The University of Melbourne

School of Computing and Information Systems

COMP90051
Statistical Machine Learning
2020 Semester 2 – Final Exam

Identical examination papers: None.

Exam duration: 180 minutes.

Reading time: 15 minutes included in 30 minutes total of reading plus upload time.

Length: This paper has 6 pages including this cover page.

Authorised materials: Lecture slides, workshop materials, prescribed reading, your own project reports.

Calculators: Permitted.

Instructions to students: The total marks for this paper is 180, corresponding to the number of minutes available. The mark will be scaled to compute your final exam grade.

This paper has three parts, A-C. You should attempt all the questions.

This is an open book exam. You should enter your answers in a Word document or PDF, which can include typed and/or hand-written answers. You should answer each question on a separate page, i.e., start a new page for each of Questions 1–9 – parts within questions do not need new pages. Write the question number clearly at the top of each page. You have unlimited attempts to submit your answer-file, but only your last submission is used for marking.

You must not use materials other than those authorised above. You are not permitted to communicate with others for the duration of the exam, other than to ask questions of the teaching staff via the discussion board. Your computer, phone and/or tablet should only be used to access the authorised materials, enter or photograph your answers, and upload these files.

Library: This paper is to be lodged with the Baillieu Library.

Continued overleaf ...

COMP90051 Statistical Machine Learning Final Exam

Semester 2, 2020

Total marks: 180

Students must attempt all questions

Section A: Short Answer Questions [60 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each question in 1-3 lines, with longer responses expected for the questions with higher marks.

Question 1: [60 marks]

(a) Explain why *maximum likelihood estimation*, *max a posteriori* and *empirical risk minimisation* are all instances of *extremum estimators*. [5 marks]

(b) We saw two *bias-variance decompositions* in the subject: (i) for *parameter estimation* and (ii) for *supervised regression*. Why was there a third *irreducible error* term in case (ii) but not in case (i)? [5 marks]

(c) Both the *learning with experts* and *multi-armed bandit* settings involve learning to optimise *cumulative rewards* (or equivalently losses). What is a key difference between the settings? [5 marks]

(d) Describe how the *frequentist* and *Bayesian* approaches differ in their modelling of *unknown parameters*. [5 marks]

(e) A model family $\mathcal{F}$'s *growth function* $S_{\mathcal{F}}(m)$ could potentially be 2^m, growing exponentially larger with increasing sample size m. Why would this be bad news for the *PAC bound with growth function* for $\mathcal{F}$? [5 marks]

(f) Why doesn't *max a posteriori* estimation require computation of the *evidence* through costly *marginalisation*, while computing the *posterior distributions* does? [5 marks]

(g) Consider as an alternate to the popular *square loss* in *supervised regression*, the function $\ell(y; \hat{y}) = (\hat{y} - y)^5$ measuring loss between label $y \in \mathbb{R}$ and prediction $\hat{y} \in \mathbb{R}$. Is this loss a good idea or a bad idea? Why? [10 marks]

(h) Suppose you have trained a *soft-margin SVM with a RBF kernel*, but the performance is poor on both training and validation sets. How will you change the hyperparameters of the SVM to improve the performance? List two strategies. [5 marks]

(i) Compared with *Root Mean Square Propagation (RMSProp)* algorithm, what's the main drawback of *Adaptive Gradient (AdaGrad)* algorithm? [5 marks]

(j) What strategy makes the *Variational Autoencoder (VAE)* capable of applying gradient descent through the samples of latent representation z to the encoder? [5 marks]

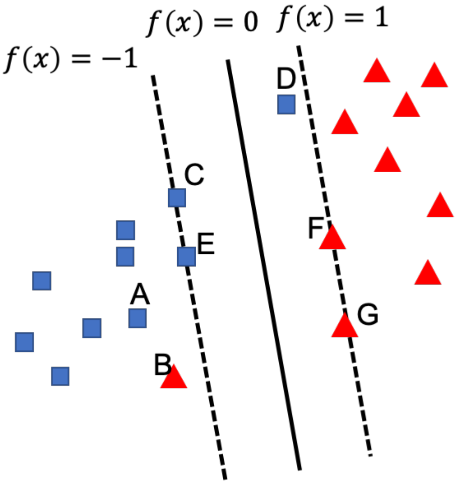
(k) For a *graph neural network* described as $\mathbf{h}_{(i,j)}^{(l)} = g^{(l)}\left(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}; \mathbf{x}_{(i,j)}\right)$, $\mathbf{h}_j^{(l+1)} = f^{(l)}\left(\sum \mathbf{h}_{(i,j)}^{(l)}\right)$, where f and g are the network functions, $\mathbf{h}$ and $\mathbf{x}$ are the hidden states and input data, explain in words if the hidden states and input data are located on the nodes, the edges, or both? [5 marks]

