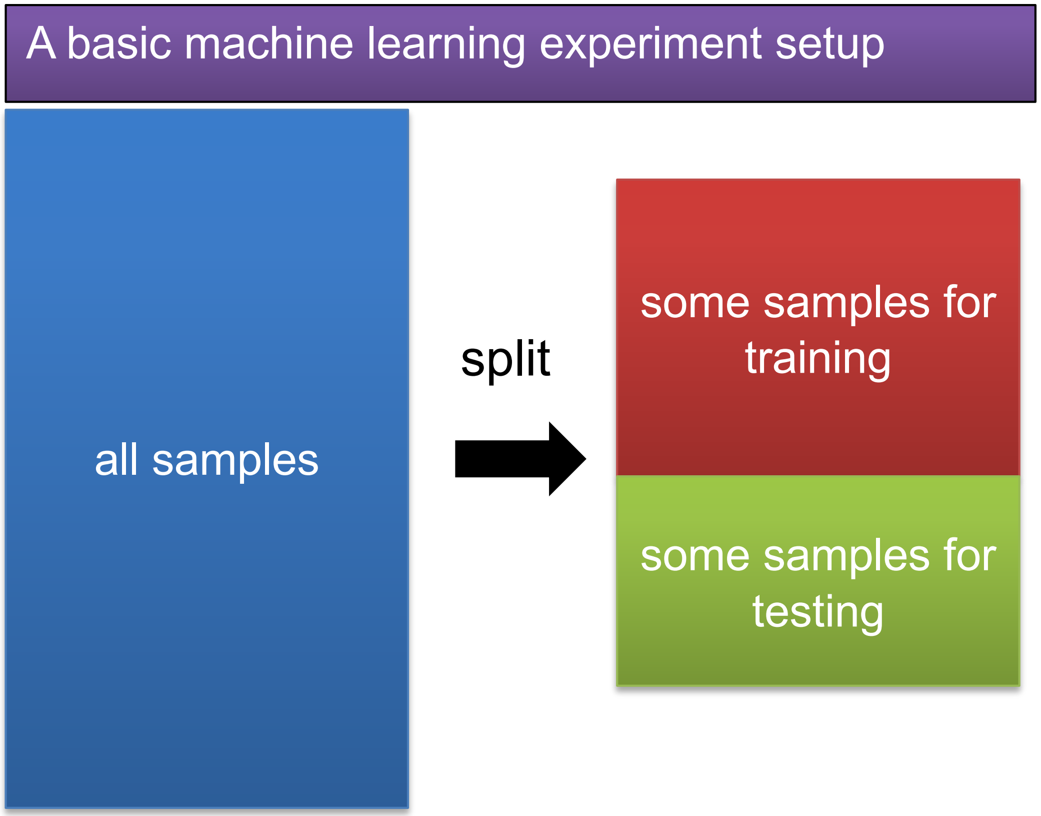
- **Bias Issue**:
  - Accuracy of the training samples can be a poor estimator of the accuracy of unseen samples.
  - It can provide optimistically biased estimate of the hypothesis over future unseen samples.
- To deal with the bias issue, it is better to choose a new set of test examples independent of the training examples.
- **Variance Issue**:
  - Accuracy of a new set of test samples can still vary from the true accuracy, depending on the makeup of a particular set of test samples.
  - Smaller set of test samples can result in higher variance.

Given a set of finite samples, how to train and evaluate a machine learning model?

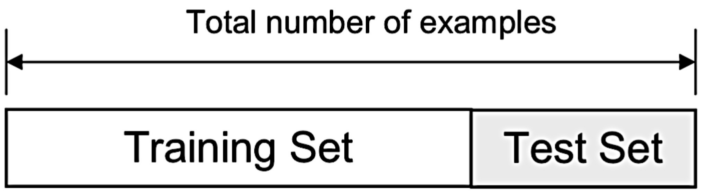


Holdout Method

- **Holdout method**: Split up your dataset into a training and test set.
  - Train your model using the training set.
  - Estimate your model error using the test set.



