THE UNIVERSITY OF MELBOURNE

Workshop 8

COMP20008

Elements of Data Processing

Zijie Xu

Agenda

- A2 group contract
- Machine learning
  - Supervised machine learning
- Feature selection with mutual information
- k-Nearest Neighbours (kNN)
- Linear regression

A2 group contract

- Everyone should have a team by now
- Group contract due in one week
  - Check your group name A2-Project X $X \in \{54,55,56,57,58,59\}$
  - Check if everyone in your team is allocated correctly
  - Canvas -> People -> Project groups

Machine learning

- Machines learn from data to make predictions/decisions
- For our subject, ML has two paradigms
  - Unsupervised: Learns pattern from unlabelled data
  - Supervised: Learns to classify/predict from labelled data
- We look at two types of supervised ML tasks
  - Classification: Predicts a discrete variable
  - Regression: Predicts a continuous variable

Supervised machine learning

- Objective: approximate f such that $y = f(X)$ from data
  - X: features, independent variables, input variables
  - y: label, response, dependent variable, target variable, output variables
  - f: model/ML algorithm
- Pipeline
  - Gather and pre-process data
  - Split data into training/validation/testing set
  - Select a ML algorithm
  - Train algorithm with training set
  - Optimise hyperparameters using validation set
  - Evaluate performance by predicting with testing set

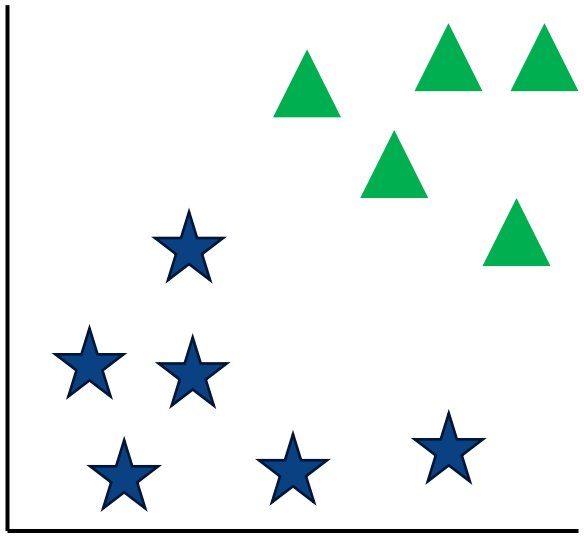
Feature selection with mutual information

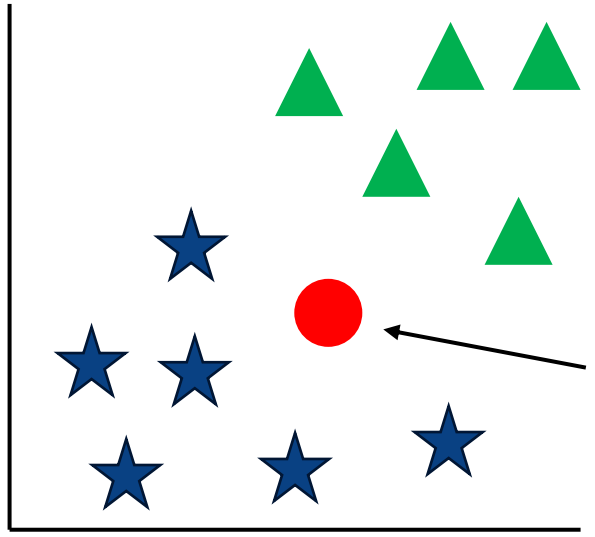
- Sometimes not all features are useful/relevant for models
- Feature selection: choosing a subset of features is enough
- One possible way is to use mutual information/information gain
  - Mutual information < threshold -> excluded from analysis

$$\mathrm{MI}(X,Y) = H(Y) - H(Y|X)$$
$$= H(Y) - \sum_{x_i \in X} \Pr(X = x_i)\, H(Y|X = x_i)$$

k-Nearest Neighbours (kNN)

- Predict label of new data by looking at the labels of nearest neighbours
  - Supervised classification algorithm
- Assumption: Similar things are near to each other
- Hyper-parameters: k, distance metric





Linear regression

- Predict value of target variable by using a weighted sum of features
  - Supervised regression algorithm
- $\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$
- The model learns the coefficients that minimises errors/loss function
  - E.g. Linear regression minimises RSS = $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$
- Assumptions:
  - Linear relationship between X and y
  - Features are independent with each other

THE UNIVERSITY OF MELBOURNE

Thank you

More Resources: Canvas