University of Bath

DEPARTMENT OF COMPUTER SCIENCE
EXAMINATION

CM50264: MACHINE LEARNING 1 — Solution

Assessment Available from: Friday, 28th January 2022, 9:30am
Latest Submission Time: Friday, 28th January 2022, 12:30am

Please read the Guidance for Students (`https://www.bath.ac.uk/topics/exams-and-assessments`) before attempting this exam. The Guidance contains information about submitting your exam attempt.

This is an open book exam. You may refer to your own course and revision notes and look up information in offline or online resources, for example textbooks or online journals.

This exam starts at: 09:30 on 28th January 2022.

This exam is designed to take approximately 2 hours to complete.

You will have an additional 60 minutes of submission time for checking your work, collating your answers and uploading files. You are advised to allow sufficient time for minor technical issues when submitting your work.

The exam will close at the end of the submission time, after which you will not be able to submit an attempt.

Which questions should be answered: All of them.

Filenames: If you are required to upload a file as part of your exam attempt, to maintain anonymity please use the following naming convention for your file: `CandidateNumberUnitCodeQuestionNumber.pdf` (e.g. `01234AR10001Q2a.pdf`). If the exam only requires one file to be submitted, you do not need to include the question number(s).

Additional materials needed to complete the assessment: None.

Further instructions: None.

Academic Integrity for Remote Exams

When you registered as a student you agreed to abide by the University's regulations and rules, and agreed that you would access and read your programme handbook. These documents contain references to, and penalties for, unfair practices such as collusion, plagiarism, fabrication or falsification. The University's Quality Assurance Code of Practice, QA53 Examination and Assessment Offences (`https://www.bath.ac.uk/publications/qa53-examination-and-assessment-offences/`), sets out the consequences of committing an offence and the penalties that might be applied.

By submitting your exam as instructed, you confirm that:

1. You have not impersonated, or allowed yourself to be impersonated by, any person for the purposes of this assessment.
2. This assessment is your original work and no part of it has been copied from any other source except where due acknowledgement is made.
3. You have not previously submitted this work for any other unit/course.
4. You give permission for your assessment response to be reproduced, communicated, compared and archived for plagiarism detection, benchmarking or educational purposes.
5. You understand that plagiarism is the presentation of the work, idea or creation of another person or organisation as though it is your own. It is a form of cheating and is a very serious academic offence that may lead to disciplinary action.
6. You understand that this assessment is undertaken without invigilation, and that you have not communicated with and will not communicate with anyone concerning this assessment before the deadline for submission unless it is expressly permitted by the assessment instructions.
7. No part of this assessment has been produced for, or communicated to, you by any other person, unless it is expressly permitted by the assessment instructions.

If you have any questions about the exam you should contact the exams helpline. Information and contact details can be found on our help and advice webpage (`https://www.bath.ac.uk/guides/exams-and-assessments-get-help-and-advice`)

1. **True or False?**

   (1) Overfitting issues are less likely with a larger feature space [1]

   (2) Increasing the regularisation parameter λ in lasso regression leads to sparser regression coefficients [1]

   (3) For a fixed size of the training and test set, increasing the complexity of the model always leads to reduction of the test error [1]

   (4) Overfitting exists in supervised learning, but not in unsupervised learning [1]

   (5) Cross-validation may reduce overfitting [1]

   (6) The back-propagation algorithm learns a globally optimal neural network with hidden layers [1]

   (7) A trained machine learning model that obtains 90% accuracy on a test set will achieve the same (90%) or higher accuracy on another test set [1]

   (8) Validation data will be used in model training [1]

Solution:

(1) False. The more the number of features, the higher the complexity of the model and hence greater its ability to overfit the training data.

(2) True. Larger regularisation parameter penalises non-zero coefficients more, leading to sparser solution.

(3) False. Increasing complexity of model for a fixed training and test set leads to overfitting the training data and reduction in training error, but not test error.

(4) False. Unsupervised learning model can also overfit.

(5) True. Overfitting is avoided by correctly estimating performance for different hyperparameter combinations; cross-validation improves the accuracy of that estimate and hence may help avoid it.

(6) False. BP cannot guarantee to find the global optimal solution.

(7) False. There's no guarantee.

(8) True. It is used to fit the hyperparameters, which is part of training the model.

2. **Multiple choice questions**

Please record all valid answers for each question. Full marks will only be given for the correct set, half marks if there is a maximum of one addition/omission.

