COMP20008 2020 SM1 Workshop Week 10
Experimental design

Part A

1. What is the difference between supervised and unsupervised learning?
2. What is the difference between training data and testing data?
3. What is the purpose of k-fold cross validation?
4. Suppose Alice takes a dataset D with 100 instances, 4 features, plus a class label feature. She computes the correlation of each of the 4 features with the class label using mutual information and discards the two features with lowest correlation. She now has a processed version D' of the dataset (2 features, class label and 100 instances). She splits D' into two - 80% training (80 instances) and 20% testing (20 instances). She learns a decision tree model on the training set and evaluates the model accuracy on the testing set. She reports the accuracy as being 90%. Why might this estimate of 90% accuracy be over-optimistic? Give reasons.

Part B

The Aurin repository contains a large number of datasets that data scientists can use to help answer important questions in society. The following exercise illustrates a possible simple scenario and is designed to get you thinking about how you might use the Aurin urban data repository that is hosted by the University.

Suppose our question is *Are we building enough green spaces in Victoria to ensure a healthy population?*

- **Question 1:** Who would be interested in an answer to this question and why?
- Log in to the Aurin portal `https://aurin.org.au`
- Select Victoria as your region of interest.
- Browse through the available datasets and see what data is available.
- Add the dataset "2015 Local Government Area (LGA) Statistical Profiles". You should select all the attributes to include. This dataset includes information about number of people reporting high blood pressure across different regions in the State. We will use this as a measure of people's health.

- Add the dataset “LGA Visit to green space (once per week)”. You should select all the attributes to include. This dataset contains information about number of people who visit local green space each week, across different regions in the State.
- Download each of these datasets as a CSV file.
- **Question 2:** What feature would you use to join these datasets together?
- **Question 3:** Describe two different techniques you could use to help identify a relationship between the visits to green space and reports of high blood pressure.
- **Question 4**: Describe how you could use the data to make a prediciton about people’s overall health based on the information available. Describe the steps you might use to evalulate your prediction.
- **Question 5:** What challenges do you think might arise in studying these questions?
- **Question 6:** What are the next steps you might take when studying these questions?
- **Question 7:** (If you have time) Implement one of these techniques and report the results.

Dimensionality Reduction:

you can increase new feature

- Reasons: plotting, reducing inefficiency, *curse of dimensionality

points form in clusters in low dimension will become far away in high dimension

- PCA *

- Performance evaluation (how well it perform)
① Metrics ② Training and test set ③ sampling and cross validation

feature selection: choosing the "right" features to keep in data set

COMP20008 2020 SM1 Workshop Week 10

Experimental design

wrapper: find all subsets of features and rank them

faster greedy: start with 0 feature, try to add one feature at a time (yield the highest accuracy gain)

decremental: start with all features, try to remove one feature at time (less impontent one)

Part A

1. What is the difference between supervised and unsupervised learning?

   - supervised learning: we have a target (eg. in classification, it is the class label)
   (- unsupervised learning: no target)

2. What is the difference between training data and testing data?

   - training data: used to build a model
   - test data: used to test a model's performance.

3. What is the purpose of k-fold cross validation?

   split a data set into different parts, train and test a moder for each split and report the average

   * to make sure that the model does not have access set when it's trained

   if we allow the test set to be involved in the training process the performance of the model will be unrealistically high (overfitting)

4. Suppose Alice takes a dataset D with 100 instances, 4 features, plus a class label feature. She computes the correlation of each of the 4 features with the class label using mutual information and discards the two features with lowest correlation. She now has a processed version D' of the dataset (2 features, class label and 100 instances). She splits D' into two - 80% training (80 instances) and 20% testing (20 instances). She learns a decision tree model on the training set and evaluates the model accuracy on the testing set. She reports the accuracy as being 90%. Why might this estimate of 90% accuracy be over-optimistic? Give reasons.

   - over-optimistic accuracy based on very small test set
   + feature selection has been done using BOTH the training data and test data
   - It's possible that feature selection done on the training data only would have selected diffent feature

   reduce the likehood of overfitting and selection bias, provide a more accurate assessment of the model's performance

   · For a fair assessment, the test data should not be involved in the model generation in any way. split the test-train should be the first thing to be done

Part B

The Aurin repository contains a large number of datasets that data scientists can use to help answer important questions in society. The following exercise illustrates a possible simple scenario and is designed to get you thinking about how you might use the Aurin urban data repository that is hosted by the University.

Suppose our question is *Are we building enough green spaces in Victoria to ensure a healthy population?*

- **Question 1:** Who would be interested in an answer to this question and why?
- Log in to the Aurin portal `https://aurin.org.au`
- Select Victoria as your region of interest.
- Browse through the available datasets and see what data is available.
- Add the dataset "2015 Local Government Area (LGA) Statistical Profiles". You should select all the attributes to include. This dataset includes information about number of people reporting high blood pressure across different regions in the State. We will use this as a measure of people's health.

- Add the dataset "LGA Visit to green space (once per week)". You should select all the attributes to include. This dataset contains information about number of people who visit local green space each week, across different regions in the State.

- Download each of these datasets as a CSV file.

- **Question 2:** What feature would you use to join these datasets together?

  LGA code

- **Question 3:** Describe two different techniques you could use to help identify a relationship between the visits to green space and reports of high blood pressure.

  ① scatter plot ② pearson, mutual information, γ^2

- **Question 4**: Describe how you could use the data to make a prediciton about people's overall health based on the information available. Describe the steps you might use to evalulate your prediction.

- **Question 5:** What challenges do you think might arise in studying these questions?

  ① missing value ② difficult to choose the most relevant ③ record is small ↓ split training & test is hard

  ④ how to identify person's health

- **Question 6:** What are the next steps you might take when studying these questions?

  ① acquire more data ② domain knowledge

- **Question 7:** (If you have time) Implement one of these techniques and report the results.