Xi'an Jiaotong-Liverpool University
西交利物浦大學

Department of Computer Science and Software Engineering

Semester 1
2017-18

CSE315 Machine Learning

Final Exam

for Year 4 students

Examiner:

Instructions to Candidates:

1) This is a closed-book examination, which is to be written without books, tapes, or notes.

2) Total time allowed: Two hours.

3) Total marks available: 100 (worth 70% of the overall module assessment).

4) Answer ALL questions in the booklet provided.

5) It is not necessary to copy the questions in the answer booklet.

6) All materials must be returned to the invigilator upon completion of the exam. Failure to do so will be deemed as academic misconduct and will be dealt with according to the University's policy.

THIS PAPER IS NOT TO BE REMOVED FROM THE EXAM ROOM.

Question 1. (15 marks) (Regression) Developing a model to predict permeability could save significant resources for a pharmaceutical company, while at the same time more rapidly identifying molecules that have a sufficient permeability to become a drug:
Start R and use these commands to load the data:

```
> library(AppliedPredictiveModeling)
> data(permeability)
```

The matrix *fingerprints* contains the 1,107 binary molecular predictors for the 165 compounds, while *permeability* contains permeability response.

(a) The fingerprint predictors indicate the presence or absence of substructures of a molecule and are often sparse meaning that relatively few of the molecules contain each substructure. Filter out the predictors that have low frequencies using the nearZeroVar function from the caret package.
(b) Split the data (*fingerprints and permeability*) into a training and a test set, pre-process the data (i.e. centering and scaling), and tune a PLS model (with parameter tune length of 40) by 5 times repeated cross-validation. (Hint: Use the functions createDataPartition, train, and trainControl in package {caret})

Question 2. (15 marks) (SVM and MARS) For the Tecator data,

```
> library(caret)
> data(tecator)
```

There are two data matrix imported, *absorp is the* absorbance data for 215 samples. And *endpoints* report the percentages of water, fat and protein.

```
> fat = endpoints[,2] # what we want to predict is fat, so take the percentage of fat out of endpoints
> absorp = data.frame(absorp)
```

Please build Support Vector Machines (SVM) and Multivariate Adaptive Regression Splines (MARS) models after splitting and preprocessing the data. (Note: Set the parameter tune length for SVM to 20. For the MARS parameters, set them as .degree=1:2, .nprune=2:38.)

Question 3. (10 marks) For Markov Decision Processes (MDPs), please fill in the following five blanks by choosing appropriate terms from candidates 1 to 5.

Given the current state and decision, any probabilistic statement about the future of the process is completely unaffected by proving any information about the history of the process. Because, we are dealing with a ___A.___, and the ___B.___ depend on only current state and decision. Furthermore, the immediate expected cost also depends on only current state and decision.

The two most important methods for deriving optimal policies for Markov decision processes are C.____ and ___D.___ . Under the discounted cost criterion, the ___E.___ provides a quick way of approximating an optional policy.

Candidate 1. policy improvement algorithms

Candidate 2. method of successive approximations
Candidate 3. Markov chain
Candidate 4. new transition probabilities
Candidate 5. leaner programming

Question 4. (15 marks) Determine which is the best approach for each problem, and please explain your reasoning.

a. *supervised learning*
b. *unsupervised clustering*
c. *data query*

1. What is the average weekly salary of all female employees under forty years of age?
2. Determine the characteristics of a successful used car salesperson.
3. Do meaningful attribute relationships exist in a database containing information about credit card customers?
4. Determine whether a credit card transaction is valid or fraudulent.

Question 5. (13 marks) Please state how to use the genetic algorithm in feature selection task for machine learning. Clarify your statement in terms of *chromosome*, *fitness*, *the size of search space*, *initialization step*, *crossover operation*, *mutation operation*, and how to escape from local optima.

Question 6. (20 marks) There are several important characteristics for different predictive modeling techniques. For example,

1. Do they allow number of predictors greater than number of samples (n < p)?
2. What kind of pre-processing is needed before modeling (such as centering and scaling, filtering predictors with near-zero variance, highly-correlated variable(s) removal, or no need to do any pre-processing)?
3. Is the model interpretable?
4. Does it have the automatic feature selection mechanism embedded in the model building process?

Please compare above characteristics of the following four predictive modeling methodologies - *artificial neural networks (ANN)*, *k-nearest neighbours (kNN)*, *logistic regression*, and *bagged trees*. Create a summary table as follows and put your reasoning accordingly.

	Allow n < p	Pre-Processing	Interpretable	Automatic Feature Selection
ANN	Yes or No, and your reasoning ….	Centering & Scaling, NZV, Corr, or None, and your reasoning ….	Yes or No, and your reasoning ….	Yes or No, and your reasoning ….
kNN				
Logistic regression				
Bagged trees				

Question 7. (12 marks) State the process of CART algorithm for building a regression tree, starting with the entire data set *S*.The statement should include *Sum of Squared Errors (SSE)*, *recursive partition*, *tree growing step*, *cost-complexity tuning or pruning*.

~ THIS IS THE END OF THE EXAM PAPER ~