UNIVERSITY OF LONDON

BSc/MSc EXAMINATION 2022

For Internal Students of
Royal Holloway

DO NOT TURN OVER UNTIL TOLD TO BEGIN

CS3920/CS5920: Machine Learning
CS3920R/CS5920R: Machine Learning – for FIRSTSIT/RESIT CANDIDATES

Time Allowed: **TWO hours**

Please answer **ALL** questions

CS calculators are permitted

1. (a) Give two examples of practical problems that can be solved using machine learning. [4 marks]

   (b) What is meant by *overfitting* in machine learning? What is meant by *underfitting*? In your answer, make sure to define and explain the role of generalization accuracy. [7 marks]

   (c) Describe the method of cross-validation for evaluating generalization accuracy. It is sufficient to consider only the case of classification. [8 marks]

   (d) Consider the following code in `scikit-learn`:

```
from sklearn.model_selection import cross_val_score
from sklearn.datasets import load_iris
from sklearn.neighbors import KNeighborsClassifier
iris = load_iris()
knn = KNeighborsClassifier(n_neighbors=1)
cross_val_score(knn, iris.data, iris.target, cv=3)

array([0.962, 0.971, 1.000 ])
```

   Explain what the code is doing. Compute the overall estimate of the generalization accuracy. [7 marks]

   (e) Explain why the cross-validation procedure leads to a downward bias in the estimate of the generalization performance. [2 marks]

   (f) Consider a classifier whose learning curve is summarized in the following table.

size of the training set	accuracy
25,000	0.70
30,000	0.80
35,000	0.85
40,000	0.90
45,000	0.95
50,000	0.98
55,000	0.99

   What is the downward bias of 10-fold cross-validation for a training set of size 50,000? What is the downward bias of 4-fold cross-validation for a training set of size 40,000? [10 marks]

2. (a) Define the Lasso approach to linear regression, making sure to define the machine-learning or statistical concepts used in your definition (such as residual sum of squares, if relevant) and to explain how predictions are made. [7 marks]

   (b) Explain the role of the parameter α for the Lasso. What happens when $\alpha = 0$? When $\alpha \to \infty$? [8 marks]

   (c) Applying the Lasso to a training set, we have found the model $\hat{y} = x[0] - x[1] + 5$ (where $x[0]$ stands for the first feature and $x[1]$ for the second feature of x). Compute the test R^2 of this model on the following test set of size 2:

   - sample $x_1^* = (0, 1)$ labelled as $y_1^* = 2$,
   - sample $x_2^* = (2, 1)$ labelled as $y_2^* = 6$.

   [10 marks]

3. (a) In the context of conformal prediction, what is the smallest possible p-value for a training set of size 10? [3 marks]

   (b) Consider the following training set in a binary classification problem:

   - positive samples: 1, 3, and 4;
   - negative samples: 6, 7, and 10.

   Therefore, each unlabelled sample consists of only one feature. We are also given two test samples, 2 and 5. Using the conformity measure defined as the distance to the nearest sample of the opposite class, perform the following calculations for each of the two test samples.

   - Compute the p-values for postulated labels 0 and 1.
   - Compute the point prediction, confidence, and credibility (if needed, briefly explain your tie-breaking strategy). Do your results agree with your prior expectations? Why or why not?
   - Suppose the true label is 1. Compute the average false p-value for this test sample.

   [34 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

VV