Section B: Method & Calculation Questions [84 marks]

In this section you are asked to demonstrate your conceptual understanding of methods that we have studied in this subject, and your ability to perform numeric and mathematical calculations.

Question 2: Support Vector Machines [14 marks]

Assume that the following figure is an illustration of the *decision boundary* learned by a *linear soft-margin SVM* on a two-class data set, with points from one class labelled with blue squares and points from the other class labelled with red triangles.



(a) Identify the training points that are *support vectors* (by index A–G). [6 marks]

(b) The *Lagrangian function* of the *soft-margin SVM* is:

$$L(w,b,\lambda,\beta,\xi) = \frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\lambda_i g_i(w,b,\xi) + \sum_{i=1}^{n}\beta_i(-\xi_i)$$
$$g_i(w,b,\xi) = 1 - \xi_i - y^{(i)}(w^T x^{(i)} + b)$$

For each of the four points A,B C and D, is $\lambda_i = 0$, $0 < \lambda_i < C$, $\lambda_i = C$ or $\lambda_i > C$? [4 marks]

For each of the four points A,B C and D, is $\xi_i = 0$, $0 < \xi_i < 1$, $1 < \xi_i < 2$ or $\xi_i > 2$? [4 marks]

Question 3: Newton-Raphson for Linear Regression [10 marks]

Recall that *linear regression*'s objective can be written as $L(\boldsymbol{\theta}) = \frac{1}{2}\|\mathbf{X}\boldsymbol{\theta} - \mathbf{y}\|_2^2$, where $\mathbf{X}$ is a $n \times d$ matrix of training instances in rows, $\mathbf{y} \in \mathbb{R}^n$ are the training labels, and $\boldsymbol{\theta} \in \mathbb{R}^d$ are the parameters to learn by minimising $L(\boldsymbol{\theta})$. Show that for any starting $\boldsymbol{\theta}_0 \in \mathbb{R}^d$, a single step of Newton-Raphson optimisation of $L(\boldsymbol{\theta})$ recovers the *normal equation* exact solution. (Hints for taking vector/matrix-valued derivatives with respect to $\boldsymbol{\theta}$, for constants matrix $\mathbf{A}$ and vector $\mathbf{b}$. Hint 1: $\nabla(\mathbf{A}\boldsymbol{\theta} + \mathbf{b}) = \mathbf{A}$. Hint 2: $\nabla_2(\mathbf{A}\boldsymbol{\theta}) = \mathbf{A}$. Hint 3: $\nabla(\|f(\boldsymbol{\theta})\|_2^2) = 2(\nabla f(\boldsymbol{\theta}))' f(\boldsymbol{\theta})$. Where prime $'$ denotes transpose, ∇ the gradient, and ∇_2 the Hessian as in class, here always with respect to $\boldsymbol{\theta}$.)

Question 4: Convolutional Operation [10 marks]

(a) Assume that in the following figure (a) is an *image patch* and (b) is a *kernel* that is used to perform *convolution* (*stride = 2, no padding*) on the patch. Show the *output feature map* after convolution. [4 marks]

1	0	1	0
0	1	1	0
0	0	0	1
0	0	1	0

(a)

1	0
0	1

(b)

(b) Assume that (a) and (b) in the following figure are two image patches. Now consider a 3×3 *binary weight matrix* (each element of the matrix is either 0 or 1) that is used as the kernel to perform convolution (*stride = 1, no padding*) on the two image patches. Assume that after convolution, the output feature maps are K_a and K_b. What kernel can make the difference between K_a and K_b most significant (sum of all elements of $|K_a - K_b|$ is maximum)? List two such kernels. [6 marks]

0	1	0
1	1	1
0	1	0

(a)

