UNIVERSITY OF SOUTHAMPTON COMP6208W1

SEMESTER 2 EXAMINATION 2021/22

ADVANCED MACHINE LEARNING

Duration 120 mins (2 hours)

This paper is a WRITE-ON examination paper.

You ***must*** write your Student ID on this Page and must not write your name anywhere on the paper.

All answers should be written within the designated boxes in this examination paper and sufficient space is provided for each question.

If, for some reason, space is required to complete or correct an answer to a question, use the "Additional Space" provided on the facing or adjacent page to the question. Clearly indicate which question the answer corresponds to.

No credit will be given for answers presented elsewhere and without clear indication of to what question they correspond. Blue answer books may be used for scratch; they will be discarded without being looked at.

Answer all parts of the question in section A (40 marks)
and ALL three questions from section B (20 marks each)

Student ID:

Question	Mark	*Arithmetic checked*	*Double Marked*
A1	/40		
B2	/20		
B3	/20		
B4	/20		
Total:			

University approved calculators MAY be used.

10 page examination paper

Section A

A 1

(a) In the bias variance dilemma explain (1) what does the variance measure and (2) why adding a regularisation term might reduce the variance. [5 marks]

(Tests basic theoretical knowledge.)

1 The variance measures the expected (squared) variation between the prediction of a learning machine trained on a particular dataset and the prediction of the mean machine.

2 By adding a regularisation term we reduce the variability of the machine so that its response is less dependent on the training set.

(b) Explain how *gradient boosting* works? [5 marks]

(Test basic knowledge of algorithms.)

Gradient boosting is an ensemble learning technique (usually using shallow decision trees). A strong learner is built iteratively by training a weak learner on the difference between the target value and the current strong learner. This weak learner is added to the strong learner to produce the new strong learner.

(c) (1) Give a definition of a positive definite kernel and (2) explain why kernels for Gaussian Processes must be positive definite. [5 marks]

(First part is more sophisticated book work—with multiple solutions. The second part requires understanding what positive definite means.)

(i) A positive definite kernel function, $K(x, y)$, will satisfy the constraint

$$\int \int f(\boldsymbol{x})\, K(\boldsymbol{x}, \boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y} > 0$$

for any non-zero function $f(x)$. (Alternatively answers would include that the eigenvalues are all positive or that the Gram matrix for every set of points is positive definite.)

(ii) The kernel function must be positive definite for a Gaussian process as it represents the covariance function which needs to be positive definite. If not there would be directions where the probability diverges and there would be no meaningful probability density.

(d) Show that if $\lambda > 0$ is an eigenvalue of $\mathbf{C} = \mathbf{X}\,\mathbf{X}^{\mathsf{T}}$ then it is also an eigenvalue of $\mathbf{D} = \mathbf{X}^{\mathsf{T}}\mathbf{X}$, where $\mathbf{X}$ is a matrix. [5 marks]

(Algebraic questions requiring a good grasp of linear algebra that underlies are lot of duality in machine learning.)

If $\lambda \neq 0$ is an eigenvalue of C then there exists some eigenvector v ($\neq 0$) such that $\mathbf{C}\,v = \lambda\, v$. Multiplying on the left by $\mathbf{X}^\mathsf{T}$ then

$$\mathbf{X}^\mathsf{T}\mathbf{C}\,v = \lambda \mathbf{X}^\mathsf{T} v$$

but $\mathbf{X}^\mathsf{T}\mathbf{C} = \mathbf{X}^\mathsf{T}(\mathbf{X}\,\mathbf{X}^\mathsf{T}) = (\mathbf{X}^\mathsf{T}\mathbf{X})\,\mathbf{X}^\mathsf{T} = \mathbf{D}\,\mathbf{X}^\mathsf{T}$ (where we have used the associativity of matrix multiplication and the definitions of C and D). Thus

