UNIVERSITY OF BRISTOL

January 2022 Examination Period

Department of Computer Science

3rd Year Examination for the Degrees of
Bachelor in Computer Science
Master of Engineering in Computer Science

COMS30033
Machine Learning

TIME ALLOWED:
2 Hours
plus 30 minutes to allow for collation and uploading of answers.

Answers

Other Instructions

1. Instruction 1: The exam is divided into two parts (Part 1 and Part 2). The first contains 10 short questions worth 5 marks each and the second 5 long questions worth 10 marks each. Both parts cover the full material taught in the unit.
2. Instruction 2: Note that sharing information with colleagues is strictly forbidden and that we have a set of measures in place to identify cases of plagiarism.
3. Instruction 3: This is NOT a open book exam.

Part 1: Short questions (5 marks each)

Question 1 (5 marks)

What are the key differences between the different forms of machine learning (unsupervised, supervised and reinforcement learning)?

Solution: The key difference is on how many teaching signals they need. Whereas unsupervised learning only requires the data itself, supervised learning requires specific teaching signals (targets) (3 marks). Reinforcement learning is somewhere in between as it requires teaching signals but those are implicitly provided by the environment in the form of rewards (2 marks).

Question 2 (5 marks)

Which of the following models (Figure 1) overfit or/and underfit the data and *why*?

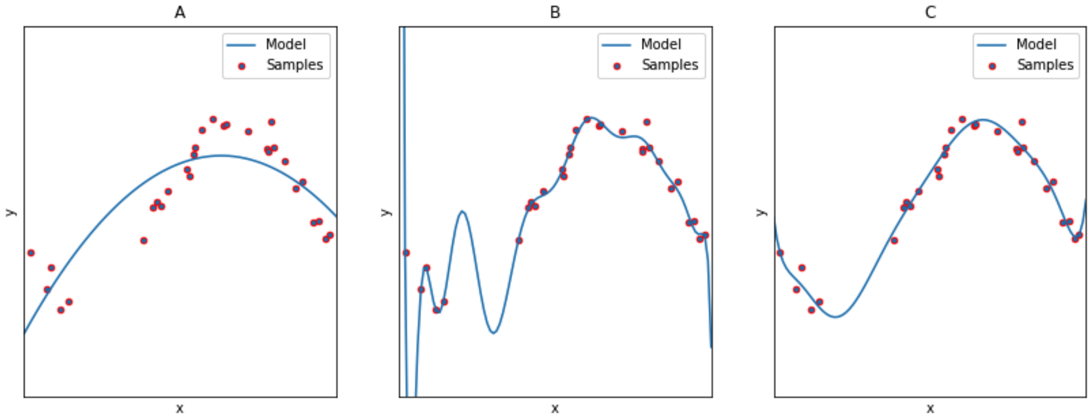


Figure 1: Examples of data (red scatter plot) with three different models (A,B,C; blue solid line).

Solution: Model A can be considered as a case of underfitting to the data as important aspects of the data are not captured (2 marks). Whereas the model in B can be considered a case of overfitting as the model tries to capture every single variation in the data (2 marks). Finally the model in C provides a good fit to the data, with a model complexity that is between model A and B (1 mark).

Question 3 (5 marks)

Figure 2 shows a Bayesian network structure (i.e. directed acyclic graph). Write down all pairs of variables which are independent conditional on A.

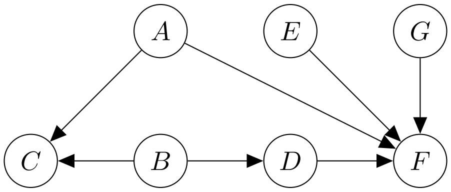


Figure 2: Directed acyclic graph for Question 3

Solution: BE, BG, CE, CG, DE, DG, EG

Question 4 (5 marks)

Let S be the sample covariance matrix of some data where S is:

$$\begin{pmatrix} 3 & -2 \\ -2 & 4 \end{pmatrix}$$

Let $u_1 = (-0.6154, 0.7882)$ be the first principal component of the data (approximated to 4 significant figures). What is the variance of the data when projected onto one dimension using this principal component? (Give you answer using 4 significant figures.)

