CS3420

THE UNIVERSITY OF WARWICK

Third Year Examinations: Summer 2019

Machine Learning

Time allowed: 2 hours.

Answer **FOUR** questions, **TWO** questions from Section A and **TWO** questions from Section B.

Use one answer book for your Section A answers and a separate answer book for your Section B answers.

Read carefully the instructions on the answer book and make sure that the particulars required are entered on **each** answer book.

Approved calculators are allowed.

Continued

Section A Answer **TWO** questions

---

1. (a) Explain the concept of k-fold cross-validation with the aid of a diagram. Describe the case of Leave-One-Out-Cross-Validation (LOOCV). List one advantage and one disadvantage of LOOCV. [3]

   (b) You are building an image classifier that assigns an image of 640×640 resolution into one of 5 classes. You have being given a balanced dataset $\mathcal{D} = \{\mathbf{x}_n, t_n\}_{n=1}^{N}$ with $N = 1000$, $t_n \in \{1, 2, 3, 4, 5\}$ and $\mathbf{x}_n \in \Re^D$ where $D = 640 \times 640 = 409,600$.

   i) Describe two data augmentation techniques you would use with this dataset and explain why they would be appropriate? [3]

   ii) What is an appropriate likelihood or activation function for this problem? [3]

   iii) Explain the curse of dimensionality within the context of this problem and dataset [3]

   (c) Derive the optimal least squares parameter value, $\widehat{\mathbf{w}}_{\text{OLS}}$, for the average training loss

   $$\mathcal{L} = \frac{1}{N} \sum_{n=1}^{N} (t_n - \mathbf{w}^{\text{T}} \mathbf{x}_n)^2$$

   [6]

   (d) In the context of Bayesian Linear Regression show that if the prior over the parameters $p(\mathbf{w}) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0)$ where $\boldsymbol{\Sigma}_0 = \frac{\sigma^2}{N\lambda}\mathbf{I}$, then the *maximum-a-posteriori* solution is equivalent to the Ridge regression estimate $\widehat{\mathbf{w}}_{\text{Ridge}}$. [7]

Continued

2. (a) Explain what is the i.i.d. assumption in the context of supervised learning. Give an example where that assumption is violated and explain why. [6]

   (b) What is overfitting and how can we detect overfitting of our models? [4]

   (c) Give the name and mathematical description of 3 different activation functions. What is the problem of *vanishing gradients* and how can it be addressed? [6]

   (d) If the unnormalised univariate density of the random variable $y \in \Re^+$ given by $p(y|a,b) \propto y^{a-1}\exp(-by)$ has its maximum at $\hat{y} = \frac{a-1}{b}$, derive its Laplace approximation $q(y|a,b)$. Note that $a, b \in \Re$ are the two parameters of the density. When is the Laplace approximation optimal? [9]

---

3. (a) i. Give the likelihood function for the Gaussian Mixture Model, define all the terms and explain the main underlying assumptions of the model. [4]

ii. What is regularisation and what are its benefits in supervised learning? Give the regulariser for LASSO and Ridge Regression. [4]

(b) i. List 3 limitations of Principal Component Analysis (PCA) [3]

ii. List 3 limitations of k-means clustering [3]

iii. Should you standardise your data before applying PCA? Explain your reasoning. [2]

(c) Give the k-means algorithm [3]

(d) Describe PCA and give its mathematical relation to an eigenvalue decomposition problem. When is it useful to apply PCA and how would you choose the number of principal components to use? [6]

Continued

Section B Answer **TWO** questions

4 Consider the following dataset in Figure 1 and a k-NN classifier.

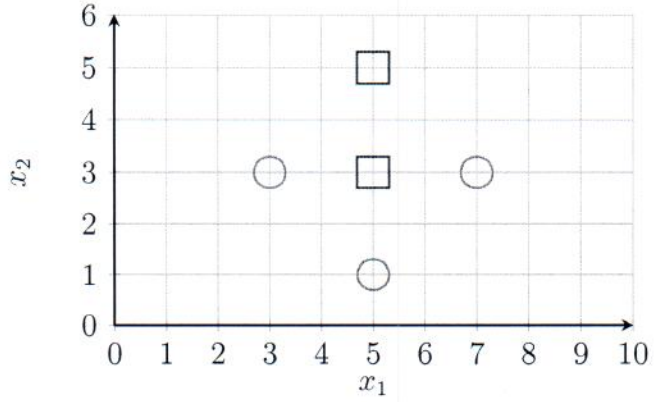

Figure 1: 2-D dataset with 2 classes (○, □) for k-NN

(a) Sketch the decision boundary for 1-NN and shade all the area in the figure based on class assignment when using:

i) the Euclidean distance. [2]

ii) the Manhattan distance. [4]

