Workshop 10

COMP20008 Elements of Data Processing

Learning outcomes

By the end of this class, you should be able to:

- explain how data is used to train and evaluate machine learning algorithms
- formulate a plan to answer a business/research question using data-driven methods

Part A

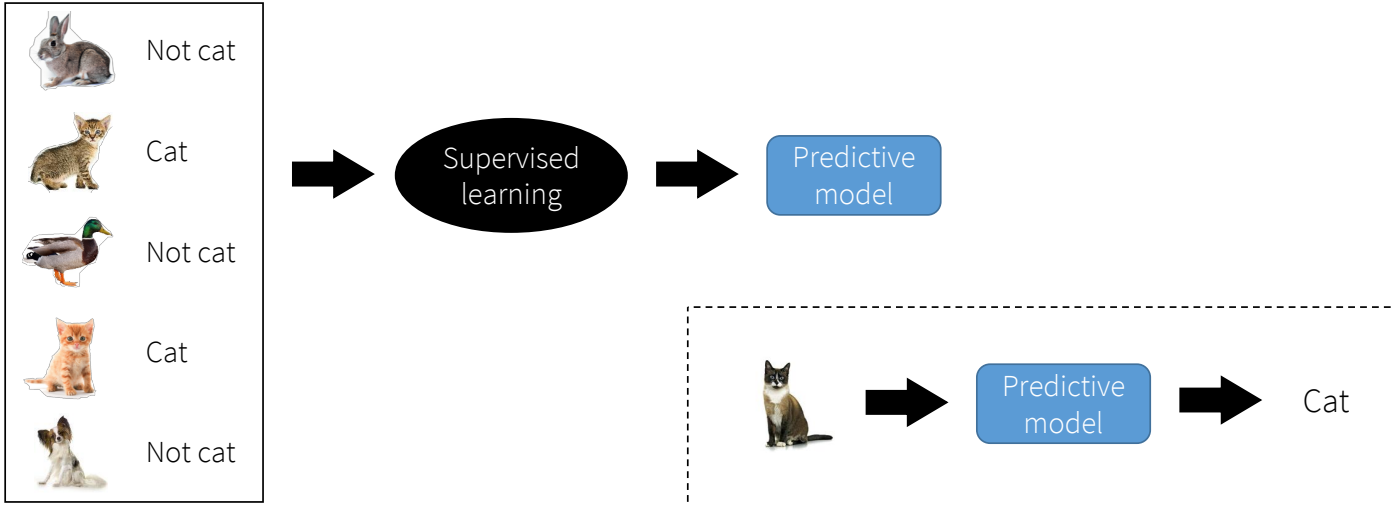
Experimental design concepts

Q1: Supervised vs. unsupervised learning

What is the difference between supervised and unsupervised learning?

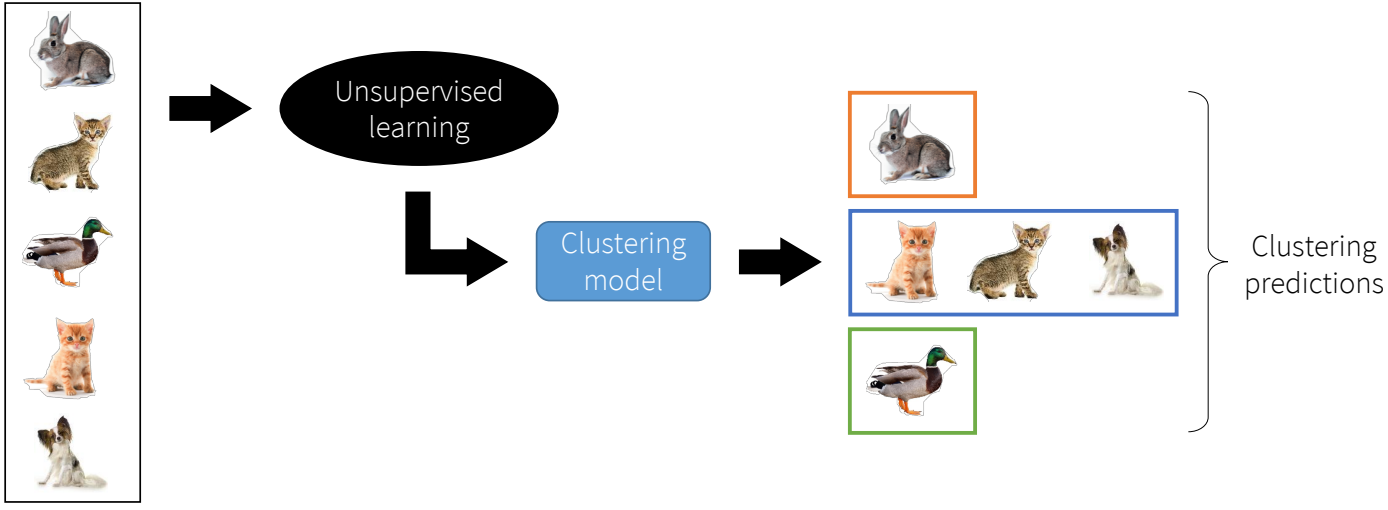
Q1: Supervised vs. unsupervised learning

What is the difference between supervised and unsupervised learning?



