ROYAL HOLLOWAY, UNIVERSITY OF LONDON
BSc/MSci EXAMINATION 2021

CS3920: MACHINE LEARNING

CS3920R: MACHINE LEARNING – for FIRSTSIT/RESIT CANDIDATES

Time allowed: **TWO hours**

Please answer **ALL** questions.

- Handwrite your answers on paper, and write your candidate number and the module number at the top of each page. Photograph/scan the pages and keep the original paper versions, as they may be required by the examiners.
- For each question you attempt, please clearly state the question number.
- Please DO NOT include your name or Student ID anywhere on your work.
- **Academic Misconduct:** We will check all assignments for academic misconduct. Suspected offences will be dealt with under the College's formal Academic Misconduct procedures. Please remember:
  - The work submitted is expected to be your own work and only your work. You may not ask for help from any source, or copy anyone else's work.
  - You must not give help to anyone else, including sending them any parts of the questions or copies of your solutions.
  - You must not discuss the questions or solutions with anyone else.
- **Submitting your work:**
  - Your document must be submitted through Moodle using the submission link in the module Moodle page. If possible please convert your document into a PDF document to make the submission process quicker and easier.
  - Emailed submissions will not be accepted.
  - **You must complete your exam upload within 1 hour of the exam finish time.**

1. (a) What is meant by *feature engineering* in machine learning? [1 mark]

   (b) You are given a classification problem with one feature x and the following training set:

x	y
-2	A
-1	A
0	B
2	B
4	C
5	C

   As usual, y is the label. This is a multi-class classification problem with possible labels A, B, and C. The test samples are 0, 1, and -5. Find the 1-Nearest Neighbour prediction for each of the test samples. Use the standard Euclidean metric. If you have encountered any ties, discuss briefly your tie-breaking strategy. [5 marks]

   (c) Engineer an additional feature for this dataset, namely x^2. Therefore, your new training set still has 6 labelled samples in its training set and 3 unlabelled samples in its test set, but there are two features, x and x^2. Find the 1-Nearest Neighbour prediction for each of the test samples in the new dataset. [16 marks]

   (d) What is meant by a kernel in machine learning? [2 marks]

   (e) How can the distance between the images of two samples in the feature space be expressed via the corresponding kernel? [2 marks]

   (f) You are given the same training set as before,

x	y
-2	A
-1	A
0	B
2	B
4	C
5	C

   and only one test sample, 1. The learning problem is still multi-class classification with possible labels A, B, or C. Using kernelized Nearest Neighbours algorithm with kernel $K(x, x') = (x \cdot x')^2$, compute the 3-Nearest Neighbours prediction for the test sample. If applicable, describe your tie-breaking strategy. [10 marks]

2. (a) Suppose that, when using grid search with cross-validation to select the parameters `C` and `gamma` of the Support Vector Machine (SVM), you have obtained these results for the accuracy of the algorithm:

		gamma				
		0.25	0.5	1	2	4
C	0.25	0.90	0.92	0.95	0.92	0.91
	0.5	0.89	0.93	0.95	0.92	0.91
	1	0.89	0.93	0.95	0.92	0.91
	2	0.89	0.93	0.95	0.92	0.90
	4	0.89	0.93	0.95	0.92	0.90

(As usual, the accuracy is defined as 1 minus the error rate.) Is this a suitable grid for selecting the optimal values of the two parameters? Explain why. If it is not suitable, describe at least one way of improving it. [7 marks]

(b) Give an example of a grid that is too crude and thus does not allow an accurate estimate of the optimal values of the parameters `C` and `gamma` of the SVM. [7 marks]

(c) Give an example of a grid that clearly does not cover the optimal values of the parameters `C` and `gamma` of the SVM. Briefly explain why your example achieves its goal. [7 marks]

3. (a) What do you regard as the main advantage of inductive conformal prediction as compared with cross-conformal prediction? [1 mark]

   (b) What do you regard as the main advantage of cross-conformal prediction as compared with inductive conformal prediction? [1 mark]