(b) What is the *training error* for 2-NN with the Euclidean distance? In the case of ties, assign it to class □. In the case of equidistant neighbours, use the majority rule. [4]

(c) What is the *leave-one-out-cross-validation* (LOOCV) error for $k = 1$ (1-NN) for the following distances? In the case of ties, assign it to class □. In the case of equidistant neighbours, use the majority rule.

i) the Euclidean distance. [2]

ii) the Manhattan distance. [4]

(d) Give the expected error on an unseen sample under the squared error loss as a function of the *Bias* and *Variance* (Bias-Variance decomposition). Formally define and explain each resulting term of the decomposition. [7]

(e) "In k-NN the resulting fitted function is always piece-wise linear." *True or False?* [2]

Continued

5 Consider the dataset in Table 1 and an ID3 classifier. The classification target is variable T and you are given three binary attributes A, B, C for the following 6 data points.

A	B	C	T
0	0	1	0
1	0	0	0
0	1	0	1
0	1	1	1
1	1	0	1
1	1	1	0

Table 1: The dataset for ID3. Three attributes (A, B, C), six data points and the classification target T.

(a) Compute the *Information Gain* (IG) and the *drop in Gini impurity* for all attributes A, B, C [6]

(b) Which attribute would ID3 select first based on drop in Gini impurity and why? [1]

(c) Write down the entire decision tree constructed by ID3, without any pruning. Show all your calculations. [8]

(d) Use the induced ID3 to classify the new data point $\mathbf{x}_{\text{new}} = [1\ 0\ 1]$ [2]

(e) Give 2 reasons why ID3 is not guaranteed to always find the optimal decision tree [2]

(f) What are the differences between Bagged Decision Trees and the Random Forest? [6]

If two attributes have the same IG, choose the first one in *alphabetical order*.

Hints: i) The drop in Impurity for an impurity measure $i(N)$ is given as: $\Delta i(S, A) = i(S) - \sum_b \frac{|S_{u_b}|}{|S|} i(S_{u_b})$. Information Gain uses Entropy as the impurity measure. ii) Remember that $\lim_{x \to 0}(x \log x) = 0$.

---

6 The probabilistic output of a binary classifier on a test set of 10 observations is depicted in the left column of Table 2, together with the true labels in the right column.

$P(t^* = +1 \mid \mathbf{x}^*, \dots)$	True Class
0.9	+1
0.8	+1
0.6	+1
0.5	+1
0.3	+1
0.8	-1
0.5	-1
0.3	-1
0.3	-1
0.2	-1

Table 2: Left: Class probabilities from the classifier for class $+1$. Right: True class labels.

(a) Compute the Receiver Operating Characteristic (ROC) curve at the following decision thresholds: $\{0.1, 0.4, 0.7\}$. [6]

(b) What is the range of values the decision threshold can take in order for the classifier to get: i) 2 False Negatives on this test set? ii) 2 False Positives on this test set? [4]

(c) The probabilistic classifier that produced this output is a Logistic Regression.

i) Derive the logistic function starting from the log-odds definition. [4]

ii) Define the Likelihood and the Maximum Likelihood Estimator for this model. [3]

(d) Starting again from log-odds ratios derive the multi-class generalisation of the logistic function for C classes, called the *softmax*: $P(t_n = j|\mathbf{x}_n, \mathbf{w}) = \frac{\exp\{\mathbf{w}_j^{\mathrm{T}}\mathbf{x}_n\}}{\sum_{c=1}^{C}\exp\{\mathbf{w}_c^{\mathrm{T}}\mathbf{x}_n\}}$ [5]

(e) Assume now that the probabilistic classifier that produced this output is a Naive Bayes model with a Gaussian class-conditional likelihood. Give the maximum likelihood estimates for the parameters of the model. [3]

Hints: Specificity = TN/ (TN + FP), Sensitivity = TP / (TP + FN)

End