Brief Exam Overview

Date and time: [redacted]

Materials: Closed book, UoW-approved calculator

Composition of my part:
- Some multiple choice questions
- Some multipart questions that may ask you to
  - Explain a concept or a technique, compare or contrast two or more techniques;
  - Execute a data mining technique "by hand" on a tiny dataset;
  - Evaluate some aspect of a technique;
  - Derive or explain a mathematical result;
  - Interpret program output.

Sample question types: general

- Describe one advantage and one disadvantage of Technique A compared to Technique B for a given scenario.
- Suggest n different Data Mining techniques which could be applied to achieve Task X for scenario Y.
- Explain the limitations of evaluating model performance using the same data which have been used to fit the model.
- Provide examples of issues that can arise during data cleaning.
- Prove a mathematical result or explain a mathematical concept.

Sample question types: specific

- Use computer generated output to classify or make a numerical prediction for given instance.
- Construct and/or interpret confusion matrices, error rates, ROC charts or other evaluation tools.
- Explain and/or apply concepts underlying specific Data Mining tools, *e.g.*
  Entropy for decision trees, P-values for regression, linear separation and margin for Support Vector Machines, stress and goodness-of-fit for Multidimensional Scaling.

Study recommendations

- I strongly recommend studying homework questions: imagine that the R output were provided for you; interpret and discuss it.
- Strengths and weaknesses of many of the methods are discussed in the lecture notes.
- All required mathematical concepts are given in the lecture notes.

Exam technique

- Provide reasons and explanations for your answers; sometimes there is no definite right or wrong answer for a data mining problem.
- Avoid simply quoting chunks of lecture notes, this does not demonstrate that you have understood anything.
- Relate your answers to the application.
- Common sense and adaptability are important qualties of a data miner; be prepared to use these skills in the exam.

Visualisation Techniques (Week 2)

- Generally depend on the types of variables being visualised

Quantitative: One: Detailed / big n: histogram, density plot
Compact / small n: boxplot, box-percentile plot
Two: Scatterplot **and interpreting it**
Many: Scatterplot matrix, parallel coordinate plot (if small n)

Categorical: One: barplot presentation is needed)
Two: parallel barplot (for relationships), stacked barplot (for proportions)
Two or more: mosaic plots

MDS (Week 10)

- Takes a matrix of *distances* among data points.
- Computes coordinates on a lower dimension reproduce these distances.
- Can be used to visualise patterns in data, find unusual observations, etc.

General classifier assessment (Week 7)

- Given the true values of the outcome variable and the classifier's predictions or guesses, we can construct a *confusion matrix*:

		Prediction		
		A	B	C
Truth	A			
	B			
	C			

  - *Accuracy* of a classifier is what fraction of the the data is predicted correctly (i.e., on the diagonal).
  - We can look at relative frequencies of off-diagonal cells to see which groups get "confused" most often and how.
  - Know how to construct the confusion matrix from the `R` output of classifiers (e.g., from an `rpart` output).

Binary classification terms

Study these by heart:

True positive rate (TPR): (a.k.a. *recall*, *sensitivity*) proportion of correctly classified instances within the special category: $\frac{TP}{TP+FN}$.

False positive rate (FPR): proportion of incorrectly classified instances within the "negative" category: $\frac{FP}{FP+TN}$.

Precision: (a.k.a. *positive predictive value*) proportion of positive classifications that actually are in the special category: $\frac{TP}{TP+FP}$.

F-Measure: (a.k.a. *F1 score*) harmonic mean of precision and sensitivity: $\frac{2TP}{2TP+FP+FN}$.

ROC Charts

- Standard way of classifying an instance using predicted probabilities: classify as belonging to the most likely class, ($p \geq 0.5$ in binary case).
- However the cut-off doesn't have to be 0.5.
- To construct ROC chart, first sort instances according to *confidence*, *i.e.* predicted probability of being positive.
- Then use each observed confidence as the cut-off, and plot resulting true positive and false positive rates as steps on the chart.
- Ideally, chart should rise steeply on the left.

Support Vector Machines (Week 7)

- Take classes $y_i = -1$ or $+1$ and predictor vectors $\boldsymbol{x}_i$.
- Finds an optimal hyperplane of separation between the different y_is.
- Through a "kernel trick", the optimisation problem can be rewritten to consider more complex separation.
- Can be extended to more than two classes.
- Show an understanding of tuning parameters that allow one to not over-fit or under-fit the data.
- Show an understanding of the maths behind the optimisation problem of a linear SVM (but not the dual problem).