Solution: 5.562

Question 5 (5 marks)

Consider running the Metropolis-Hastings algorithm where the target distribution p is a mixture of Gaussians and we use a symmetric proposal distribution. Let z^t be the current state and let z^* be a proposed new state. Suppose that $p(z^*) \geq \epsilon p(z^t)$ for some value $\epsilon > 0$. If possible, write down the acceptance probability as a function of ϵ, otherwise explain why it is not possible.

Solution: The acceptance probability is $\min(1, p(z^*)/p(z^t))$, but we can not express this as a function of ϵ, since ϵ is only a bound on this value, and we are not told it is above 1.

Question 6 (5 marks)

Suppose you were using EM to estimate the parameters of a mixture of 3 Gaussians. Let x_i be a training datapoint where $\mathcal{N}(x_i|\mu_1, \Sigma_1) = 0.3$, $\mathcal{N}(x_i|\mu_2, \Sigma_2) = 0.4$, $\mathcal{N}(x_i|\mu_3, \Sigma_3) = 0.2$. μ_j and Σ_j are the current parameter values for the jth Gaussian. Suppose the current parameter values for the mixing coefficients were $\pi_1 = 0.2$, $\pi_2 = 0.4$ and $\pi_3 = 0.4$. Compute (up to 3 significant figures) the *responsbility* γ_{ij} for $j = 1, 2, 3$, i.e. the probability that x_i was generated from the jth Gaussian.

Solution: $\gamma_{i1} \propto 0.3 \times 0.2 = 0.06$, $\gamma_{i2} \propto 0.4 \times 0.4 = 0.16$, $\gamma_{i3} \propto 0.2 \times 0.4 = 0.08$. We have $0.06 + 0.16 + 0.08 = 0.3$. So $\gamma_{i1} = 0.06/0.3 = 0.200$, $\gamma_{i2} = 0.16/0.3 = 0.533$, $\gamma_{i3} = 0.08/0.3 = 0.267$.

Question 7 (5 marks)

Here is a dataset with 2 datapoints where each row is a datapoint:

$$\begin{pmatrix} 2 & 0 \\ 4 & -5 \end{pmatrix}$$

Let $\phi_1(x) = x^T x$ and $\phi_2(x) = x_1 - 2x_2$. Write down the Gram matrix for this data using the kernel associated with the feature map ϕ where $\phi(x) = (\phi_1(x), \phi_2(x))$.

Solution:

$\phi(\boldsymbol{x}_1) = \phi((2,0)) = (\phi_1((2,0)), \phi_2((2,0))) = (4,2)$.

$\phi(\boldsymbol{x}_2) = \phi((4,-5)) = (\phi_1((4,-5)), \phi_2((4,-5))) = (41,14)$.

We have: $(4,2)^T(4,2) = 20$, $(4,2)^T(41,14) = 192$ and $(41,14)^T(41,14) = 1877$.

So the Gram matrix is $\begin{pmatrix} 20 & 192 \\ 192 & 1877 \end{pmatrix}$

Question 8 (5 marks)

You are using a linear dynamical system to predict a continuous state variable, z_n. For the current time-step n, your prior distribution over the state is a Gaussian:

$$p(z_n | x_1, ..., x_{n-1}) = \mathcal{N}(0, 20^2), \tag{1}$$

with mean of zero and variance of 20^2. You observe a noisy measurement of the state, $x_n = 50$, where $x_n = z_n + \epsilon_n$ and the noise ϵ_n is zero-mean, Gaussian distributed:

$$p(\epsilon_n) = \mathcal{N}(0, 1^2). \tag{2}$$

You use the Kalman filter to update your distribution over z_n, giving you a posterior:

$$p(z_n | x_1, ..., x_n) = \mathcal{N}(\mu, \sigma^2). \tag{3}$$

(a) Is the value of the posterior mean, μ, closer to 0 or 50?

(b) Is the value of σ greater or less than 20?

Solution:

(a) [1 point for answer, 2 for explanation] μ is closer to 50. The prior variance is 20, which much higher than the noise variance, which is 1. This means that the posterior will be closer to the observation. Calculations using Bayes' rule can alternatively be shown as working.