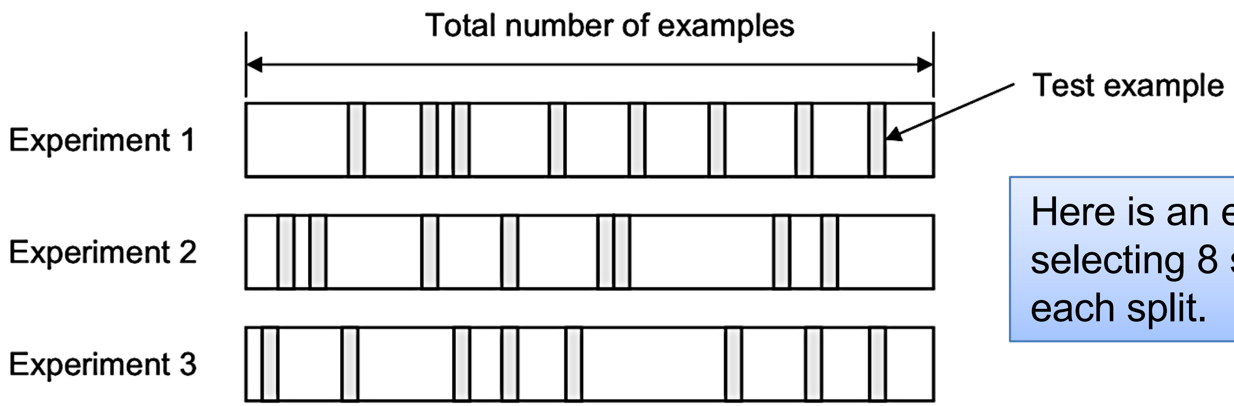
Drawback of Holdout Method



- Drawback of holdout:
  - If the dataset is small, we may not be able to set aside a portion of the dataset for testing.
  - The holdout estimate of error rate can be misleading if we happen to get an “unfortunate” split (sample error $\neq$ true error).
- Better methods?

Random Subsampling

- Perform K data splits of the entire dataset.
  - Each split randomly selects a fixed number of samples for testing, and uses the remaining samples for training.
  - For each data split, the classifier is trained from scratch using the training samples, and its error rate is estimated with the testing samples (denoted by E_i for the i-th split).



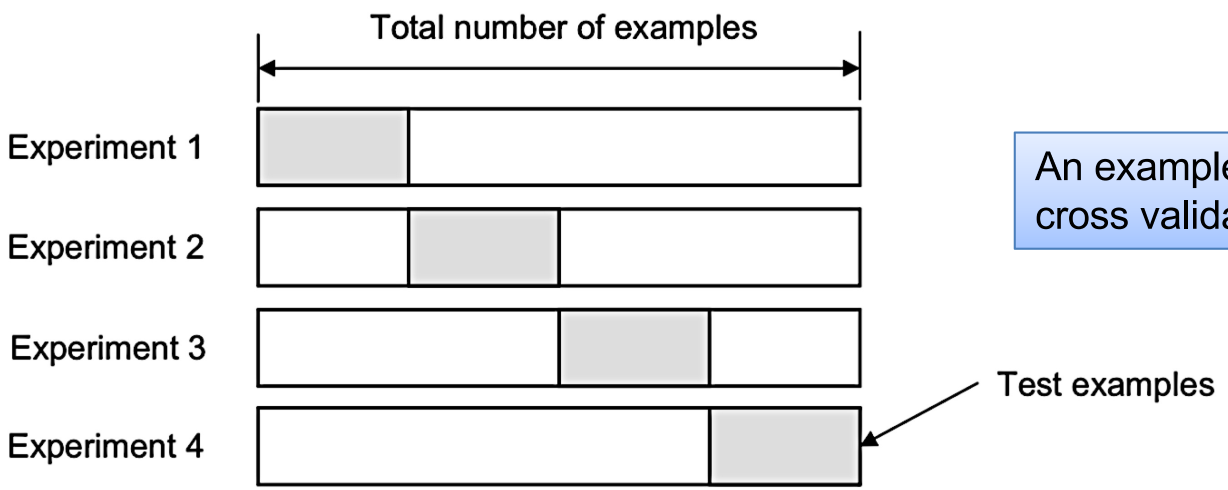
Here is an example of randomly selecting 8 samples for testing in each split.

