The University of Melbourne

School of Computing and Information Systems

COMP90051
Statistical Machine Learning
2019 Semester 2 – Practice

Identical examination papers: None

Exam duration: 180 minutes (for real exam)

Reading time: 15 minutes

Length: This paper has 9 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 the real paper is 50 (this practice is just 22). This paper has four parts, A-D. You should attempt all 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: The real paper is to be lodged with the Baillieu Library.

Student id:

Continued overleaf . . .

Formulae

Bias of an estimator $\hat{\theta}$ of population parameter θ: $B_\theta(\hat{\theta}) = \mathbb{E}_{\mathbf{X}\sim\theta}[\hat{\theta}(X_1, \ldots, X_n)] - \theta$
Variance of $\hat{\theta}$: $Var_\theta(\hat{\theta}) = \mathbb{E}_\theta[(\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}])^2]$

Logistic function: $f(t) = \frac{1}{1+e^{-t}}$

Naïve Bayes joint distribution on one labelled example $(\mathbf{x}, y)$ in some d-dimensional feature space: $p(\mathbf{x}, y) = p(y)\prod_{j=1}^{d} p(x_j \mid y)$ typically for Boolean features in which case $p(y = true) = \theta$ and each $p(x_j \mid y) = \theta_{j,y}^{x_j}(1 - \theta_{j,y})^{1-x_j}$

Linear regression normal equations: $\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

Convolution $\mathbf{b} = \mathbf{a} * \mathbf{w}$ defined as $b_i = \sum_{\delta=-c}^{C} a_{i+\delta} w_{\delta+C+1}$, where C is the size/length of filter $\mathbf{w}$.

SVM hard-margin learner primal: $\arg\min_{\boldsymbol{w},b} ||\boldsymbol{w}||$ s.t. $y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1$, $\forall i$
Hard-margin dual: $\arg\max_{\boldsymbol{\lambda}} \Sigma_{i=1}^{n}\lambda_i - \frac{1}{2}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}\lambda_i\lambda_j y_i y_j \boldsymbol{x_i}\cdot\boldsymbol{x_j}$ s.t. $\lambda_i \geq 0\ \forall i$ and $\Sigma_{j=1}^{n}\lambda_j y_j = 0$
Dual recovery of b, for any support vector j: $b = {y_j}^{-1} - \sum_{i=1}^{n}\lambda_i y_i \mathbf{x}_i'\mathbf{x}_j$
Making prediction on instance $\boldsymbol{x}$ via sign of $b + \mathbf{w}'\mathbf{x} = b + \Sigma_{i=1}^{n}\lambda_i y_i \boldsymbol{x_i}'\boldsymbol{x}$

SVM soft-margin primal: $\arg\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i$ s.t. $\forall i,\ y_i(\mathbf{w}'\mathbf{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geq 0$
Hinge loss: $\ell_h(y, \hat{y}) = \max\{0, 1 - y\hat{y}\}$
Soft-margin dual: $\arg\max_{\boldsymbol{\lambda}} \Sigma_{i=1}^{n}\lambda_i - \frac{1}{2}\Sigma_{i=1}^{n}\Sigma_{j=1}^{n}\lambda_i\lambda_j y_i y_j \boldsymbol{x_i}\cdot\boldsymbol{x_j}$ s.t. $C \geq \lambda_i \geq 0\ \forall i$ and $\Sigma_{j=1}^{n}\lambda_j y_j = 0$

Polynomial kernel of degree p: $K(\mathbf{u}, \mathbf{v}) = (1 + \mathbf{u}\cdot\mathbf{v})^p$
RBF kernel of bandwidth σ: $K(\mathbf{u}, \mathbf{v}) = \exp\left(\frac{-||\mathbf{u}-\mathbf{v}||^2}{2\sigma^2}\right)$
Positive semidefinite matrix $\mathbf{K}$, if $\mathbf{v}'\mathbf{K}\mathbf{v} \geq 0$ for all $\mathbf{v} \neq \mathbf{0}$.
Positive definite if $\mathbf{v}'\mathbf{K}\mathbf{v} > 0$ for all $\mathbf{v} \neq \mathbf{0}$.