$$\mathbf{D}\,\mathbf{X}^\mathsf{T} v = \lambda \mathbf{X}^\mathsf{T} v$$

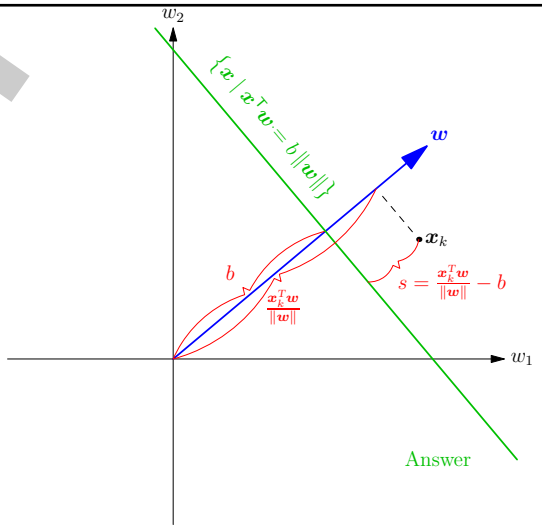
or $\mathbf{D}u = \lambda\, u$ where $u = \mathbf{X}^\mathsf{T} v \neq 0$ is an eigenvector of D with eigenvalue λ.

(e) Sketch the set of points $\{\boldsymbol{x} \,|\, \boldsymbol{x}^\mathsf{T}\boldsymbol{w} = b\,\|\boldsymbol{w}\|\}$, which generate a separating plane and explain why

$$y_k\left(\frac{\boldsymbol{x}_k^\mathsf{T}\boldsymbol{w}}{\|\boldsymbol{w}\|} - b\right) \geq \gamma$$

implies that all points $(\boldsymbol{x}_k, y_k)$ with $y_k \in \{-1, 1\}$ will lie a distance of γ or more from the separating plane. [5 marks]

(Test understanding of constraints in SVMs.)



The value $\frac{\boldsymbol{x}_k^\mathsf{T}\boldsymbol{w}}{\|\boldsymbol{w}\|} - b$ measures the shortest separation between a point $\boldsymbol{x}_k$ and the separating plane defined by w and b. This is negative if the point lies below the separating plane. For data with $y_k = 1$ the constraint implies that all points lie a distance of γ or greater above the separating plane, while for points with $y_k = -1$ the constraints imply that all points lie a distance of γ below the separating plane.

(f) Explain how to do model selection in Bayesian inference. [5 marks]

(Test fundamental understanding.)

Bayesian inference is predicated on a model, $\mathcal{M}$, through

$$p(\boldsymbol{\theta}|\mathcal{D}, \mathcal{M}) = \frac{p(\mathcal{D}|\boldsymbol{\theta}, \mathcal{M})\, p(\theta|\mathcal{M})}{p(\mathcal{D}|\mathcal{M})}.$$

The marginal likelihood or evidence $p(\mathcal{D}|\mathcal{M})$ is a measure of how well the model, $\mathcal{M}$, explains the evidence. To select between models we can choose the model with the highest marginal likelihood (or we can put a prior on the models and compute the posterior probability of the model given the data).

(g) If $\|\boldsymbol{x}\|$ is a proper norm use the triangular inequality ($\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$), linearity of a norm ($\|a\,\boldsymbol{x}\| = a\,\|\boldsymbol{x}\|$) and the definition of convexity, to show that the norm is convex. [5 marks]

(The students have not seen this. The algebra is simple, but the students need a good understanding to do the proof.)

For any vectors, x, and y and any scale $a \in [0, 1]$ then

$$\|a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}\| \leq \|a\,\boldsymbol{x}\| + \|(1-a)\,\boldsymbol{y}\| = a\,\|\boldsymbol{x}\| + (1-a)\,\|\boldsymbol{y}\|$$

where the first inequality follows from the triangular inequality and the second equality from the linearity of the norm. However,

$$\|a\,\boldsymbol{x} + (1-a)\,\boldsymbol{y}\| \leq a\,\|\boldsymbol{x}\| + (1-a)\,\|\boldsymbol{y}\|$$

is the defining equation of convexity.

(h) Show that positive semi-definite kernels form a convex set. You may assume elementary properties of positive semi-definite kernels. [5 marks]

(The students have not seen this and it requires combining knowledge taught in different lectures.)