- The final error estimate is computed by $E = \frac{1}{K}\sum_{i=1}^{K} E_i$

This slide is prepared based on Lecture 13, Introduction to Pattern Analysis, R. Gutierrez-Osuna.

K-fold Cross Validation

- Divide the entire dataset into K partitions.
  - For each of the K experiments, use (K-1) partitions for training and the remaining one for estimating the error rate E_i.



An example of 4-fold cross validation

- The final error estimate is computed by $E = \frac{1}{K}\sum_{i=1}^{K} E_i$
- Advantage: all the examples in the dataset are eventually used for both training and testing.

This slide is prepared based on Lecture 13, Introduction to Pattern Analysis, R. Gutierrez-Osuna.

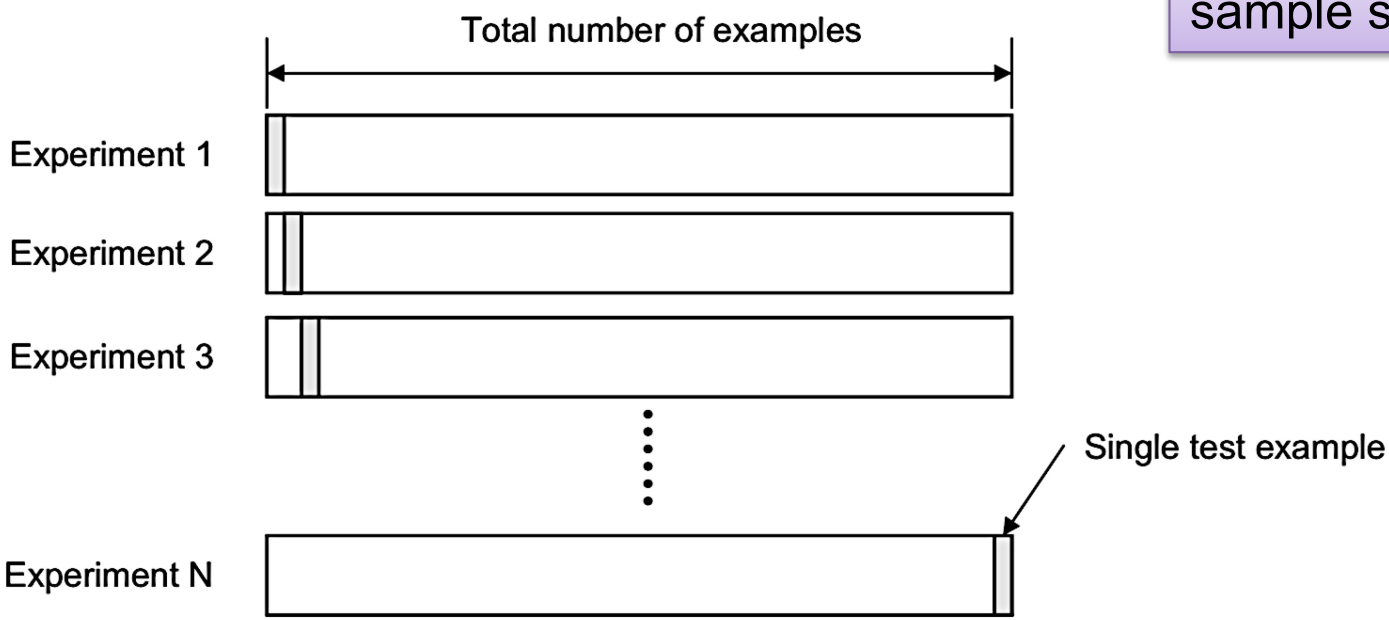
Comments on K-fold CV

- Each sample is used as the testing samples only once, but as the training samples K-1 times.
- All the test sets are independent, but there is overlapping between training sets.