1-D Gaussian: $N(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$
d-dimensional Gaussian: $N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}))$
Beta distribution: $\text{Beta}(\theta; \alpha, \beta) = \frac{\gamma(\alpha+\beta)}{\gamma(\alpha)\gamma(\beta)}\theta^{\alpha-1}\beta^{\beta-1}$
Bernoulli distribution: $\text{Ber}(x; \theta) = \theta^x(1 - \theta)^{1-x}$

UCB arm i value estimate after observing rewards: $Q_{t-1}(i) = \hat{\mu}_{t-1}(i) + \sqrt{\frac{\rho\log(t)}{N_{t-1}(i)}}$ where $N_{t-1}(i)$ is the number of observations of arm i, $\hat{\mu}_{t-1}(i)$ is the average of observed rewards on arm i so far; $\rho > 0$ a hyperparameter. Unobserved arms are initialised with value estimate Q_0 a hyperparameter.

 Continued overleaf . . .

COMP90051 Statistical Machine Learning Practice Exam

Semester 2, 2019

Total marks: 50 in real exam based on more questions; 22 in this practice exam

Students must attempt all questions

Section A: Short Answer Questions [6 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 [6 marks]

(a) What are the *free parameters* of a *Gaussian mixture model*? What algorithm is used to fit them for *maximum likelihood estimation*? [2 marks]

Acceptable 1 mark: For a Gaussian mixture with k components the parameters are probabilities for $(k-1)$ components, a mean vectors for each of the k components, and a symmetric positive-definite covariance matrix for each of the k components.
Acceptable 1 mark: The EM algorithm is appropriate for maximum likelihood estimates.

(b) In words or a mathematical expression, what quantity is minimised by *linear regression*? [1 mark]

Acceptable: The residual sum of errors
Acceptable: The mean-squared error
Acceptable: $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ (terms are true and estimated labels) or this times a constant

(c) In words or a mathematical expression, what is the *marginal likelihood* for a *Bayesian probabilistic model*? [1 mark]

Acceptable: the joint likelihood of the data and prior, after marginalising out the model parameters
Acceptable: $p(\mathbf{x}) = \int p(\mathbf{x}|\theta)p(\theta)d\theta$ where $\mathbf{x}$ is the data, θ the model parameter(s), and $p(\mathbf{x}|\theta)$ the likelihood and $p(\theta)$ the prior
Acceptable: the expected likelihood of the data, under the prior

(d) In words, what does $\Pr(A, B \mid C) = \Pr(A \mid C)\Pr(B \mid C)$ say about the *dependence* of A, B, C? [1 mark]

Acceptable: A and B are conditionally independent given C.

(e) In words or a mathematical expression, what is the *chain rule* of probability? [1 mark]

Acceptable: $\Pr(X_1, \ldots, X_k) = \prod_{i=1}^{k} \Pr(X_i \mid X_{i+1}, \ldots, X_k)$
Acceptable: For any joint likelihood, and ordering on the random variables, expansion of joint as product of conditionals of each random given variables coming after (alternatively: before) in ordering.

Continued overleaf ...

Section B: Method Questions [4 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: Ensemble Methods [2 marks]

Please write the following in your script book, and there connect each dot on the left with one dot on the right, to create the best possible correspondence

Stacking	○	○	Weighted-voting over weak classifiers	[0.5 marks]
Boosting	○	○	Trees with reduced feature sets	[0.5 marks]
Bagging	○	○	Bootstrap aggregation	[0.5 marks]
Random forests	○	○	Meta-classifier over base classifiers	[0.5 marks]

Acceptable (0.5 mark each):
Stacking ↔ Meta-classifier over base classifiers
Boosting ↔ Weighted-voting over weak classifiers
Bagging ↔ Boostrap aggregation
Random forests ↔ Trees with reduced feature sets

Question 3: Kernel methods [2 marks]

(a) Consider a 2-dimensional *dataset*, where each point is represented by two *features* and the *label* (x_1, x_2, y). The features are binary, the label is the result of XOR function, and so the data consists of four points $(0, 0, 0)$, $(0, 1, 1)$, $(1, 0, 1)$ and $(1, 1, 0)$. Design a *feature space transformation* that would make the data *linearly separable*. [1 mark]

Acceptable: new feature space (x_3), where $x_3 = (x_1 - x_2)^2$

(b) Why do the *primal* and *dual* optima for the *hard/soft-margin support vector machines* coincide? [1 mark]

Acceptable: As strong duality holds due to convexity of the objectives.