(b) [1 point for answer, 1 for explanation] σ is less than 20. The noisy observation has decreased our uncertainty in the value of z_n, which is reflected in a lower posterior variance. Calculations using Bayes' rule can alternatively be shown as working.

Question 9 (5 marks)

Bagging trains classifiers on different subsets of the dataset. How can this improve the performance of the ensemble?

Solution:

[3 points] To increase the diversity of classifiers so that their errors are less strongly correlated, i.e., so that they each make mistakes on different data points.

[2 points] If the base models in the ensemble have uncorrelated, zero-mean errors, the expected error rate of the ensemble will be 1/M x the expected error rate of the average individual, where M is the number of base models.

Question 10 (5 marks)

Crowdsourcing is often used to annotate datasets for training and evaluating machine learning systems. List at least three ways in which we can improve the quality of the labels obtained from crowdsourcing.

Solution:

Give 2 points (up to maximum of 5 total) for each point they make correctly:

Increase the number of annotators (crowdworkers) who label each data point

Ask multiple workers to label each data point.

Use voting to combine annotations (e.g., class labels) from multiple annotators.

Use averaging to combine numerical annotations (e.g., ratings from 1 to 10) from multiple annotators.

Weight annotators' votes using a probabilistic model (e.g., Dawid and Skene).

Use machine learning to distinguish spammers from more accurate annotators.

Clarify the instructions for the annotators.

Part 2: Long questions (10 marks each)

Question 11 (10 marks)

You are given a very simple neural network (Figure 3) with only one input x and one output neuron y with a sigmoid activation function σ. Consider that the network is currently receiving the following data point $x = 0.1$ and that its target should be $t = 1$. Assume a standard mean squared error, $E = \frac{1}{2}(y - t)^2$.

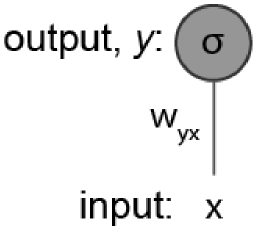


Figure 3: Schematic of a simple artificial neural network.

(a) To perform gradient descent on the weights in your network you must compute $w_{yx} = w_{yx} - \frac{\partial E}{\partial w_{yx}}$. What is the numeric value for $\frac{\partial E}{\partial w_{yx}}$? Assume that the current $w_{yx} = 0.9$ and that there is no bias. Your answer only needs to be approximate (e.g. 0.06 instead of 0.062348). Make sure to explain the equations and steps used to arrive at your solution. (5 marks)

(b) What is the curse of dimensionality and how do artificial neural networks models deal with it? (3 marks)

(c) Are artificial neural networks parametric or non-parametric models? Explain *why*. (2 marks)

Solution:

1. Using the chain rule you should obtain: $\frac{\partial E}{\partial w_{yx}} = (y - t)\sigma'(w_{yx}x)x$ (3 marks), where $\sigma'(x) = \sigma(x) \times (1 - \sigma(x))$ is the derivative of the sigmoid activation function. The exact solution is: $(\sigma(0.9 \times 0.1) - 1) \times \sigma'(0.9 \times 0.1) \times 0.1) \sim -0.0119$ (2 marks).

2. In practice the curse of dimensionality is not a problem because data has inherent smoothness properties and is often localized to small regions of the bigger possible space (1 mark). Neural networks like many ML models implicitly exploit this property as they find input-output mappings that work within this low-dimensional space in which the data lies (1 mark). Indeed, if you test neural networks outside this data with a data point that lies outside the training space (i.e. a generalization test) they (like other models) typically fail (1 mark).

3. ANNs are typically parametric models as their number of parameters does not grow with more data. However, there exist some (recent) variants that are non-parametric (2 marks).