(1) Which approaches may reduce overfitting? [2]

(a) Increasing the quantity of training data
(b) Improving regularisation
(c) Increasing the complexity of the model
(d) Reducing the complexity of the model

(2) Which of the following are true about bagging? [2]

(a) In bagging a random sample of the input points is drawn with replacement
(b) Bagging is primarily about decreasing the bias of a model
(c) Bagging does not work for logistic regression because you get exactly the same decision boundary each time
(d) Bagging decision trees with one data point per leaf will not give a lower training error than an ordinary decision tree

(3) Which of the following statements are true? [2]

(a) The k-means algorithm (EM) has no guarantee it will converge to the global optimum
(b) Removing features will always improve the performance of a clustering algorithms such as k-means
(c) L1 regularisation is better than L2 for linear regression
(d) Machine learning works automatically, and there is no need for the user to manually set model hyperparameters

(4) Of the following statements which are true for a k-NN classifier? [2]

(a) k-NN does not require an explicit training step
(b) The decision boundary of k-NN is nonlinear
(c) Increasing k will smooth the decision boundary
(d) Classification accuracy improves as k is increased

Question 2 continues on next page ...

Question 2 continued ...

(5) Which of the following hyperparameters influence the overfitting/underfitting of an artificial neural network? (assume there are sufficient epochs for model training to have converged) [2]

(a) Number of neurons/nodes
(b) Number of layers
(c) Initialisation of node weights
(d) Batch size
(e) Learning rate

(6) Which of the following is true for Maximum Likelihood Estimates (MLE)? [2]

(a) There may not be a MLE
(b) There is always a MLE
(c) If a MLE exists, it may not be the only optimal solution
(d) If a MLE exists, it is the only optimal solution

(7) Of the following, which can always be used to construct a standard neural network? [2]

(a) k-NN
(b) Linear regression
(c) Logistic regression
(d) Hidden Markov model

Solution:

(1) a, b, d

(2) a, d

(3) a

(4) a, b, c

(5) a, b, d, e

(6) a, c

(7) b, c

3. **Random forests and a friend**

(a) When training a decision tree, if a feature is uncorrelated with the target variable will it be selected for splits? Explain your answer. [2]

(b) For optimising a split give one reason why you might choose Gini impurity rather than information gain? Give one reason why you might choose information gain instead of Gini impurity? [2]

(c) If you are evaluating a data point with a missing feature using a random forest what are your options for handling a split on that feature? [2]

(d) Would you use bagging if you have a simulator that can generate infinite training data? Explain your answer. [2]

Boosted trees are an alternative to random forests that replace bagging with boosting. Boosting sequentially trains trees so that each tree reduces the error of the current model (there is no randomness but depth is limited). The first decision tree, $f_0(\cdot)$, is trained normally, to minimise the error of

$$y = f_0(x).$$

The second tree, $f_1(\cdot)$, is then trained to reduce the error of the first tree,

$$\epsilon_1 = f_1(x)$$

where ϵ_1 is the error, $\epsilon_1 = y - f_0(x)$ for regression. The model now predicts

$$y' = f_0(x) + f_1(x).$$

This continues for an arbitrary number of trees.

(e) Will performance on the training set improve with each successive tree? What about the testing set? Why? [2]

(f) The above model has been given for regression. What would you do if given a classification problem? [2]

Solution:

(a) "Yes, but only due to spurious correlation." Will also accept: "No, because splitting on this feature will not result in any information gain." No explanation = no marks. Presuming Pearson correlation (i.e. linear) is not penalised, as it demonstrates an understanding of the above.

(b) Gini is faster than information gain due to not having to evaluate log. Information gain works for problems that are not classification.

(c) You can either go down both paths and return a weighted average or select a branch at random, with the probability proportional to the quantity of the training set that went down each branch. (Using the mean etc. or predicting the value with another ML model should technically also be included, but as the question is about RF I'll accept if only RF specific approaches are given)

(d) No. Bagging approximates having many unique data sets, but infinite data means we actually have that.

(e) Yes for training set, as by construction it always reduces error. No for testing set, as it will still suffer from overfitting. (to get marks the yes/no statements must be explicitly attached to the train/test set)

(f) Output a vector of (unnormalised) class membership probabilities from each tree, so the error is continuous and hence can be corrected with each additional tree. (there are many other answers; the point is that reducing the error at each step requires a continuous representation of class membership, not a discrete one — if someone has acknowledged that without fully fleshing out a solution they can expect one mark)

4. **Optimisation**

Consider random variables x as the input and y as the label for a machine learning model, and an unknown coefficient w to be estimated. We observe N samples $D = \{(x_1, y_1), \cdots, (x_n, y_n), \cdots, (x_N, y_N)\}$ where n is the sample index. The loss function for a batch of N samples can be written as:

$$L(w) = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{2}(y_n - wx_n)^2 + \frac{\lambda}{2}\|w\|^2$$

where λ is the regularisation coefficient and $L(w)$ is a regularised loss.