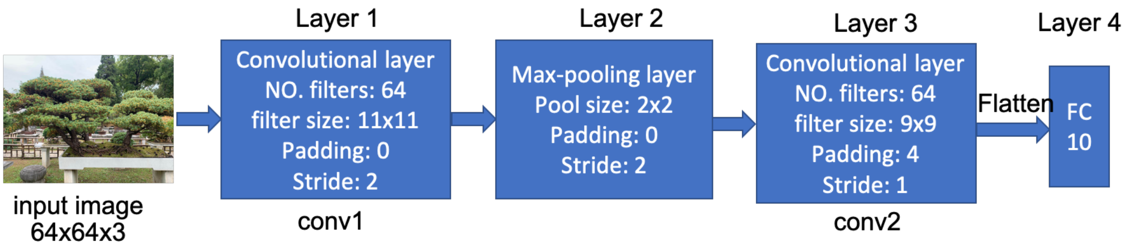
1	0	1
0	1	0
1	0	1

(b)

Question 5: CNN Architectures [16 marks]

Assume that you train a *convolutional neural network (CNN)* on a GPU. The size of each input image is 64x64x3 (3 channels). The following figure shows the settings of the layers in the CNN. As shown, there are four layers (blue blocks) in the CNN. Specifically, there are two *convolutional layers* (conv1 and conv2). Between them there is a *max-pooling layer*. The output feature map of conv2 is flattened into a vector, which is then fed to a *fully-connected layer (FC)*. The FC layer contains 10 output units.

Note: The padding size is the number of rows/columns on each size of the input. For example, if the size of an input is 5×5 and the padding size is 1, then after padding, each dimension of the input becomes $5 + 2 \times 1 = 7$, so the input becomes 7×7.



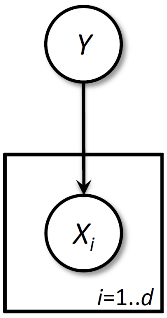
(a) Write the sizes of output feature maps of the Layer 1, Layer 2 and Layer 3 (write in the form 'width × height × channels') [6 marks]

(b) How many parameters in *each of the four layers* (exclude the bias) in the CNN? Show your working. [4 marks]

(c) How to replace conv2 to keep the same size of *receptive field* of conv2 with fewer parameters? Write three different replacements. [6 marks]

Question 6: Naïve Bayes as a Linear Classifier [14 marks]

Consider a standard *naïve Bayes binary classifier* on d binary features, with model specified by the following *plate diagram* where $\Pr(Y = y) = p^y(1-p)^{1-y}$ for $y \in \{0,1\}$ and $\Pr(X_i = x \mid Y = y) = p_{i,y}^x(1-p_{i,y})^{1-x}$ for $x, y \in \{0,1\}$ and $i \in \{1,\ldots,d\}$. Where this model has already been trained, we have the $2d+1$ naïve Bayes parameters $p, p_{1,0}, p_{1,1}, \ldots, p_{d,0}, p_{d,1}$ all specified.



Naïve Bayes makes classifications for a new *test instance* $\mathbf{x} = (x_1, \ldots, x_d)$ (breaking ties always in favour of class $y = 1$) using the *posterior predictive distribution*, as

$$f(\mathbf{x}) = \arg\max_{y\in\{0,1\}} \Pr(Y = y \mid \mathbf{X} = \mathbf{x}) \enspace .$$

(a) Show that this classifier is a *linear classifier* i.e. there exists some $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$ such that for any test instance $\mathbf{x}$, $f(\mathbf{x}) = 1$ if and only if $\mathbf{w}'\mathbf{x} \geq b$. (Hint 1: It might be useful to use $f(\mathbf{x}) = 1$ if and only if $\Pr(Y = 1 \mid \mathbf{X} = \mathbf{x}) \geq \Pr(Y = 0 \mid \mathbf{X} = \mathbf{x})$. Hint 2: You may assume that all $2d + 1$ naïve Bayes parameters are in $(0, 1)$.) [8 marks]

(b) Write down the specific $\mathbf{w}, b$ as a function of the $2d + 1$ parameters. [6 marks]

Question 7: Vapnik-Chervonenkis Dimension [20 marks]

The two parts of this question related to the *VC dimension* but are otherwise unrelated.