- Low number of K results in insufficient training-testing trials.
- High number of K results in small testing set potentially with high variance.
- Some standard settings: 10-fold CV, 5-fold CV.

Leave One Out

- Leave one out (LOO) is a special case of k-fold CV.
  - We have a total of N samples.
  - LOO is an N-fold cross validation.

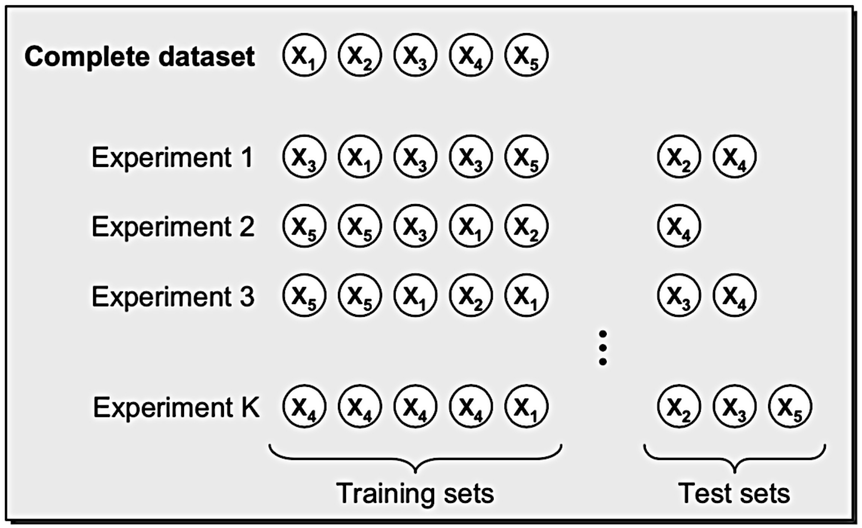
Suitable for small sample set.



This slide is prepared based on Lecture 13, Introduction to Pattern Analysis, R. Gutierrez-Osuna.

Bootstrap

- Bootstrap is based on **sampling with replacement**.
- Repeat the following process K times:
  - Randomly select (with replacement) M samples and use these for training.
  - The remaining samples that were not selected are for testing. The number of testing samples can change over repeats.



<u>Sampling with Replacement</u>:

Choose a sample from the given set, put that sample back to the set, and then choose another sample.

