Assessment Information

SIT720: Machine Learning

Assessment 4: Machine Learning Project

This document supplies detailed information on assessment tasks for this unit.

Key information

- Due: Wednesday 26th September 2018 by 11.30pm (AEST)
- Weighting: 30%

Learning Outcomes

This assessment assesses the following Unit Learning Outcomes (ULO) and related Graduate Learning Outcomes (GLO):

Unit Learning Outcome (ULO)	Graduate Learning Outcome (GLO)
ULO 2: Perform linear regression, classification using logistic regression and linear Support Vector Machines.	**GLO 1:** Discipline knowledge and capabilities **GLO 5:** Problem solving
ULO 3: Perform non-linear classification using Support Vector Machines with kernels, Decision trees and Random forests.	**GLO 1:** Discipline knowledge and capabilities **GLO 5:** Problem solving
ULO 4: Understand the concept of maximum likelihood and Bayesian estimation.	**GLO 1:** Discipline knowledge and capabilities **GLO 5:** Problem solving
ULO 5: Construct a multi-layer neural network using backpropagation training algorithm.	**GLO 1:** Discipline knowledge and capabilities
ULO 6: Perform model selection and compute relevant evaluation measure for a given problem.	**GLO 2:** Communication

Purpose

This assessment is an extensive machine learning project. Students will be given a specific data set for analysis and will be required to develop and compare various classification techniques. Each student must demonstrate skills acquired in data representation, classification and evaluation.

Instructions

- the dataset consists of training and testing data in "train" and "test" folders. Use training data: X_train.txt labels: y_train.txt and testing data: X_test.txt labels: y_test.txt. There are other files that also come with the dataset and may be useful in understanding the dataset better.
- Please read the pdf file "dataset-paper.pdf" to answer Part 1.

Task A: *Understanding the data*

Answer the following questions briefly, after reading the paper

- What is the objective of the data collection process?
- What human activity types does this dataset have? How many subjects/people have performed these activities?
- How many instances are available in the training and test sets? How many features are used to represent each instance? Summarize the type of features extracted in 2-3 sentences.
- Describe briefly what machine learning model is used in this paper for activity recognition and how is it trained. How much is the maximum accuracy achieved?

(3 Marks)

Assessment Information

Task B: *K-Nearest Neighbor Classification*

Build a K-Nearest Neighbor classifier for this data.

- Let K take values from 1 to 50. For choosing the best K, use 10-fold cross-validation. Choose the best value of K based on model F1-score.
- Show a plot of cross-validation accuracy with respect to K.
- Using the best K value, evaluate the model performance on the supplied test set. Report the confusion matrix, multi-class averaged F1-score and accuracy.

(5 Marks)

Task C: *Multiclass Logistic Regression with Elastic Net*¶

Build an elastic-net regularized logistic regression classfier for this data.

- Elastic-net regularizer takes in 2 parameters: alpha and l1-ratio. Use the following values for alpha: 1e-4,3e-4,1e-3,3e-3, 1e-2,3e-2. Use the following values for l1-ratio: 0,0.15,0.5,0.7,1.
- Choose the best values of alpha and l1-ratio using 10-fold cross-validation, based on model F1-score.
- Draw a surface plot of F1-score with respect to alpha and l1-ratio values.
- Use the best value of alpha and l1-ratio to re-train the model on the training set and use it to predict the labels of the test set. Report the confusion matrix, multi-class averaged F1-score and accuracy.

(5 Marks)

Task D: *Support Vector Machine (RBF Kernel)*

Build a SVM (with RBF Kernel) classfier for this data.

- SVM with RBF takes 2 parameters: gamma (length scale of the RBF kernel) and C (the cost parameter). Use the following values for gamma: 1e-3, 1e-4. Use the following values for C: 1, 10, 100, 1000.
- Choose the best values of gamma and C using 10-fold cross-validation, based on model F1-score.
- Draw a surface plot of F1-score with respect to gamma and C.
- Use the best value of gamma and C to re-train the model on the training set and use it to predict the labels of the test set. Report the confusion matrix, multi-class averaged F1-score and accuracy.

(6 Marks)

Task E: *Random Forest*

Build a Random forest classifier for this data.

- Random forest uses two parameters: the tree-depth for each decision tree and the number of trees. Use the following values for the tree-depth: 300,500,600. Use the following values for the number of trees: 200,500,700.
- Choose the best values of tree-depth and number of trees using 10-fold cross-validation, based on model F1-score.
- Draw a surface plot of F1-score with respect to tree-depth and number of trees.
- Use the best value of tree-depth and number of trees to re-train the model on the training set and use it to predict the labels of the test set. Report the confusion matrix, multi-class averaged F1-score and accuracy.

(6 Marks)

Task F: *Discussion*

Write a brief discussion about which classification method achieved the best performance. Your thoughts on the reason behind this. What method performed the worst? Could you do better or worse than the results in the dataset paper? Do you have any suggestions to further improve model performances?

(5 Marks)

Assessment Information

Submission details

Deakin University has a strict standard on plagiarism as a part of Academic Integrity. To avoid any issues with plagiarism, students are strongly encouraged to run the similarity check with the *Turnitin* system, which is available through Unistart. A Similarity score MUST NOT exceed 39% in any case.
Late submission penalty is 5% per each 24 hours from 11.30pm, 26th of September. No marking on any submission after 5 days (24 hours X 5 days from 4pm 14th of September)
Be sure to downsize the photos in your report before your submission in order to have your file uploaded in time.

