The University of Melbourne

School of Computing and Information Systems

COMP90051
Statistical Machine Learning
November 2017

Identical examination papers: None

Exam duration: 180 minutes

Reading time: 15 minutes

Length: This paper has 7 pages including this cover page.

Authorised materials: None

Calculators: Not permitted

Instructions to invigilators: The examination paper is to remain in the examination room. Please provide extra script books on request.

Instructions to students: The total marks for this paper is 50. This paper has four parts, A-D. You should attempt all 11 questions.

Please ensure your student number is written on all script books and answer sheets during writing time. Please start the answer to each question on a new page in the script book. The left-hand unlined pages of script books are for draft working and notes and *will not be marked*.

Mobile phones, tablets, laptops, and other electronic devices, wallets and purses must be placed beneath your desk. All electronic devices (including mobile phones and phone alarms) must be switched off and remain under your desk until you leave the examination venue. No items may be taken to the toilet.

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

Student id:

COMP90051 Statistical Machine Learning
Exam

Semester 2, 2017

Total marks: 50

Students must attempt all questions

Section A: Short Answer Questions [13 marks]

Answer each of the questions in this section as briefly as possible. Expect to answer each sub-question in a couple of lines per mark.

Question 1: General Machine Learning [13 marks]

(a) Name two problems in *frequentist supervised learning* that can be addressed using *regularisation*, and in each case, explain why. [2 marks]

(b) Consider the *frequentist supervised learning* setup. Is it always possible to achieve zero training error by increasing *model complexity*? Explain using an example(s), and discuss whether zero training error is a good thing. [2 marks]

(c) Name an advantage of using *lasso regularisation* method over the *ridge regression*. [1 mark]

(d) *Perceptrons* and *hard-margin SVMs* are both *linear classifiers*. What is the difference between the two methods? [1 mark]

(e) State what mathematical operation is implemented by the *back-propagation* algorithm, as used in learning *artificial neural networks*. [1 mark]

(f) In a sentence each, describe what *probabilistic inference* and *statistical inference* are for. [2 marks]

(g) Explain why *Bayesian model selection* tends to prefer simpler models over complex ones, and state a situation when a more complex model would be preferred. [2 marks]

(h) Describe the process of *Bayesian sequential updating*, making reference to the *prior*, *likelihood* and *posterior* distributions. [1 mark]

(i) *Gaussian mixture modeling* and *k-means algorithm* can both be used for *clustering*. Name another similarity between these methods. [1 mark]

Section B: Method Questions [17 marks]

In this section you are asked to demonstrate your conceptual understanding of a subset of the methods that we have studied in this subject.

Question 2: Linear Regression [6 marks]

(a) Outline the steps for setting the value of *regularisation* parameter λ in *ridge regression* using *cross-validation* or *held-out validation*. [2 marks]

(b) Why can't one find the value of λ in the same way weights $\boldsymbol{w}$ are found? [1 mark]

(c) Show that the standard method for training a linear regression model (minimising the sum of squared errors) is equivalent to the *maximum likelihood estimate* for the graphical model, $y \sim \mathbf{x}'\mathbf{w} + \epsilon$, where y are the outputs, $\mathbf{x}$ the inputs, $\mathbf{w}$ the model parameters and ϵ is zero-mean Gaussian noise. Present your answer mathematically, and show the stages of your working. [3 marks]

In your answer to the above, may find the formulation of the Gaussian distribution handy:

$$\mathcal{N}(v|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(v-\mu)^2\right\} .$$

Question 3: Support Vector Machines [4 marks]

(a) A method is called *non-parametric* if it does not rely on a fixed number of *parameters*. Explain why a *SVM* can be considered *non-parametric*. [2 marks]

(b) Consider a definition of a *kernel* as the *dot product* in some *feature space*. Show that a sum of two *kernels* is a *kernel*. [2 marks]

Question 4: Ensemble Methods [3 marks]

Outline how the *bagging approach* to *ensemble learning* works. What is the key assumption underlying the method? Explain why this leads to bagging often making better predictions than just using a single classifier.

Question 5: Probabilistic Models [4 marks]

(a) Bayesian models often use a *conjugate prior*. With the aid of an example, state what it means to be *conjugate*, and explain why it is of practical importance. [2 marks]

(b) The *Maximum a Posteriori (MAP)* method is sometimes referred to as "poor man's Bayes." Explain how MAP is different to full Bayesian inference, and describe a situation in which the two methods will produce different results. [2 marks]

Section C: Numeric Questions [10 marks]

In this section you are asked to demonstrate your understanding of a subset of the methods that we have studied in this subject, in being able to perform numeric calculations.

Question 6: Artificial Neural Networks [2 marks]

Artificial neural networks (ANN) are capable of representing arbitrary *Boolean functions*. Draw an *ANN* that represents a *Boolean function* defined by the table below. Specify the value of each *weight*, and indicate which *activation function* is used for each node. Explicitly draw *bias* terms if these are used. Note x_1, x_2, x_3 are the inputs to the network.

x_1	x_2	x_3	output
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	0
1	0	0	0
1	0	1	0
1	1	0	1
1	1	1	0

Question 7: Gaussian Mixture Models [3 marks]

Assume a *Gaussian mixture model* with parameters $\boldsymbol{\theta}$. In your answers to both parts of this question below you can leave the Gaussian equation as $\mathcal{N}(\boldsymbol{x}|\ldots)$, replacing "..." with the parameters of a Gaussian.