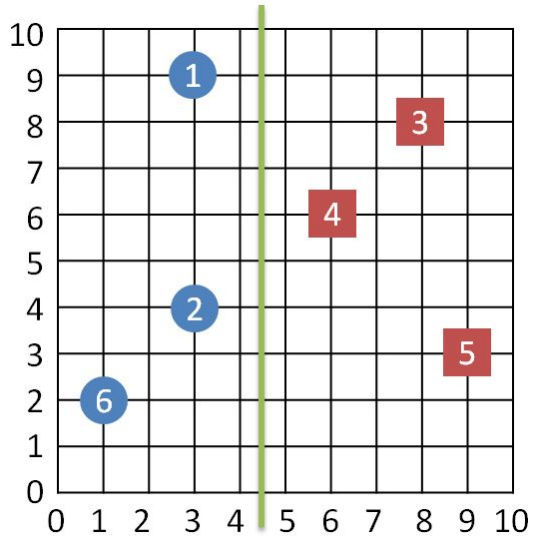
Continued overleaf . . .

Section C: Calculation Questions [7 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 and mathematical calculations.
NOTE: in the real exam, a small number of questions from this section will be a bit harder/longer than others.

Question 4: Kernel Methods [2 marks]

Consider the data shown below with *hard-margin linear SVM decision boundary* shown between the classes. The right half is classified as red squares and the left half is classified as blue circles. Answer the following questions and explain your answers.



(a) Which points (by index 1–6) would be the *support vectors* of the *SVM*? [1 mark]

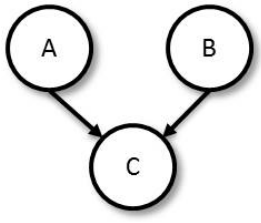
Acceptable: Points 1, 2, 4 would all be support vectors as they all lie on the margin.

(b) What is the value of the *hard margin SVM loss* for point 3? [1 mark]

Acceptable: zero, since the point is on the right side of the boundary and is outside the margin.

Question 5: Statistical Inference [3 marks]

Consider the following directed PGM



where each random variable is Boolean-valued (True or False).

(a) Write the format (with empty values) of the conditional probability tables for this graph. [1 mark]

```
------------
Pr(A=True)
------------
```

 Continued overleaf . . .

```
?
------------

------------
Pr(B=True)
------------
?
------------

------------------
A B Pr(C=True|A,B)
------------------
T T ?
T F ?
F T ?
F F ?
------------------
```

(b) Suppose we observe n sets of values of A, B, C (complete observations). The maximum-likelihood principle is a popular approach to training a model such as above. What does it say to do? [1 mark]

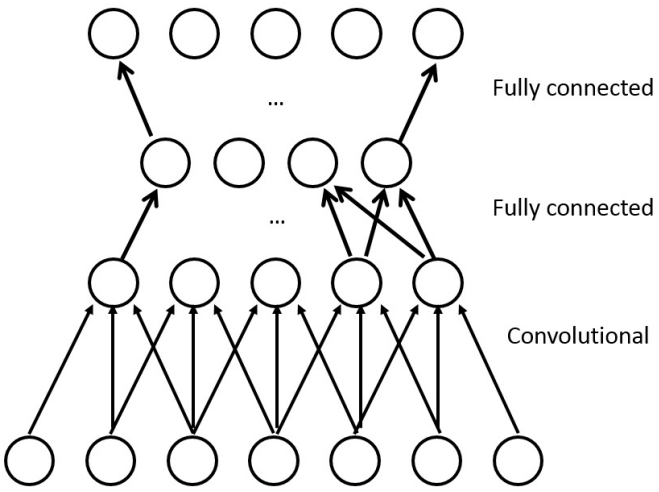
Acceptable: It says to choose values in the tables that maximise the likelihood of the data.
Acceptable: $\arg\max_{tables} \prod_{i=1}^{n} \Pr(A = a_i) \Pr(B = b_i) \Pr(C = c_i \mid A = a_i, B = b_i)$

(c) Suppose we observe 5 training examples: for (A, B, C) — $(F, F, F); (F, F, T); (F, T, F); (T, F, T); (T, T; T)$. Determine maximum-likelihood estimates for your tables. [1 mark]

Acceptable: The MLE decouples when we have fully-observed data, and for discrete data as in this case — where the variables are all Boolean — we just count.
The $Pr(A = True)$ is 2/5 since we observe A as true out of five observations. Similarly for B we have the probability of True being 2/5. Finally for each configuration TT, TF, FT, FF of AB we can count the times we see C as True as a fraction of total times we observe the configuration. So we get for these probability of $C = True$ as 1.0, 1.0, 0.0, 0.5 respectively.

