THE UNIVERSITY OF HONG KONG

FACULTY OF ENGINEERING
DEPARTMENT OF COMPUTER SCIENCE

Final Examination

COMP7404 Computational Intelligence and Machine Learning
(For sub-classes A & C)
17 December 2020, 20:00 - 22:00

- This is an open book examination. Candidates are permitted to use any online/printed/written materials in the examination
- You may only use Python libraries that were shown in code examples on the lecture slides
- All answers must be your own
- Use of a calculator is allowed. Only approved calculators as announced by the Examinations Secretary can be used in this examination. It is candidates' responsibility to ensure that their calculator operates satisfactorily, and candidates must record the name and type of the calculator used on the front page of the examination script
- This examination consists of two parts: Part A and Part B. Answer ALL questions.
- Write your answer to Part A as a PDF and your answer to Part B as a *.ipynb file. Zip both files into a single *.zip file for submission to OLEX

Part	Question	Max. Mark
A	1	6
	2	12
	3	8
B	1	6
	2	6
	3	6
	4	4
	5	2
	Total	50

Part A

A.1:

A.1.1 (3 marks): Say we work on patient screening in a medical application (True means patient has a medical problem). Which of the following situation(s) would you like your classifier to have? Provide an explanation. (FP: false positive, FN: false negative, TP: true positive, TN: true negative, >>: much greater)

a) FP >> FN
b) FN >> FP
c) FN = FP × TP
d) TN >> FP
e) FN × TP >> FP × TN
f) All of the above

A.1.2 (3 marks): Suppose that there are a total of 44 machine learning related documents out of a total of 400 documents available in HKU library (HKUL). Now, suppose that Find@HKUL (the HKUL search engine) retrieves 10 documents after a user enters 'machine learning' as a query, of which 6 are machine learning related documents. Write down the recall and precision for this example.

A.2:

A.2.1 (3 marks): Say we have three models in an ensemble and got the following results on a binary classification task.

Name of Model	Class 1 Probability	Class -1 Probability
A	0.48	0.52
B	0.97	0.03
C	0.49	0.51

Apply soft voting and hard voting to this ensemble.

A.2.2 (3 marks): If AdaBoost underfits the training data, which hyperparameter should be changed and how?

A.2.3 (3 marks): For binary classification, which of the following statement(s) are true of AdaBoost with decision trees?

a) It can train multiple decision trees in parallel
b) It usually has lower bias than a single decision tree
c) It is popular because it usually works well even before any hyperparameter tuning

A.2.4 (3 marks): For binary classification, which of the following statement(s) are true of AdaBoost?

a) It can be applied to neural networks
b) The metalearner provides not just a classification, but also an estimate of the posterior probability
c) It uses the majority vote of learners to predict the class of a data point

A.3:

A.3.1 (2 mark for correct answer, -2 mark for incorrect answer): (True/False) Let $h_1(s)$ and $h_2(s)$ be admissible search heuristics. It follows that $h_3(s) = (h_1(s)+h_2(s))/2$ is also admissible.

A.3.2 (2 mark for correct answer, -2 mark for incorrect answer): (True/False) A CSP that is arc consistent can always be solved without any backtracking.

A.3.3 (2 mark for correct answer, -2 mark for incorrect answer): (True/False) The search heuristic $h(s) = 0$ is always consistent for every search problem.

A.3.4 (2 mark for correct answer, -2 mark for incorrect answer): (True/False) In value iteration, if a policy has converged, values must also have converged.

Part B

In Part B you are going to work on WhatsApp Message Spam Classification. A dataset of >5000 labeled messages is available at: i.cs.hku.hk/~sdirk/wa.zip. Each message is labeled as ham (legitimate) or spam.

The files contain one message per line. Each line is composed of two columns. The first column is the label (ham or spam) and the other column is the raw text of the message. Here are some examples:

ham	WHO ARE YOU SEEING?
ham	Its a part of checking IQ
spam	Did you hear about the new "Divorce Barbie"? It comes with all of Ken's stuff!
ham	No prob. I will send to your email.

Submit your solution to the following five questions as a Jupyter Notebook file (*.ipynb) consisting of your python code with descriptions and analysis. You must write your own code. We will perform strict plagiarism checks.

B.1 (6 marks): Load the dataset and split it into train and test sets of suitable sizes.

B.2 (6 marks): Select a suitable feature extractor and apply feature extraction to the dataset.

B.3 (6 marks): Apply a Decision Tree, Logistic Regression and KNN classifier. Determine suitable hyperparameters and predict generalization performance using cross validation. Use F1-score as a performance measure.

B.4 (4 marks): Create a forth classifier that is the ensemble of the three classifiers of B.3.

B.5 (2 marks): Apply your four classifiers on the test set.

END OF PAPER