F20/21DL Data Mining and Machine Learning

Part 1 Revision Lecture

Diana Bental

Structure

- A bit of (meta!) feedback first…
  - https://goo.gl/forms/l0X05hmM1wwjOvm53
- Exam Format
- Part 1 Key Topics
- Practice Questions

Exam format

- See Thursday's lecture!
- 2 hours, answer all questions
- Electronic exam – using SURPASS clients in our labs
  - Multiple choice
  - Multiple answer
  - Essays – type in answers
  - Numbers – type in answers
- Calculator tool provided

Part 1 Key Topics

- Models, representations and data types – Lectures 1-3, 5
  - Different models and data types
  - Classification, regression, clustering
  - Data types – categorical vs numeric
  - Tables, rules, decision trees, nearest neighbour
  - Numeric predictions: Linear regression, decision boundaries, regression trees
  - Advanced: Problems of learning complex relationships using relational techniques

Part 1 Key Topics

- Computations for simple classification methods:
  - Perform nearest neighbour classification using sum of squares
    - Lecture 4
  - Perform ZeroR on a simple example
    - Laboratory 1
  - Perform OneR on a simple example
    - Laboratory 1

Part 1 Key Topics

- Data Preparation - Lecture 4
  - Discretization Binning
    - Calculations - equal width binning and equal frequency binning on simple examples
    - When and why use these and other methods, limitations
  - Normalization (scaling)
    - Attribute normalization – when to use it, and why
    - Instance normalization – when to use it, and why
    - Calculations – perform min-max on a simple example of input

Part 1 Key Topics

- Credibility measures – Lecture 7, Example sheet
  - Calculations for simple examples (e.g. from a confusion matrix for a 2 class example):
    - Success rates (accuracy) and error/failure rates
    - Identify true positive, false positive, true negative, false negative
    - Sensitivity and Specificity
    - Recall and Precision
  - Reasons for using these and other measures, limitations and trade-offs between them
  - Other measures F-measure, Kappa, cost-based measures – purpose, when/how to use

Part 1 Key Topics

- Validation – Lecture 8
  - Resubstitution, Holdout, Cross-validation methods
    - What they are, when and how to use them
  - Other topics
    - Stratification
    - Methods for small datasets
      - Leave-one-out, bootstrap

Part 1 Key Topics

- Attribute selection – Lecture 9
  - Why do it?
  - Methods
    - Wrapper methods
      - Forward and backward methods
    - Forward methods
      - Heuristic methods
      - Correlation-based
    - Feature extraction
    - Pros and cos of different methods

Practice questions for Part 1

- Try them with Surpass!
  - Exam-style questions for Part 1 topics
  - Vision > Learning Materials > Part 1 Example Questions
    - View the pdf for instructions
  - Feedback and example answers provided
    - The pdf tells you to how to get the feedback too
  - If anything doesn't work, or doesn't make sense, then let me know.