- The final error estimate is computed by $E = \frac{1}{K}\sum_{i=1}^{K} E_i$

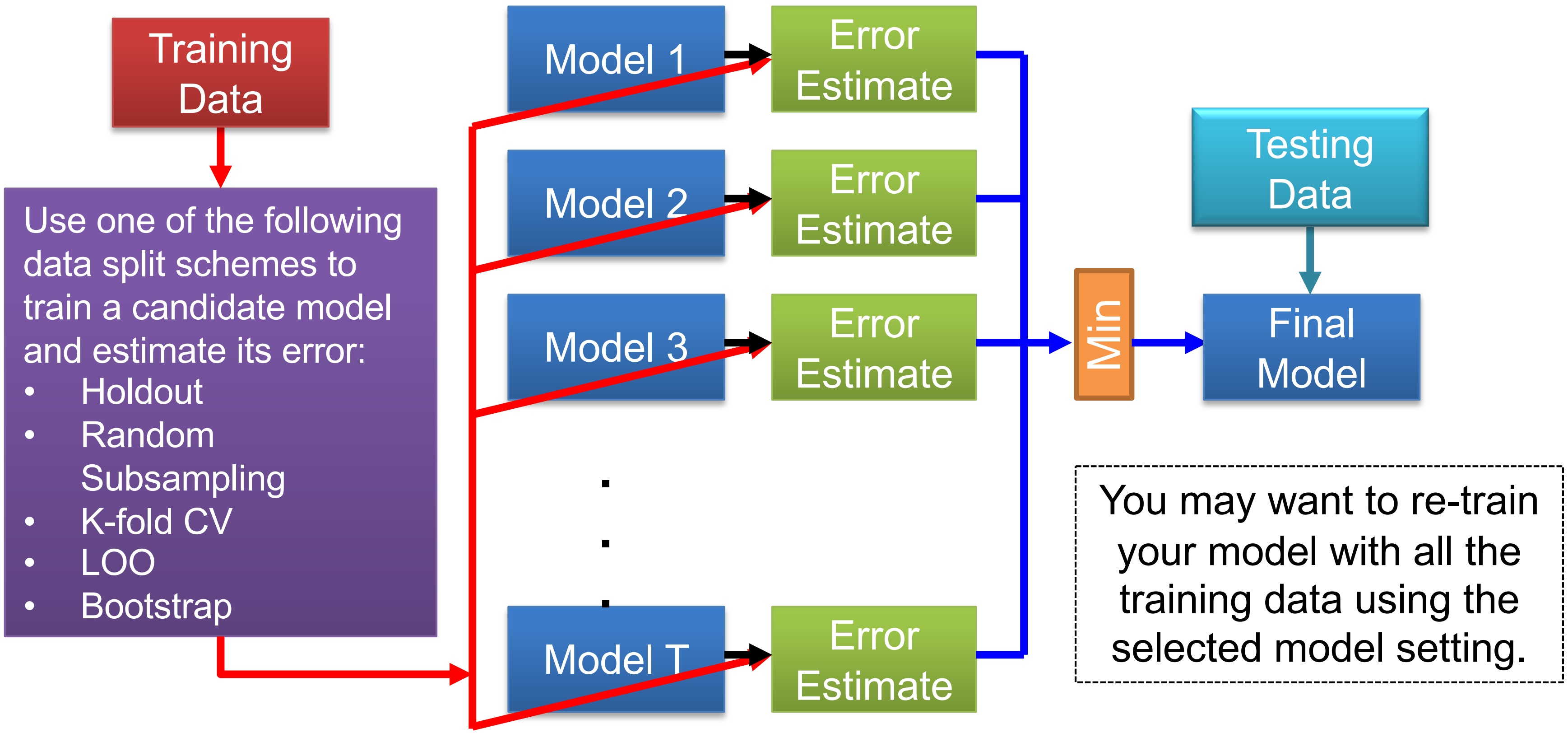
This slide is prepared based on Lecture 13, Introduction to Pattern Analysis, R. Gutierrez-Osuna.

Machine Learning Experiments

- A complete machine learning experiment includes
  1) Model training
  2) Model evaluation
  3) Model selection: Select a best model among different options (also known as hyper-parameter selection, model selection)
- Do not train, assess and select hyper-parameters using the same sample set.
- You need to split the data with an appropriate strategy, utilising, e.g., hold-out, random subsampling, K-fold CV, LOO, Bootstrap.