(a) Derive the standard batch gradient descent (with batch size N) result for $L(w)$ with respect to w. [3]

(b) Assume the general updating equation for w is defined as $w_{t+1} = w_t + \alpha p_t$, where t is the iteration index and α is a known step size, p_t denotes the direction from t to $t+1$. Write down the update equation of **stochastic** gradient descent with respect to w assuming a batch size 1. [3]

(c) Can we use Newton's method to solve this problem? Why? [3]

(d) If the sample size N increases, will we need to adjust λ? If yes, in which direction (increase/decrease) is λ more likely to go? [3]

Please demonstrate how you solve the problem. Correct answers without reasoning will be marked as 0.

Solution:

(a) Derivation with respect to w gives:

$$\begin{aligned}\frac{\partial}{\partial w}L &= \frac{1}{N}\sum_{n=1}^{N}(y_n - wx_n)(-x_n) + \lambda w \\ &= \frac{1}{N}\sum_{n=1}^{N}(wx_n - y_n)x_n + \lambda w.\end{aligned}$$

1 mark for correct mathematical starting point, 2 marks for deriving the correct result in either summation form or matrix form.

(b)

$$\begin{aligned}w_{t+1} &= w_t + \alpha\nabla L(w) \\ &= w_t - \alpha((w_t x_n - y_n)x_n + \lambda w_t) \\ &= (1 - \alpha\lambda)w_t - \alpha x_n(w_t x_n - y_n)\end{aligned}$$

1 mark for correct mathematical starting point, 2 marks for deriving the correct result. 1 mark deducted if using the summation for in the updating equation, as the question is asking for stochastic gradient descent update.

(c) In principle, yes, Newtons' method uses 2nd order derivative to find the direction. However, this is a simple univariate question (we only have one w) and its 2nd order derivative ($\frac{\partial^2 L}{\partial w^2}$) is a constant. 1 mark for yes or no, 2 marks for the correct explanations, either plain text explanations or differentiable proofs.

(d) If sample size N increases, there may be an increasing risk of underfitting. So we need to adjust the regularisation coefficient λ. Decrease λ will mitigate underfitting. 1 mark for yes or no, 1 mark for explanation, 1 mark for λ direction.

5. **Bayesian Regression**

Consider two random variables X and Y, where Y is the noisy observed data generated from X under Gaussian noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$, which is given in full as

$$P(\epsilon|0, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(0-\epsilon)^2}{2\sigma^2}\right).$$

The relationship between Y and X is

$$Y = \beta X + \epsilon,$$

where β is an unknown coefficient. We assume σ is known.

(a) Write down the full mathematical expression of the conditional probability $p(Y|X, \beta)$. [2]

Assume we are given N samples $D = \{(x_1, y_1), \cdots, (x_n, y_n), \cdots, (x_N, y_N)\}$ where n is the sample index.

(b) Write down the full mathematical expression of the objective function that maximum likelihood (ML) solves, for β. [2]

(c) Derive and write down the maximum likelihood estimate of β, referred to as β_*^{ML}. [3]

(d) Assume that the prior distribution over β (denoted as $p(\beta)$) is also Gaussian, $\beta \sim \mathcal{N}(0, \sigma_\beta^2)$. Write down the expression of the posterior distribution $p(\beta|Y, X)$ using Bayes' theorem. [3]

(e) If $\sigma_\beta \to \infty$, what is the increase/decrease trend of the error between the ML estimate β_*^{ML} and MAP estimate β_*^{MAP}? Why? [2]

Please demonstrate how you solve the problem when requested. Correct answers without reasoning will be marked as 0. For question (b) and (d), please use the given mathematical symbols only, don't invent any new random variables, vectors, or matrices.

Solution:

(a)

$$p(Y|X,\beta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y-\beta X)^2\right).$$

Only full and correct expression get 2 marks, otherwise 0 mark.