Q1: Supervised vs. unsupervised learning

What is the difference between supervised and unsupervised learning?



Q1: Supervised vs. unsupervised learning

What is the difference between supervised and unsupervised learning?

- Supervised learning: labelled data
- Unsupervised learning: unlabelled data

Q2: Training vs. test data

What is the difference between training data and testing data?

Q2: Training vs. test data

What is the difference between training data and testing data?

Usually the *purpose* is the only difference

- Training data: used to fit a model
- Test data: used to assess the model's performance

Note: the test data must not play any role in model selection, parameter tuning, etc. Otherwise the estimate of the model's performance may be biased (typically overly-optimistic).

Q3: k-fold cross validation

What is the purpose of k-fold cross validation?

Q3: k-fold cross validation

What is the purpose of k-fold cross validation?

Two uses:

1. Evaluation when labelled data is scarce
2. Parameter tuning—e.g. optimising *k* for a k-NN classifier

Q3: k-fold cross validation

What is the purpose of k-fold cross validation?

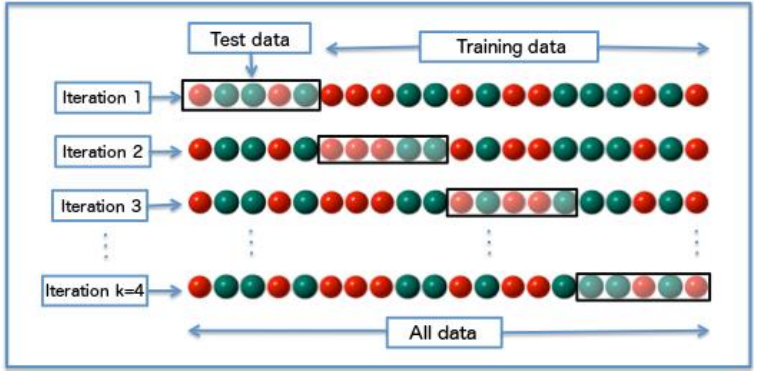


Image source: Wikimedia Commons

1. Randomly partition the data into k equal folds.
2. Repeat training and evaluation over k rounds. In each round, use one fold for testing and the remaining $k-1$ folds for training.
3. Average the performance measure over the k rounds.

Q3: k-fold cross validation

What is the purpose of k-fold cross validation?

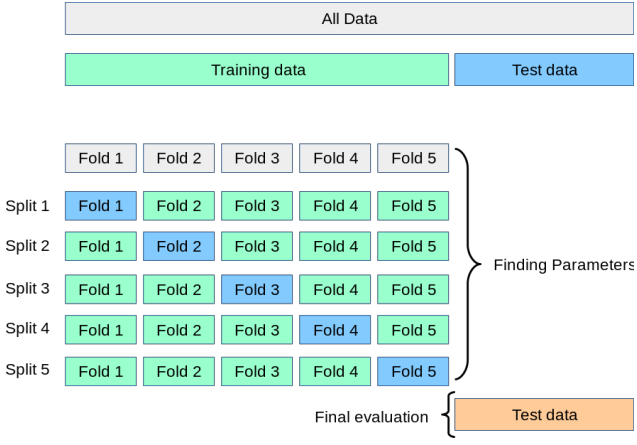


Image source: scikit-learn documentation

Grid search cross validation:

1. Randomly partition the data into k equal folds
2. Repeat training and evaluation over k rounds, using a *different parameter* in each round.
3. Choose the parameter that achieves the best performance.

Q4: Machine learning workflow

Suppose Alice has a dataset with 100 instances, 4 features and class labels. She performs the following steps:

1. Feature selection: compute the mutual information (MI) between each feature and the class label and discards the two features with the lowest MI.
2. Splitting data: randomly split the data into train/test sets with a 4:1 ratio
3. Model fitting: fit a decision tree model on the training set
4. Evaluation: compute the accuracy on the test set

She reports the accuracy as being 90%.

Why might this estimate of 90% accuracy be overly optimistic? Give reasons.

Q4: Machine learning workflow

Why might this estimate of 90% accuracy be overly optimistic? Give reasons.