(a) Write down the probability $P(\boldsymbol{x}_i|\boldsymbol{\theta})$ of a point $\boldsymbol{x}_i \in \boldsymbol{R}^d$. Identify what parameters are included in vector $\boldsymbol{\theta}$ and state any constraints that apply to the parameters. [1 mark]

(b) Write down the conditional probability of point i originating from cluster c given data $P(z_i = c|\boldsymbol{x}_i, \boldsymbol{\theta})$. [2 marks]

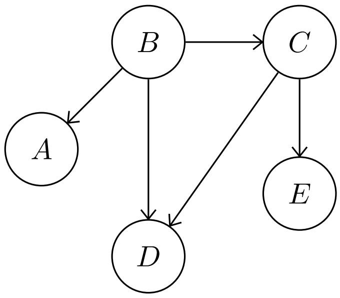
Question 8: Dimensionality Reduction [2 marks]

Let $\boldsymbol{x}$ be a vector, and A be a matrix so that $A\boldsymbol{x}$ is defined. Let α be a real number. Recall that if $A\boldsymbol{x} = \alpha\boldsymbol{x}$, then $\boldsymbol{x}$ is called an *eigenvector* of A with the corresponding *eigenvalue* α.

Consider a set of m *high-dimensional points mapped* onto a *line* (i.e., to 1D). Let $\boldsymbol{x}$ be an $m-$*dimensional vector* representing 1D coordinates of the mapped points. According to *Laplacian eigenmaps* method, the best mapping is the one that minimizes $\boldsymbol{x}'L\boldsymbol{x}$, where L is the *Laplacian* of the *similarity graph* constructed from *high-dimensional points*, and $\boldsymbol{x}$ is restricted to have a fixed norm, e.g., $\boldsymbol{x}'\boldsymbol{x} = 1$. Show that the optimal $\boldsymbol{x}^*$ is an *eigenvector* of L.

Question 9: Probabilistic Inference [3 marks]

Consider the following directed *probabilistic graphical model* (PGM) over five binary-valued random variables, denoted A, B, C, D and E.



(a) State the form of the joint probability density function, $P(A, B, C, D, E)$. [1 mark]

(b) State all the *independence relations* that hold between B and E, considering both *marginal* and *conditional* independence. [2 marks]

Section D: Design and Application Questions [10 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. Expect your answer to each question to be from one third of a page to one full page in length. These questions may require significantly more thought than those in Sections A–C and should be attempted only after having completed the earlier sections.

Question 10: Movie Recommendation [5 marks]

Your task is to design an automatic *movie recommendation* system for a company that provides an on-line movie streaming service. This company maintains a large *collection* of movies, and each movie can be accessed for a small fee. The service is only available for *registered users*. The users preview the movies and can pay for access to the full movie. For each *user*, the system keeps track of what movies were previewed and what movies were purchased. The aim of your system is to provide personalised *movie recommendations* to each user, as to maximise the number of purchases. Note that your system will only be applied for users that already have a history of a large number of previews. With all these considerations in mind, outline the design of your *movie recommendation* system.

(a) Formulate the *movie recommendation task* as a *supervised learning* problem. Explain, what your training instances are, and what is the target variable. What features you will use, and how do you construct them from the data? [2 marks]

(b) You have decided to use a kernel classifier for this purpose. Outline why using a kernel might convey an advantage for this problem over other methods, such as a linear classifiers. [1 mark]

(c) It is also possible to use *unsupervised learning methods* to assist with the task. What *unsupervised learning methods* might you apply for this task, and how? [2 marks]

Question 11: Modelling diseases and symptoms [5 marks]

Influenza is a common disease, commonly known as the flu. It manifests with symptoms including muscle pain, coughing, fever and sneezing. A doctor will make a judgement about whether or not a patient has the flu, on the grounds of an individual's symptoms. The doctor may also make use of a clinical diagnostic test. These tests provide more reliable information than using symptoms alone, but are not perfect, in that they can return false positive or false negative results.

(a) Draw a directed *probabilistic graphical model (*PGM*)* to best model the above scenario, based on the following random variables: F, whether the individual has the flu; S_i, the presence of symptom type i, $i = 1 \dots I$; and T the result of the test. All random variables are binary valued. You may choose to include other random variables, as needed. [2 marks]

(b) Now incorporate D, the doctor's diagnosis. How might the PGM differ if the doctor is known to be perfectly accurate, or inaccurate, when making their diagnosis based on the evidence? You may choose to include other random variables, as needed. [1 mark]

(c) In making a diagnosis, the PGM can be used to compute a distribution over F. State the desired distribution, based on an individual attending the doctor with specific symptoms. [1 mark]

(d) The parameters of the PGM are in the form of conditional probability tables. Explain how these parameters can be learned from data, and what kind of data would be required. [1 mark]

— End of Exam —