(b) The full expression of the ML solution is

$$\arg\max_{\beta} \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n-\beta x_n)^2\right).$$

Answers in the following forms also get full mark:

$$\arg\max_{\beta} \prod_n \exp\left(-\frac{1}{2\sigma^2}(y_n-\beta x_n)^2\right).$$

$$\arg\max_{\beta} -\frac{1}{2}\sum_n (y_n-\beta x_n)^2.$$

(c) The objective function now is

$$f(\beta) = \prod_n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(y_n-\beta x_n)^2\right)$$

Finding the optimal of the above objective function is equivalent to finding optimal for the below objective function by ignoring other irrelevant terms in differentiation:

$$f(\beta) = -\frac{1}{2}\sum_n (y_n-\beta x_n)^2$$

So now we solve:

$$\frac{\partial}{\partial\beta}\left(-\frac{1}{2}\sum_n (y_n-\beta x_n)^2\right) = 0$$

$$\sum_n (y_n-\beta x_n)(-x_n) = 0$$

$$\sum_n \beta x_n^2 - \sum_n x_n y_n = 0$$

$$\beta = \frac{\sum_n x_n y_n}{\sum_n x_n^2}$$

1 mark for writing a correct objectives, 1 mark for taking the correct derivative, 1 mark for getting the correct β result. If the answer is $\beta = \frac{\sum_n y_n}{\sum_n x_n}$, it is wrong, we cannot cancel out x_n.

(d) Using Bayes' theorem:

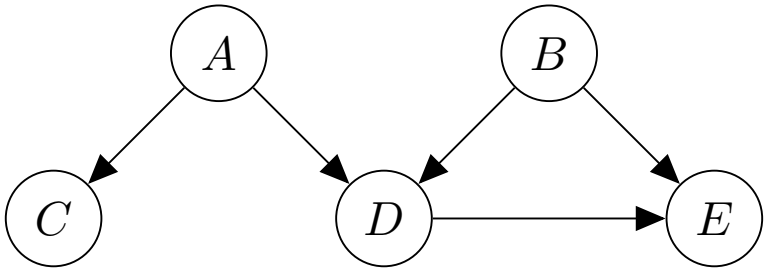
$$\begin{aligned} p(\beta|Y,X) &\propto p(Y|X,\beta)p(\beta|X) \\ &\propto p(Y|X,\beta)p(\beta) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(Y-\beta X)^2\right) \frac{1}{\sqrt{2\pi\sigma_\beta^2}} \exp\left(-\frac{\beta^2}{2\sigma_\beta^2}\right) \end{aligned}$$

1 mark for the correct Bayes' theorem expression, 1 mark for correct Gaussian mathematical expression, 1 mark for correct posterior expression.

(e) The error between two estimates will decrease. This is because σ_β increase will give a wider (or 'flatter') prior distribution, and in ML we actually assume a flat uniform prior for β. A wider prior in MAP will be closer to the ML prior assumption. 1 mark for correct answer, 1 mark for why.

6. **Bayes Networks**

Consider the Bayes network given below, where the only (conditional) independence assumptions between the random variables A to E are those enforced by the shape of the network:



(a) Write down the full joint probability of all random variables $A, \ldots, E$ in terms of the distributions implied by the Bayes network. [2]

(b) For this question only assume that all random variables $A, \ldots, E$ are binary. What is the minimum total number of elements required in probability tables to store the joint probability of variables in this Bayes network? Assume the tables need only to store the positive results for each binary variable. Show your working. [2]

(c) Based on the Bayes network, what can you say about dependencies between each of the following pairs of variables? Choose the type of dependency from {*none*, *conditionally independent given ...*, *independent*}

(1) C and D [1]
(2) B and C [1]
(3) A and E [1]
(4) D and E [1]

(d) Consider the Bayes network as above where the arrow between D and E is reversed, i.e. is now pointing from E to D. What is the new dependency type for each of the pairs of variables from the previous subquestion (c)?

(1) C and D [1]
(2) B and C [1]
(3) A and E [1]
(4) D and E [1]

Solution:

(a) Following the formula: $p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|parents(X_i))$, we have:

$$p(A, B, C, D, E) = p(A)p(B)p(C|A)p(D|A, B)p(E|B, D)$$

1 mark for correct equation, 1 mark for correct reasoning.

(b) From the formula for the full joint probability, we have

- $p(A)$ would need to store 1 element $p(A = true)$;
- Similarly to $p(A)$, $p(B)$ would need to store 1 element $p(B = true)$;
- $p(C|A)$ would need to store 2 elements: $p(C = true|A = true)$ and $p(C = true|A = false)$;
- $p(D|A, B)$ would need to store 4 elements for each of 4 combinations of values that binary A and B can take:

A	B	D
$true$	$true$	$p(D = true\|A, B)$
$true$	$false$	$p(D = true\|A, B)$
$false$	$true$	$p(D = true\|A, B)$
$false$	$false$	$p(D = true\|A, B)$

- Similarly to $p(D|A, B)$, $p(E|B, D)$ would need to store 4 elements for each of 4 combinations of values that binary B and D can take.