Let $\mathcal{P}$ be the set of positive semi-definite kernels. To show this forms a convex set we require that for any two elements $K_1(\boldsymbol{x}, \boldsymbol{y}), K_2(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ then the linear combination

$$K_3(\boldsymbol{x}, \boldsymbol{y}) = a\,K_1(\boldsymbol{x}, \boldsymbol{y}) + (1-a)\,K_2(\boldsymbol{x}, \boldsymbol{y})$$

with $a \in [0, 1]$ must also be in $\mathcal{P}$. If $a = 0$ then $K_3 = K_2 \in \mathcal{P}$, while if $a = 1$ the $K_3 = K_1 \in \mathcal{P}$. Otherwise $a, (1-a) > 0$. If $c > 0$ and $K(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ then $c\,K(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$. So that $a\,K_1(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$ and $(1-a)\,K_2(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$. But the sum of two kernels in $\mathcal{P}$ will also be in $\mathcal{P}$ so $K_3(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{P}$.

End of question A1

Section B

B 2

(a) We consider a regression problem where the data $(\boldsymbol{x}, y)$ is distributed according to $\gamma(\boldsymbol{x}, y)$. We consider a learning machine that makes a prediction $\hat{f}(\boldsymbol{x}|\boldsymbol{\theta})$, where the parameters, $\boldsymbol{\theta}$ are trained using a stochastic algorithm that returns parameters distributed according to a probability density $\rho(\boldsymbol{\theta})$. We can define the mean machine as $\hat{m}(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}\sim\rho}\Big[\hat{f}(\boldsymbol{x}|\boldsymbol{\theta})\Big]$. We assume that

$$\mathbb{E}_{(\boldsymbol{x},y)\sim\gamma}\Big[\left(\hat{m}(\boldsymbol{x}) - y\right)^2\Big] = B, \qquad \mathbb{E}_{(\boldsymbol{x},y)\sim\gamma}\Big[\mathbb{E}_{\boldsymbol{\theta}\sim\rho}\Big[\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}) - \hat{m}(\boldsymbol{x})\Big)^2\Big]\Big] = V.$$

That is, we can define a bias B and variance V. We now consider ensembling n machines

$$\hat{f}_n(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n}\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i)$$

where $\boldsymbol{\theta}_i$ are drawn independently from $\rho(\boldsymbol{\theta})$. Compute the expected generalisation error of $\hat{f}_n(\boldsymbol{x})$.

[10 marks]

(This mixes two calculates the students have seen (the bias variance dilemma calculation and the effect of averaging n random variables). It is a good test of the students understanding.)

We note that $\mathbb{E}_{\boldsymbol{\theta}}\Big[\hat{f}_n(\boldsymbol{x})\Big] = \hat{m}(\boldsymbol{x})$ so the bias variance dilemma calculation still holds

$$\mathcal{E}_n = \mathbb{E}_{\boldsymbol{\theta}}\Big[\mathbb{E}_{(\boldsymbol{x},y)}\Big[\Big(\hat{f}_n(\boldsymbol{x}) - y\Big)^2\Big]\Big] = \mathbb{E}_{\boldsymbol{\theta}}\Big[\mathbb{E}_{(\boldsymbol{x},y)}\Big[\Big((\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x})) + (\hat{m}(\boldsymbol{x}) - y)\Big)^2\Big]\Big]$$
$$= \mathbb{E}_{(\boldsymbol{x},y)}\Big[\left(\hat{m}(\boldsymbol{x}) - y\right)^2\Big] + \mathbb{E}_{\boldsymbol{\theta}}\Big[\mathbb{E}_{(\boldsymbol{x},y)}\Big[\Big(\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x})\Big)^2\Big]\Big]$$

where the cross term vanishes as

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\Big(\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x})\Big)\ \big(y - \hat{m}(\boldsymbol{x})\big)\Big] = \big(y - \hat{m}(\boldsymbol{x})\big)\,\mathbb{E}_{\boldsymbol{\theta}}\Big[\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x})\Big] = 0.$$

Thus $\mathcal{E}_n$ consists of a bias B and a new variance V_n where

$$V_n = \mathbb{E}_{\boldsymbol{\theta}}\Big[\mathbb{E}_{(\boldsymbol{x},y)}\Big[\Big(\hat{f}_n(\boldsymbol{x}) - \hat{m}(\boldsymbol{x})\Big)^2\Big]\Big] = \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\left(\frac{1}{n}\sum_{i=1}^{n}\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\Big)\right)^2\right]\right]$$
$$= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\frac{1}{n^2}\sum_{i=1}^{n}\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\Big)^2 + \frac{1}{n^2}\sum_{i,j=1 j\neq i}^{n}\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\Big)\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_j) - \hat{m}(\boldsymbol{x})\Big)\right]\right]$$
$$= \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbb{E}_{(\boldsymbol{x},y)}\left[\frac{1}{n^2}\sum_{i=1}^{n}\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i) - \hat{m}(\boldsymbol{x})\Big)^2\right]\right] = \frac{V}{n}$$