Decision Trees (Week 8)

- Make sure you know how to interpret the output from `rpart` and `ctree`, and also draw and understand the trees.
- Know how to calculate the information gain from a split (decision) from classifier output.
- Understand how random forests are constructed from decision trees.

The Regression Problem (Week 9)

- *Regression* (or numeric prediction) is the task of learning a target function f which maps each attribute set $\boldsymbol{x}$ to a numeric output (response) variable y.
- Consider a data set of n observations:

$$\{(\boldsymbol{x}_i, y_i), i = 1, 2, \ldots, n\}.$$

  Usually $\boldsymbol{x}_i$ consists of multiple attributes.
- Let $\hat{y}_i = f(\boldsymbol{x}_i)$ denote the predicted (fitted) value for observation i, e.g., in a linear model $\hat{y}_i = \hat{\boldsymbol{\beta}}\boldsymbol{x}_i$.

Performance Measures

Learn and understand the following:

Mean Squared Error: $\text{MSE} = \frac{\text{SSE}}{n-p-1}$, where $\text{SSE} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $p =$number of predictors.

Mean Absolute Error: $\text{MAE} = \frac{1}{n-p-1}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

Coefficient of determination (R^2): $1 - \frac{\text{SSE}}{\text{SST}}$, where $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$

- Generally easier to interpret than MSE.
- Always increases (or doesn't decrease) with more predictors.

Adjusted R^2: $R^2_{\text{adj}} = R^2 - (1 - R^2)\frac{p}{n-p-1} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$

Linear Regression

$$\hat{y}_i = \boldsymbol{x}_i\boldsymbol{\beta} \equiv \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p}$$

- $x_{i,k}$ =kth predictor of ith observation.
- Can give indication of statistical significance for each predictor in the model.
- "Linear" means linear in βs, not xs:
  - x can be categorical via dummy variables.
  - Transform xs for better fit
  - Add x^2, x^3, etc. to model curves
  - Transforming y is also possible, changing interpretation.
  - Interactions between different xs can be added.
- Understand the maths behind obtaining the least squares estimates $\hat{\boldsymbol{\beta}}$ of the linear model.

Automatic Model Selection

Stepwise regression to try adding and removing predictors from the model to see if they improve a criterion.

All subsets regression to try to fit all possible combinations of predictors.

- Criteria include adjusted R^2, as well as several others (AIC, BIC, etc.) that work for a bigger variety of statistical models.

Logistic Regression

- Regression for binary outcomes: used in classification, but also inference.
- Models

$$\Pr(Y_i = 1) = \text{squash}(\beta_0 + \beta_1 x_{i,1} + \cdots + \beta_i x_{i,p})$$

  where $\text{squash}(x) = 1/(1 + e^{-x})$
- Similar considerations to linear regression.

Regression Trees

- Same as classification, but instead of predicting class probabilities, predict mean outcome.
- Branches to reduce variation within-leaf.
- A *model tree* is a variation involving the fitting of linear regression models at each leaf.

Probabilistic classification (Week 10)

- Given categorical variable Y and predictors $\boldsymbol{X}$, we want to estimate $P(Y = y|\boldsymbol{X})$ for different possible values of y.

Bayes's rule (for events): $P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|B^{\complement})P(B^{\complement})}$

Bayes's rule (for discrete variables):

$$P(Y = y \mid X = x) = \frac{P(X{=}x|Y{=}y)P(Y{=}y)}{\sum_{y' \in \mathcal{Y}} P(X{=}x|Y{=}y')P(Y{=}y')}$$

Conditional independence: A is *conditionally independent* of C given B if $P(A \cap C|B) = P(A|B)P(C|B)$; equivalently, $P(A|B \cap C) = P(A|B)$.

Bayes classifier: $\hat{y} = \arg\max_y P(y|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y)P(y)}{\sum_{y' \in \mathcal{Y}} P(\boldsymbol{x}|y')P(y')}$: which value of y has the highest probability given $\boldsymbol{x}$?

Naive Bayes

A Bayes classifier that assumes elements of $\boldsymbol{X}$ are independent given $\boldsymbol{Y}$.

1. Estimate $P(y)$ for each y from the data.
2. Estimate $P(x_i|y)$ (distribution of element of $\boldsymbol{x}$, x_i, for each y).
3. "Update" the probability of y using $P(x_i|y)$:

$$P(y|\boldsymbol{x}) \approx \frac{P(y)\prod_{i=1}^{d} P(x_i|y)}{\sum_{y' \in \mathcal{Y}} P(y')\prod_{i=1}^{d} P(x_i|y')}.$$

- Quantitative x_is accommodated by either a normal distribution or by discretisation.
- Know how to compute $P(y|\boldsymbol{x})$ from the individual probabilities.