(a) Consider an *input domain* of five points $x_1, \ldots, x_5$, and *binary classifier family* $\mathcal{F}$ defined by the table of *dichotomies* given below. Calculate $\mathrm{VC}(\mathcal{F})$ and show a corresponding shattered set of points. [6 marks]

x_1	x_2	x_3	x_4	x_5
1	0	1	0	0
0	0	0	0	1
0	0	0	1	0
0	0	0	1	1
1	1	0	0	0
0	1	1	0	1
0	1	0	1	0
0	1	0	1	1

(b) Consider a family $\mathcal{F}$ of *classifiers* mapping a finite domain of n instances $\{x_1, \ldots, x_n\}$ into *binary labels* $\{-1, +1\}$; and a second family $\mathcal{G}$ of classifiers acting on a *distinct* domain of m instances $\{x_{n+1}, \ldots, x_{n+m}\}$. Now define a new *product family* $\mathcal{H}$ that can act on the *combined domain* $\{x_1, \ldots, x_{n+m}\}$, where for every pair $f \in \mathcal{F}, g \in \mathcal{G}$, we define an $h \in \mathcal{H}$ such that, for each instance $x_i \in \{x_1, \ldots, x_{n+m}\}$, it outputs classification

$$h(x_i) = \begin{cases} f(x_i), & 1 \leq i \leq n \\ g(x_i), & n+1 \leq i \leq n+m \end{cases} \enspace .$$

What is $\mathrm{VC}(\mathcal{H})$ in terms of $\mathrm{VC}(\mathcal{F}), \mathrm{VC}(\mathcal{G})$? (Hint: It may be helpful to think of families $\mathcal{F}, \mathcal{G}$ as being tables of unique dichotomies in rows, with columns being instances $x_1, \ldots, x_n$ and $x_{n+1}, \ldots, x_{n+m}$ respectively. Then $\mathcal{H}$ is a table with columns $x_1, \ldots, x_{n+m}$.) [14 marks]

Continued overleaf ...

Section C: Design and Application Questions [36 marks]

In this section you are asked to demonstrate that you have gained a high-level understanding of the methods and algorithms covered in this subject, and can apply that understanding. Answers should be about 1 page in length for each question.

Question 8: COVID-19 Vaccine Selection [18 marks]

Many months have passed since the onset of COVID-19, and 5 medical laboratories have created and trialed *5 vaccines* with no significant adverse side effects after careful clinical trials. From the trials it is believed that some vaccines are more effective at preventing contraction of COVID-19 for some patients, but no one knows which. Doctors across the world must now determine *how to select the right vaccine for any given patient*, as patients stream into general practice clinics. How might you as Chief Health Officer use machine learning to help fight COVID-19 by *recommending appropriate vaccines*?

(a) Formulate the *vaccine selection task* as a non-contextual *multi-armed bandit problem*. Explain what your arms are, what rounds correspond to, and what are your rewards. [6 marks]

(b) Suppose in earlier clinical trials various measurements were taken of patients prior to test vaccination. Given access to these measurements, and effectiveness of the vaccinations in trials, how might you decide whether *contextual multi-armed bandits* might offer an improvement over your non-contextual solution? [4 marks]

(c) Suppose 100 machine learning companies approach you with their MAB solutions, that have never been previously tested. What drawback is there to trying all of the 100 approaches out on real patients in the wild? [4 marks]

(d) Now suppose the prior clinical trails were completely randomised, such that for a patient that arrived, a *uniformly random* vaccine was given to them, and the result recorded. How could you use this data to *evaluate all 100 MAB learners*? [4 marks]

Question 9: Cafe Surveillance [18 marks]

To balance economic recovery with public health safety, the City of Melbourne has decided to have an automatic system to count the number of *customers in Melbourne restaurants*. They have installed surveillance cameras in some cafes and now they wish to have an algorithm to count the number of people who are in these businesses, measure when they are using the facilities, and to categorise users by age, gender and other demographic variables. The City also has access to other real-time data sources, such as the use of parking meters, and traffic detectors on nearby main roads.

(a) Based on what you learned in this subject, which techniques would you suggest that could be useful in the implementation of this system and why? [6 marks]

(b) Describe the steps in the design and implementation of your method focusing on:

   (i) What features you would use? How you construct them from the data? [3 marks]

   (ii) What are the different sections in your system? [3 marks]

   (iii) Any model hyperparameters that you need to tune in your system development. [3 marks]

   (iv) What challenges you expect to face and how you would overcome them. [3 marks]

— End of Exam —