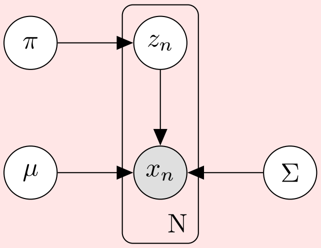
Question 12 (10 marks)

One option when learning the parameters of a Gaussian mixture model is to use EM to (attempt to) find the maximum likelihood estimates of the parameters. In this question we consider an alternative Bayesian approach

(a) Draw the Gaussian mixture model as a Bayesian network using plate notation. Do not forget to include nodes which represent the model parameters. It is fine to use a single node to represent a random variable which has matrix or vector values. State which nodes of your Bayesian network are observed and which not. You do not need to represent the number of Gaussians in the mixture model. You do not need to define priors over model parameters. (5 marks)

(b) One way to get an approximation to the posterior distribution over model parameters is sampling: either sampling from the posterior distribution itself or sampling from a sequence of distributions that get ever closer to the posterior distribution. Consider (i) rejection sampling and (ii) MCMC as methods for approximating the posterior distribution for a Gaussian mixture model. Give pros and cons of each method and state which method is preferable for this problem. (5 marks)

Solution: This is actually just Fig 9.6 in Bishop. Only the shaded node is observed.



For the second part, rejection sampling should be rejected since the observed quantity is continuous: almost surely every sample will be rejected. There is really no argument for using it. MCMC is a sensible option, but students should mention that one needs to be careful about burn-in and checking for convergence.

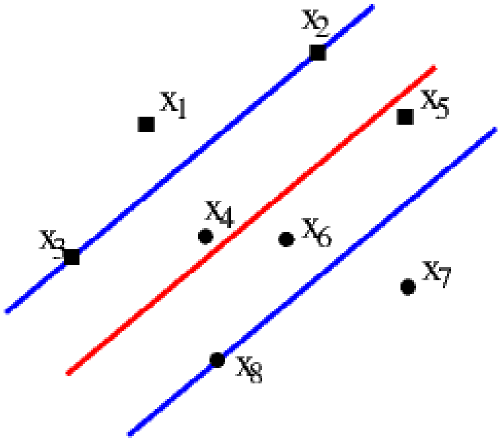


Figure 4: SVM results for Question 13

Question 13 (10 marks)

(a) Fig 4 shows the results (in the implicit feature space) of learning with an SVM. The separating hyperplane is in red and the margin is indicated by the blue lines. Training datapoints x_1, x_2, x_3 and x_5 have class label -1 and training datapoints x_4, x_6, x_7 and x_8 have class label 1. x_7 is correctly classified. Which data points are misclassified and which are support vectors? (5 marks)

(b) The following minimisation problem (or an equivalent formulation) is solved when using SVMs with soft margins:

$$\begin{aligned} \min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{n=1}^{N} \zeta_n \\ \text{subject to } t_n(w^T \phi(x_n) + b) \geq 1 - \zeta_n, \\ \zeta_n \geq 0, n = 1, ..., N \end{aligned} \tag{4}$$

Explain what each of the symbols represents and how altering the value of C affects an SVM classifier. (5 marks)

Solution: x_4 and x_5 are the only misclassified data points. All data points apart from x_1 and x_7 are support vectors. These are the support vetors: x_2, x_3, x_4, x_5, x_6 and x_8. vectors.

Here, t_n is the nth training point. w and b defines the separating hyperplane $y(x) = w^T \phi(x) + b$, ζ_n is the slack variable for datapoint t_n. If t_n lies on the right side of the margin then we can set $\zeta_n = 0$ and so incur no penalty for t_n. Otherwise there is a penalty. The parameter C regulates how big the penalty for non-zero slack variables should be.

Question 14 (10 marks)

This question is about hidden Markov models (HMM). We have a discrete text sequence labelling task with three classes. You have trained a hidden Markov model (HMM) using expectation maximization (EM) and obtained the following estimates of the initial state probabilities, π, and transition matrix, A:

$$\pi = p(z_1) = [0.1, 0.9] \qquad A = p(z_{n+1}|z_n) = \begin{matrix} 0.5 & 0.5 \\ 0.1 & 0.9 \end{matrix} \tag{5}$$

Rows in A correspond to values of z_n and columns to values of z_{n+1}.