   (c) What is the difference between conformity measures and inductive conformity measures? [1 mark]

   (d) In this question you should apply inductive conformal prediction to the following regression problem, with the training set split into two parts, training set proper and calibration set:

   - The training set proper consists of the following 5 labelled samples:
     - sample $(0, 0)$ is labelled as 0
     - sample $(0, 1)$ is labelled as 1
     - sample $(2, 0)$ is labelled as 2
     - sample $(2, 1)$ is labelled as 3
     - sample $(2, 2)$ is labelled as 4
   - The calibration set consists of the following 3 labelled samples:
     - sample $(1, 0)$ is labelled as 1
     - sample $(1, 1)$ is labelled as 2
     - sample $(1, 2)$ is labelled as 3

   Use the inductive nonconformity measure $\alpha = |y - \hat{y}|$, where y is the true label and $\hat{y}$ is the 2-Nearest Neighbours prediction computed from the training set proper. Find the prediction set at significance level $\epsilon = 30\%$ for the test sample $(4, 2)$. If relevant, explain your strategy for tie-breaking. [17 marks]

(e) In this question you should apply cross-conformal prediction to the following training set, split into two folds:

- the positive samples are $(0, 1)$, $(-1, 1)$, and $(-1, 0)$;
- the negative samples are $(-1, -1)$, $(0, -1)$, and $(1, 0)$;
- fold 1 consists of $(0, 1)$, $(-1, 1)$, and $(0, -1)$;
- fold 2 consists of $(-1, 0)$, $(-1, -1)$, and $(1, 0)$.

The test sample is $(-1, 2)$. As your inductive conformity measure, take the distance to the nearest neighbour of the opposite class.

- Draw the training and test sets on paper. Before performing any calculations, guess what point prediction will be output by the cross-conformal predictor. Give a brief explanation. [2 marks]
- Compute the two p-values for the test sample. [18 marks]
- Compute the point prediction for the test sample, its confidence, and credibility. [3 marks]

Selected formulas and `scikit-learn` keywords that may (or may not) be useful

- The optimization problems solved by Ridge Regression and the Lasso are

$$\text{RSS} + \alpha \sum_{j=0}^{p-1} w[j]^2 \to \min \quad \text{and} \quad \text{RSS} + \alpha \sum_{j=0}^{p-1} |w[j]| \to \min .$$

- The inductive conformal predictor based on the conformity measure $|y - \hat{y}|$: the prediction set is

$$\left[\hat{y}^* - \alpha_{(k)}, \hat{y}^* + \alpha_{(k)}\right], \quad \text{where } k = \lceil (1-\epsilon)(m+1) \rceil .$$

- Polynomial kernel: $K(x, x') = (1 + x \cdot x')^d$.
- Radial kernel: $K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right)$.
- Soft margin classifier: $\|w\|^2 + C \sum_{i=1}^{n} \zeta_i \to \min$ subject to

$$y_i (w \cdot x_i + b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0, \quad \text{where } i = 1, \ldots, n.$$

- The p-value in conformal prediction is

$$p(y) := \frac{\#\{i = 1, \ldots, n+1 \mid \alpha_i^y \leq \alpha_{n+1}^y\}}{n+1}$$

  and the p-value in cross-conformal prediction is

$$p(y) := \frac{\sum_{k=1}^{K} \#\{i \in S_k \mid \alpha_{i,k} \leq \alpha_k^y\} + 1}{n+1},$$

  where α are conformity scores.

- Some important classes in `scikit-learn`: `KNeighborsClassifier`, `LinearRegression`, `Ridge`, `Lasso`, `SVC`, `GridSearchCV`. Important methods for them: `fit`, `predict`, `score`.
- Important scalers and normalizer in `scikit-learn`: `StandardScaler`, `MinMaxScaler`, `RobustScaler`, `Normalizer`. Important methods for them: `fit`, `transform`, `fit_transform`.

END