since

$$\mathbb{E}_{\boldsymbol{\theta}}\Big[\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_i)-\hat{m}(\boldsymbol{x})\Big)\Big(\hat{f}(\boldsymbol{x}|\boldsymbol{\theta}_j)-\hat{m}(\boldsymbol{x})\Big)\Big]=0$$

as θ_i and θ_j are independent. Thus the expected generalisation performance is equal to $\mathcal{E}_n = B + V/n$.

(b) Consider a classification problem where $\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})$ is the probability that a learning machine with parameters $\boldsymbol{\theta}$ predicts that input $\boldsymbol{x}$ belongs to class $c \in \mathcal{C}$. Assume the training is stochastic so the probability of obtaining parameters $\boldsymbol{\theta}$ is $\rho(\theta)$. Let $\hat{m}_c(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{\theta}}\Big[\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\Big]$ be the output of the mean machine for class c. Assuming that for a data point $(\boldsymbol{x}, y)$, where y is a class label, we use a cross entropy loss

$$L(\boldsymbol{x}, y, \boldsymbol{\theta}) = -\sum_{c\in\mathcal{C}} [\![y = c]\!] \log\Big(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\Big),$$

show that the expected loss over inputs and parameters can be written as the expected loss of the mean machine plus a second loss. Use Jensen's inequality ($\mathbb{E}\left[\log(X)\right] \leq \log\big(\mathbb{E}\left[X\right]\big)$) to show the second term is positive. [10 marks]

(Students have not seen this although it is not difficult to prove.)

$$\begin{aligned}\mathcal{E} &= \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[-\sum_{c\in\mathcal{C}} [\![y = c]\!] \log\Big(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\Big)\right]\right] \\ &= \mathbb{E}_{(\boldsymbol{x},y)}\left[-\sum_{c\in\mathcal{C}} [\![y = c]\!] \log(\hat{m}_c(\boldsymbol{x}))\right] + \mathbb{E}_{(\boldsymbol{x},y)}\left[\mathbb{E}_{\boldsymbol{\theta}}\left[-\sum_{c\in\mathcal{C}} [\![y = c]\!] \log\left(\frac{\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})}{\hat{m}_c(\boldsymbol{x})}\right)\right]\right]\end{aligned}$$

The first terms acts like a bias. The second (variance-like) term is

$$\mathbb{E}_{(\boldsymbol{x},y)}\left[-\sum_{c\in\mathcal{C}} [\![y = c]\!]\, \mathbb{E}_{\boldsymbol{\theta}}\Big[\log\Big(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\Big)\Big]\right] - \mathbb{E}_{(\boldsymbol{x},y)}\left[-\sum_{c\in\mathcal{C}} [\![y = c]\!] \log(\hat{m}_c(\boldsymbol{x}))\right]$$

But using Jensen's inequality
$\mathbb{E}_{\boldsymbol{\theta}}\Big[\log\Big(\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\Big)\Big] \leq \log\Big(\mathbb{E}_{\boldsymbol{\theta}}\Big[\hat{f}_c(\boldsymbol{x}|\boldsymbol{\theta})\Big]\Big) = \log(\hat{m}_c(\boldsymbol{x}))$**. Thus this second term is positive.**