In total, we would need $1 + 1 + 2 + 4 + 4 = 12$ elements in probability tables to store the full joint probability. 2 marks for correct number of elements. 1 mark can be given if **one** type of the conditional probabilities was computed incorrectly.

(c) Recall two rules of determining conditional independence for any two variables in a Bayes network:

(i) Each node is conditionally independent of its non-descendants given its parents;

(ii) Each node is conditionally independent of all other nodes in the network given its Markov blanket: parents, children, children's parents.

(1) C and D are conditionally independent given A. This can be shown either by

using the rule (i) applied, for example, for variable C, or by:

$$p(C|A,D) = \frac{p(A,C,D)}{p(A,D)} = \frac{\sum_B \sum_E p(A,B,C,D,E)}{\sum_B \sum_C \sum_E p(A,B,C,D,E)}$$
$$= \frac{\sum_B \cancel{\sum_E} \cancel{p(A)} p(B) p(C|A) p(D|A,B) \cancel{p(E|B,D)}^{1}}{\sum_B \cancel{\sum_C} \cancel{\sum_E} \cancel{p(A)} p(B) \cancel{p(C|A)}^{1} p(D|A,B) \cancel{p(E|B,D)}^{1}}$$
$$= p(C|A) \frac{\sum_B p(B) \cancel{p(D|A,B)}^{1}}{\cancel{\sum_B p(B) p(D|A,B)}} = p(C|A)$$

(2) B and C are independent. This can be shown either by using rule (i) applied for variable B, i.e., variable B is conditionally independent of C (B-th non-descendant) given B-th parents. B does not have a parent, therefore, the independence between B and C is unconditional. This can also be shown by:

$$p(B,C) = \sum_A \sum_D \sum_E p(A,B,C,D,E)$$
$$= \sum_A \cancel{\sum_D} \cancel{\sum_E} p(A) p(B) p(C|A) \cancel{p(D|A,B)}^{1} \cancel{p(E|B,D)}^{1}$$
$$= p(B) \sum_A p(C,A) = p(B) p(C)$$

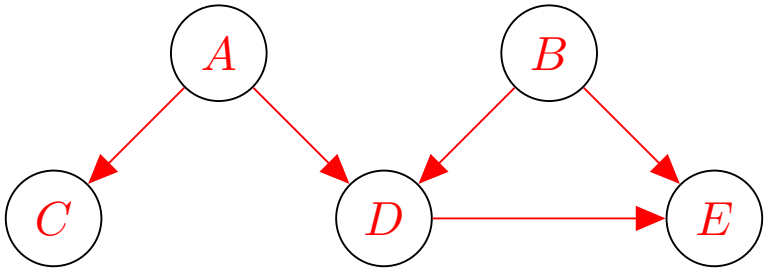
(3) A and E are conditionally independent given B and D. This can be shown either by using rule (ii) applied for variable E, or by:

$$p(A|B,D,E) = \frac{p(A,B,D,E)}{p(B,D,E)}$$
$$= \frac{p(A) \cancel{p(B)} p(D|A,B) \cancel{p(E|B,D)}}{\sum_A p(A) \cancel{p(B)} p(D|A,B) \cancel{p(E|B,D)}}$$
$$= \frac{p(A,D|B)}{p(D|B)} = p(A|B,D)$$

(4) Nothing can be said about the type of dependencies between D and E. Since D is a parent of E, generally speaking, D is not independent (conditionally or unconditionally) from E.

(for *conditionally independent* answers the variables required to achieve independence must be given to get the marks)

(d) The new Bayes network is:



The full joint probability for the new Bayes network is:

$$p(A, B, C, D, E) = p(A)p(B)p(C|A)p(D|A, B, E)p(E|B)$$

(1) C and D are still conditionally independent given A. The reasoning is the same as in subquestion (c).

(2) B and C are still independent. The reasoning is the same as in subquestion (c).

(3) A and E are now independent (unconditionally). This can be shown either by using rule (i) for variable A (since A does not have parents the rule would impose unconditional independence), or by:

$$\begin{aligned} p(A, E) &= \sum_B \sum_C \sum_D p(A, B, C, D, E) \\ &= \sum_B \sum_C \sum_D p(A)p(B)\cancel{p(C|A)}^{1}\cancel{p(D|A, B, E)}^{1}p(E|B) \\ &= p(A) \sum_B p(E, B) = p(A)p(E) \end{aligned}$$

(4) There is still nothing to be said about the type of dependency between D and E. The relationship between them has flipped and now E is a parent of D, but still generally speaking, D is not independent (conditionally or unconditionally) from E.

(for *conditionally independent* answers the variables required to achieve independence must be given to get the marks)