Extension requests

Requests for extensions should be made to Unit/Campus Chairs well in advance of the assessment due date. If you wish to seek an extension for an assignment, you will need to apply by email directly to Chandan Karmakar (karmakar@deakin.edu.au), as soon as you become aware that you will have difficulty in meeting the scheduled deadline, but at least 3 days before the due date. When you make your request, you must include appropriate documentation (medical certificate, death notice) and a copy of your draft assignment.
Conditions under which an extension will normally be approved include:

Medical To cover medical conditions of a serious nature, e.g. hospitalisation, serious injury or chronic illness. Note: Temporary minor ailments such as headaches, colds and minor gastric upsets are not serious medical conditions and are unlikely to be accepted. However, serious cases of these may be considered.

Compassionate e.g. death of close family member, significant family and relationship problems.

Hardship/Trauma e.g. sudden loss or gain of employment, severe disruption to domestic arrangements, victim of crime. Note: Misreading the timetable, exam anxiety or returning home will not be accepted as grounds for consideration.

Special consideration

You may be eligible for special consideration if circumstances beyond your control prevent you from undertaking or completing an assessment task at the scheduled time.
See the following link for advice on the application process:
http://www.deakin.edu.au/students/studying/assessment-and-results/special-consideration

Assessment feedback

The results with comments will be released within 15 business days from the due date.

Referencing

You must correctly use the Harvard method in this assessment. See the Deakin referencing guide.

Academic integrity, plagiarism and collusion

Plagiarism and collusion constitute extremely serious breaches of academic integrity. They are forms of cheating, and severe penalties are associated with them, including cancellation of marks for a specific assignment, for a specific unit or even exclusion from the course. If you are ever in doubt about how to properly use and cite a source of information refer to the referencing site above.

Plagiarism occurs when a student passes off as the student's own work, or copies without acknowledgement as to its authorship, the work of any other person or resubmits their own work from a previous assessment task.

Collusion occurs when a student obtains the agreement of another person for a fraudulent purpose, with the intent of obtaining an advantage in submitting an assignment or other work.

Work submitted may be reproduced and/or communicated by the university for the purpose of assuring academic integrity of submissions: https://www.deakin.edu.au/students/study-support/referencing/academic-integrity

Criteria	Excellent	Good	Fair	Unsatisfactory
Criteria 1: Understand the data by reading the provided research article and answer four questions asked in the Part 1 of the assignment.	Successfully answered all four questions. **3 Marks**	Successfully answered at least 2 questions and satisfactorily tried others. **2 Marks**	Successfully answered only one question. **1 Mark**	Failed to answer any question satisfactorily. **0 mark**
Criteria 2: Build a K-Nearest Neighbor classifier for this data: * Choose the best K value from given set of values and F1-score. * Show a plot of cross-validation accuracy with respect to K. * Using the best K value, evaluate the model performance using the supplied test set. * Report the results as requested in the assignment.	Successfully completed all four tasks. **5 Marks**	Successfully completed any two of the four tasks and satisfactorily tried one of the remaining tasks. **3 Marks**	Successfully completed only one of the four tasks and satisfactorily tried one of the remaining tasks. **2 Marks**	Failed to complete any given task. **0 Marks**
Criteria 3: * For L1 model, choose the best alpha value from the provide set of values. * For L2 model, choose the best lambda value from the provided set of values. * Evaluate the prediction performance on test data, report results and discuss if there is any sign of underfitting or overfitting with appropriate reasoning.	Successfully completed all three tasks. **5 Marks**	Successfully completed any two of the three tasks. **3 Marks**	Successfully completed any one of the three tasks. **2 Marks**	Failed to complete any given task. **0 Marks**
Criteria 4: Build a SVM (with RBF Kernel) classifier for this data. SVM with RBF takes 2 parameters: gamma (length scale of the RBF kernel) and C (the cost parameter). Use the following values for gamma: 1e-3, 1e-4. Use the following values for C: 1, 10, 100, 1000. Choose the best values of gamma and C using 10-fold cross-validation, based on model F1-score. Draw a surface plot of F1-score with respect to gamma and C. Use the best value of gamma and C to re-train the model on the training set and use it to predict the labels of the test set. Report the confusion matrix, multi-class averaged F1-score and accuracy.	Successfully completed all three tasks **6 Marks**	Successfully completed any two of the three tasks. **4 Marks**	Successfully completed any one of the three tasks. **3 Marks**	Failed to complete any given tasks **0 Marks**

SIT720 Machine Learning
Assessment Task 4: Individual problem-solving rubric

Criteria	Excellent	Good	Fair	Unsatisfactory
Criteria 5: Build a Random forest classfier for this data. (6 Marks) Random forest uses two parameters: the tree-depth for each decision tree and the number of trees. Use the following values for the tree-depth: 300,500,600. Use the following values for the number of trees: 200,500,700. Choose the best values of tree-depth and number of trees using 10-fold cross-validation, based on model F1-score. Draw a surface plot of F1-score with respect to tree-depth and number of trees. Use the best value of tree-depth and number of trees to re-train the model on the training set and use it to predict the labels of the test set. Report the confusion matrix, multi-class averaged F1-score and accuracy.	Successfully completed all three tasks. **6 Marks**	Successfully completed any two of the three tasks. **4 Marks**	Successfully completed any one of the three tasks. **3 Marks**	Failed to complete any given task. **0 Marks**
Criteria 6: Write a brief discussion about which classification method achieved the best performance. Your thoughts on the reason behind this. What method performed the worst? Could you do better or worse than the results in the dataset paper? Do you have any suggestions to further improve model performances?	Successfully completed all three tasks. **5 Marks**	Successfully completed any two of the three tasks. **3 Marks**	Successfully completed any one of the three tasks. **2 Marks**	Failed to complete any given task. **0 Marks**