Question 6: Artificial Neural Networks [2 marks]

How many *parameters* does the following *convolutional neural network* have? Show your working.



 Continued overleaf ...

Acceptable: The convolutional network has 3 matrix-valued parameters, adding up the sizes gives $3 + 5 \cdot 4 + 4 \cdot 5 = 43$ parameters in total (assuming that biases are not included).

Continued overleaf ...

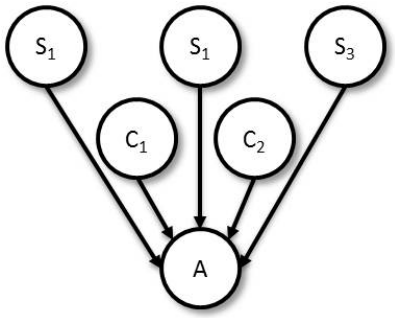
Section D: Design and Application Questions [5 marks]

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.

Question 7: Bushfire Evacuation Alerts [5 marks]

Your task is to design a system for alerting residents of the Dandenongs that they should evacuate for an impending bushfire. The Dandenongs is an area of Victoria that suffers from regular bushfires in the summer when temperatures are high, and humidity low. When fires are close to a fictional town called Bayesville, you must alert residents that they should evacuate. If fires are too close, then you should not advise evacuation as residents are safer if they stay where they are (at home).
The Country Fire Association has deployed sensors around the area that monitor whether a fire is in progress at each sensor's location; in particular if any of three sensors S_1, S_2, S_3 are 'on' then residents should evacuate. However if either of the closer sensors C_1, C_2 are 'on' then residents should stay put.

(a) Model the above problem as a *directed probabilistic graphical model* (PGM). In particular, you need not provide any probabilty tables, just the graph relating random variables S_1, S_2, S_3, C_1, C_2 and an additional r.v. A for alerting residents to evacuate. [1 mark]



(b) How many conditional probability tables should be specified for your model, and what should these tables' dimensions be? [1 mark]

Acceptable: One table per r.v. so 6 tables. All 5 sensor r.v.'s should have tables with one column — a value for the probability that entry is True (probability of False is one minus this). The CPT for A should have 6 columns, one per parent and one for A.

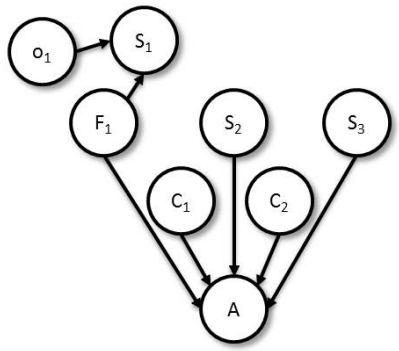
(c) Suppose you are told by the Fire Commissioner that the sensors are always accurate. What could you say about your CPTs? [1 mark]

Acceptable: This doesn't tell you how often the sensors are 'on' so nothing about the sensors' tables. However it tells us that you can trust the sensors and you can determine the CPT for A exactly as $(S_1 \vee S_2 \vee S_3) \wedge \neg(C_1 \vee C_2)$

(d) How would you change your model if the Commissioner then tells you that the sensor S_1 is not perfectly accurate? [1 mark]

Acceptable: We should add a r.v. F_1 as to whether there's really a fire at S_1, and another r.v. o_1 as to whether S_1 is operational. Now we have that the alarm should depend on the unobserved F_1.

Continued overleaf ...



(e) Given this final model, assuming you have trained it and completed all the necessarily CPTs, how would you use it to drive the alarm to evacuate? [1 mark]

Acceptable: We can use probabilistic inference — the elimination algorithm — to determine $\Pr(A = True)$ from observations of the five sensors. When doing elimination, the five sensors will be observed (no real summing there) but we will be summing over the unobserved o_1 and F_1.

— End of Exam —