- Feature selection was performed using the training and test data
- It's possible that different features would have been selected if only the training data was used
- The accuracy estimate may be biased (overly optimistic), since the test data influenced the model development

Part B

Practicing the data science process

Research question

“Are we building enough green spaces in Victoria to ensure a healthy population?”

Image source: https://theconversation.com/can-parks-help-cities-fight-crime-118322

Image source: https://www.abc.net.au/news/2019-09-06/climate-change-brisbane-rooftop-gardens/11477178

Q1: Understanding the business case

Who would be interested in an answer to this question and why?

Q1: Understanding the business case

Who would be interested in an answer to this question and why?

Interested parties:

- State government departments (DHHS and DELWP)
- Local councils
- Urban planners
- Researchers

Motivation: If a relationship between green spaces and public health can be established, it might be used as evidence to develop more green space. (But it's a complex question.)

Data collection

- We'll use data hosted by the Australian Urban Research Infrastructure Network (AURIN)
- Provides access to over 1000 multi-disciplinary datasets
- Related to towns, cities and communities
- https://portal.aurin.org.au

Collecting relevant data from AURIN

A: greenspace_visitation

Title: VicHealth - Visiting Greenspace More than Once per Week (LGA) 2011

Organisation: VIC_Govt_VicHealth

Abstract: This data reports on the % of people who have visited green space weekly or more often by 2006 Local Government Area boundaries.

Respondents who didn't know or refused to answer the question have been excluded from this variable.

Survey question: How often in the last 3 months would you have visited your local park, garden, oval or green space?

Source of the question: VicHealth Indicators Survey, 2011.

For more information please read the VicHealth Indicator Survey 2011 Selected Findings.

B: lga_stat_profiles

Title: Local Government Area (LGA) profiles data 2015 for VIC

Organisation: VIC_Govt_DHHS

Abstract: The 2015 profiles (released November 2016) include over 220 variables from a wide range of data sources. The profiles provide measures on a broad range of topics including: population, diversity, disadvantage and social engagement, housing, transport and education, health status and service utilisation, child and family characteristics and service utilisation. Rankings are provided to enable comparison of LGAs, along with the Victorian averages. The profiles are structured to provide a measure on each variable for each Local Government Area (LGA), and to also enable comparisons by providing rankings against all Local Government Areas as well as the Victorian measure. Profiles are also provided for former Department of Health regions, as well as the former Department of Human Services operational service areas and divisions. Note that all data is the most current available at the time of publication. Based on the Australian Statistical Geography Standard (ASGS), Non ABS Structures, 2015 (note, according to the ABS website LGA boundaries have not changed since 2014).

Q2: Data integration

What feature would you use to join the two datasets?

Q2: Data integration

What feature would you use to join the two datasets?

- Use the `lga_code` or `lga_name` attribute to perform a join. This is a unique identifier for the entity represented in each record.
- In Pandas, can use the `pd.merge` function or the `join` method associated with a DataFrame

Q3: Assessing correlations

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

Q3: Assessing correlations

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

- Variables of interest are:
  - `numeric` (in dataset A)
  - `ppl_reporting_high_blood_pressure_perc` (in dataset B)
- Check for statistical association using scatter plots, mutual information/Pearson correlation, significance tests, etc.

Q4: Making predictions

Describe how you could use the data to make a prediction about people's overall health based on the information available. Describe the steps you might use to evaluate your prediction.

Q4: Making predictions

Describe how you could use the data to make a prediction about people's overall health based on the information available. Describe the steps you might use to evaluate your prediction.

Difficult to solve using aggregate data. One possible approach:

- Select aggregate features that could be applied to an individual, e.g.
  - whether they report arthritis, heart disease, poor dental health
  - whether they meet physical activity guidelines
  - whether they're a current smoker
  - whether they rate their community as "active"…
- Use population rate for training and boolean (0/1) when making predictions for individuals

Q4: Making predictions

Describe how you could use the data to make a prediction about people's overall health based on the information available. Describe the steps you might use to evaluate your prediction.

- Use *fraction of people reporting fair or poor health status* as the target variable for classification (you would need to decide how to discretise into poor/good health)
- Split the data into train/test
- Fit a classification model on the training set
- Evaluate on the test set. (Even better: evaluate on surveyed individuals)

Q5: Challenges

What challenges do you think might arise in studying these questions?

Q5: Challenges

What challenges do you think might arise in studying these questions?

- Only have aggregate (LGA-level) data
- Variations in LGA between datasets due to changing boundaries
- Difficult to establish causal link
- Small dataset ⇒ conclusions may not be statistically significant

Q6: Next steps

What are the next steps you might take when studying these questions?

- Search for additional data sources, e.g. check data.melbourne.vic.gov.au and data.vic.gov.au
- Collect data about green space usage
- Collect data about green space capacity
- Collect individual-level data (e.g. consider conducting a survey)