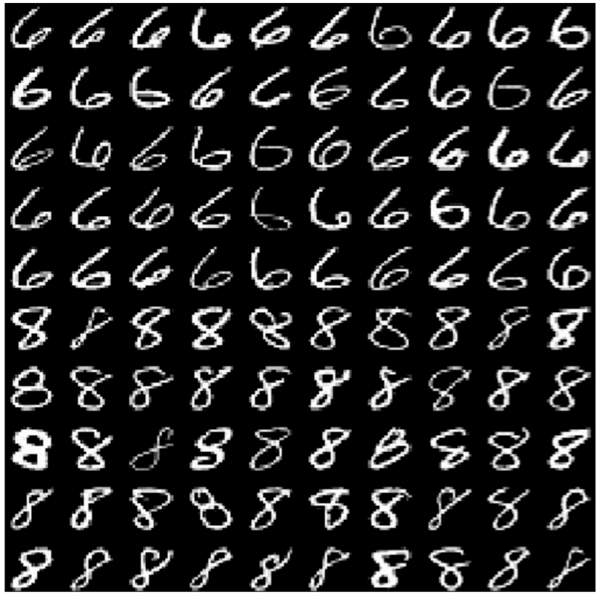
Example: Hyper-parameter Selection

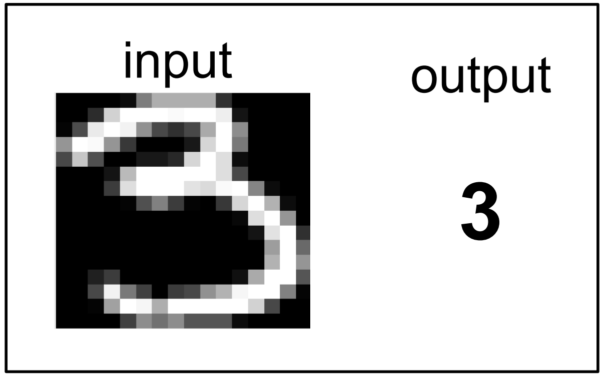
- Different hyper-parameter settings correspond to different model options.
- An example hyper-parameter selection strategy: Holdout for testing with random subsampling embedded for model selection.



Experiment: Handwriting Digit Recognition

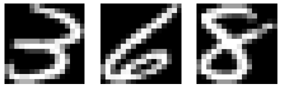
- US Postal Service handwritten digit data, https://www.dropbox.com/s/9gamjq7rpdxdi9s/postaldata.mat?dl=0.
- It includes the actual images and label variables of the digits 0-9 (500 examples per digit class, and 256 pixels per image).
- The task is to predict what digit class an input image belongs to.