The likelihood of each token in the vocabulary, x_n, given the state, z_n, is given by an emission distribution matrix, which you have also learned already using EM. In this matrix, the rows correspond to the token values, i and the columns to the state values, j. The emission distribution matrix is shown here:

$$p(x_n = i|z_n = j) = \begin{array}{ccc} x_n = i & z_n = 1 & z_n = 2 \\ \text{in} & 0.1 & 0.3 \\ \text{Merge} & 0.2 & 0.05 \\ \text{octopus} & 0 & 0.1 \\ \text{Released} & 0.01 & 0.1 \\ \text{Senegal} & 0.4 & 0.1 \\ \text{States} & 0.29 & 0.05 \\ \text{the} & 0 & 0.3 \end{array}$$

Suppose you observe the following short sequence of text tokens: ["In", "Senegal"]

(a) Compute the probability $p(z_1|x_1 = \text{“}In\text{”}, x_2 = \text{“}Senegal\text{”})$ to three decimal places.

(b) True or false: the Viterbi algorithm is an alternative to the forward-backward algorithm. Explain your answer in one or two sentences.

Solution: (a)

[3 points] Using the steps of the forward-backward algorithm.

In the forward pass, compute:

$\alpha(z_1 = k) = p(z_1 = 1, x_1 = \text{“}In\text{”}) = \pi_k * p(x_n = \text{“}In\text{”}|z_1 = 1)$

$\alpha(z_1 = 1) = 0.1 * 0.1 = 0.01$

$\alpha(z_1 = 2) = 0.9 * 0.3 = 0.27$

[2 points] In the backward pass:

$\beta(z_2 = 1) = \beta(z_2 = 2) = 1$

$\beta(z_1 = 1) = \sum_{l=1}^{K} A_{1l} p(x_2 = \text{“}Senegal''|z_2 = l)\beta(z_2 = l) = 0.5*0.4*1+0.5*0.1*1 = 0.25$

$\beta(z_1 = 2) = \sum_{l=1}^{K} A_{2l} p(x_2 = \text{“}Senegal''|z_2 = l)\beta(z_2 = l) = 0.1*0.4*1+0.9*0.1*1 = 0.13$

[2 points] Posteriors:

$p(z_1 = 1|x_1 = \text{“}In'', x_2 = \text{“}Senegal'') = \alpha(z_1 = 1)\beta(z_1 = 1)/(\alpha(z_1 = 1)\beta(z_1 = 1) + \alpha(z_1 = 1)\beta(z_1 = 2)) = 0.01 * 0.25/(0.01 * 0.25 + 0.27 * 0.13) = 0.066$

$p(z_1 = 2|x_1 = \text{“}In'', x_2 = \text{“}Senegal'') = \alpha(z_1 = 2)\beta(z_1 = 2)/(\alpha(z_1 = 1)\beta(z_1 = 1) + \alpha(z_1 = 2)\beta(z_1 = 2)) = 0.27 * 0.13/(0.01 * 0.25 + 0.27 * 0.13) = 0.934$

(b) [1 point for answer, 2 points for explanation] False: the Viterbi algorithm finds the most likely sequence of labels, whereas the forward-backward algorithm is needed for the E-step or to compute the posteriors over each state (responsibilities).

Question 15 (10 marks)

This question is about CART decision trees. Suppose we want to train a classifier to decide whether an email is important to a user or not given three features. The training data is shown below, where each row contains the data for one email:

Sender in address book	‘important’ flag	Percentage of spam phrases	Important?
N	N	10	N
N	N	10	N
Y	N	20	N
N	Y	5	N
N	Y	7	N
N	N	4	Y
Y	Y	7	Y
Y	Y	20	Y

(a) Suppose we learn a CART decision tree, where the objective used to grow the tree is to minimise cross-entropy error. No pruning is used. Draw the final learned tree and state the amount of residual error.

(b) What is pruning and why is it often used with decision trees?

Solution: (a) [5 points for correct tree diagram. Points for correct decision node + 2 for the leaves.] Top node: ’sender in address book’. If True, next node: ’has important flag’, followed by leaves: True = Important, False = not important. If top node is False, next node is: ’percentage of spam phrases < 5’. True = Important, False = not important.

[1 point] There is no residual error because the features completely separate the samples into two classes.

(b) [1 point] Pruning removes certain nodes of the tree if they meet a particular criterion.

[2 points] The criterion balances training set error against model complexity. Alternatively, they may define a criterion using an equation.

[1 point] Pruning is used to reduce overfitting.

END OF PAPER