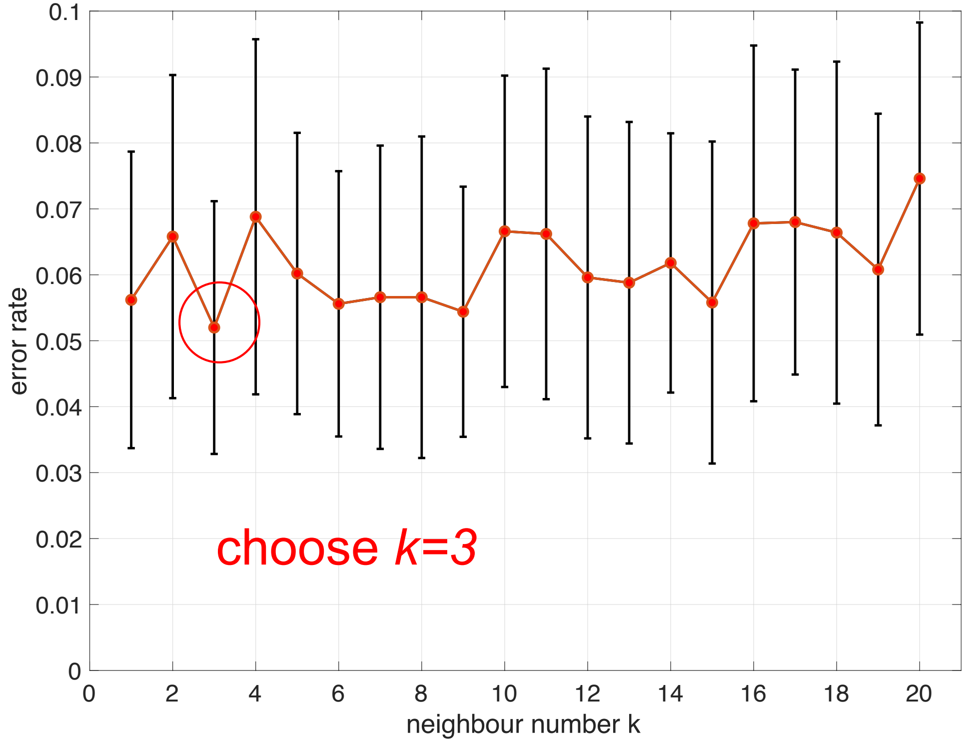


Example: Neighbour Number Selection for k-NN

- **3-class k-NN**: recognise digits 3,6 and 8.



- Given 400 training samples.
- 20 options of neighbour number k are checked (k=1,2,…20). These correspond to 20 model options.
- Random subsampling is used to estimate the performance of each model.
  - In each trial, 300 out of the 400 samples are randomly selected to train the classifier and use the other 100 to calculate the error rate.
  - Run 50 trials in total.
- k=3 is chosen with the smallest averaged error.



The mean and standard deviation of the 50 error rates is shown as an **error bar** for each hyperparameter option. There are 20 error bars in the plot.

Bias-Variance Decomposition

The University of Manchester

- Take regression as an example.
- Expected squared error for a new test sample:

$$E\left[\left(y - f\left(x\right)\right)^2\right]$$

- With some calculation, it has

$$E\left[\left(y - f\left(x\right)\right)^2\right] = \underbrace{E\left[\left(f\left(x\right) - \hat{y}\right)^2\right]}_{\text{variance error}} + \underbrace{\left(y - \hat{y}\right)^2}_{\text{bias error}}, \text{where } \hat{y} = E\left[f\left(x\right)\right]$$

- There are various explanations of the expectation range.
- One scenario is the possible choices of training samples.

Trick:

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

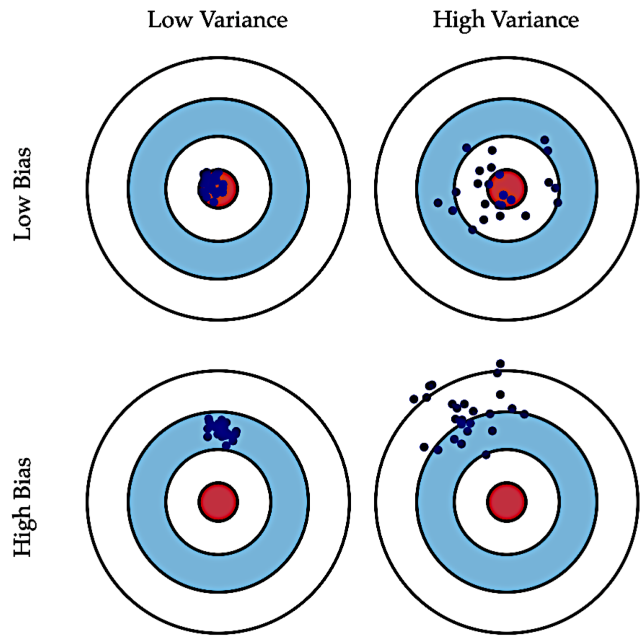
Bias-Variance Decomposition

- Bias error: $\left(y - E\left[f\right]\right)^2$

Bias error measures how much the averaged prediction is close to the true value.

- Variance error: $E\left[\left(f - E\left[f\right]\right)^2\right]$

Variance error measures how much the prediction varies among different realisations of the model.



http://scott.fortmann-roe.com/docs/BiasVariance.html

- Over-fitting: low bias error, high variance error, e.g., an over complex model that is sensitive to small fluctuations in the training set.
- Under-fitting: high bias error, low variance error, e.g., an excessively simple model.