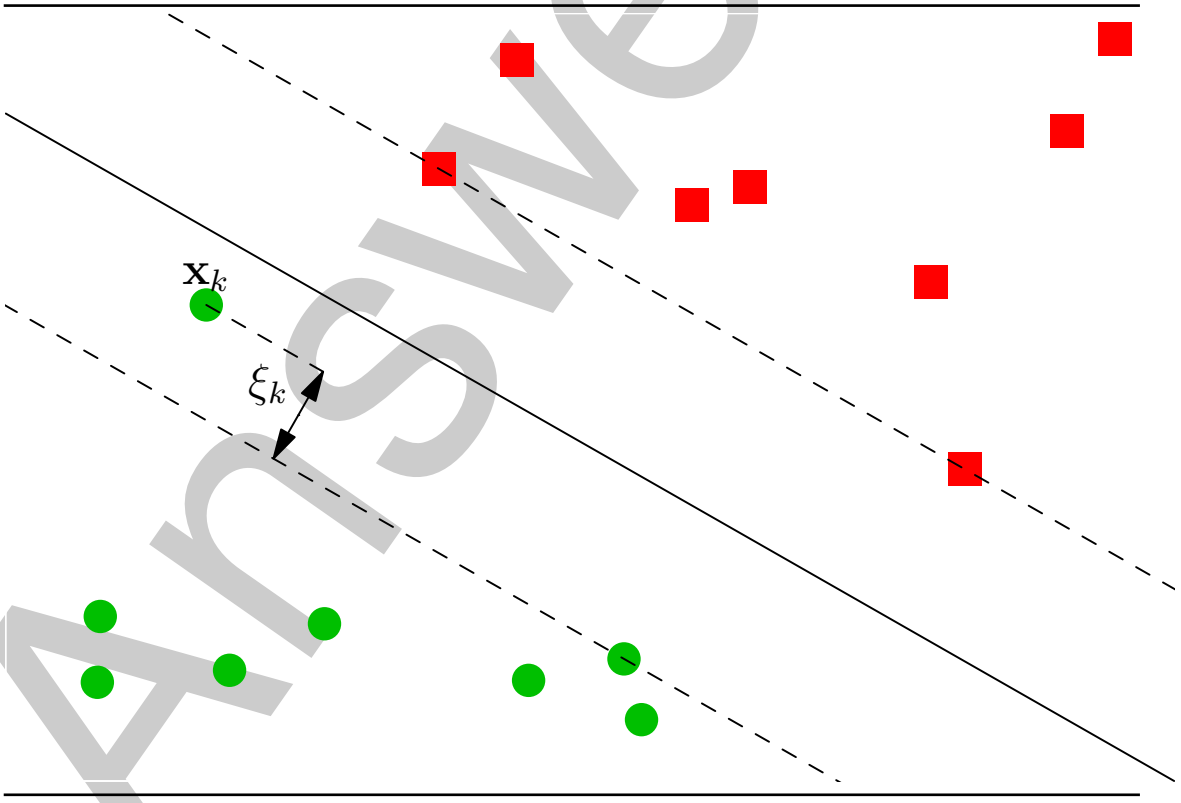
End of question B2

B 3 The Lagrangian for an SVM with constraints $y_k\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_k - b\right) \geq 1$ is given by

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{k=1}^{m} \alpha_k \left(y_k\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_k - b\right) - 1\right).$$

(a) Sketch how slack variables, ξ_k, are introduced to allow some data points to lie within the margins. [5 marks]

(This question tests understanding of the use of slack variables.)



(b) Write down how adding slack variables modifies the constraint on the $\boldsymbol{w}$ and b for an SVM. Also write down the penalty term on the slack variables and the constraint on ξ_k. [5 marks]

The new constraint on w and b is that for all inputs

$$y_k\left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_k - b\right) \geq 1 - \xi_k.$$

The penalty on the slack variables is

$$C\sum_{k=1}^{m} \xi_k$$

where we impose the constraint $\xi_k \geq 0$.

(c) Write down the modified Lagrangian and by computing the saddle-point equation with respect to the slack variables obtain an additional constraint on each Lagrange multiplier, α_k. [10 marks]

(Test ability to work with Lagrangians.)

The new Lagrangian is

$$L = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{k=1}^{m}\xi_k - \sum_{k=1}^{m}\alpha_k\left(y_k\ \left(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_k - b\right) - 1 + \xi_k\right) - \sum_{k=1}^{m}\beta_k\,\xi_k$$

where β_k are Lagrange multipliers to ensure $\xi_i \geq 0$. Note that $\beta_k \geq 0$. The saddle-point equation with respect to ξ_k give

$$\frac{\partial L}{\partial \xi_k} = C - \alpha_k - \beta_k = 0$$

or $\alpha_k = C - \beta_k$. However, since $\beta_k \geq 0$ it follows that the slack variable leads to the additional constraint $\alpha_k \leq C$.

End of question B3

B 4 We have some counts, $\mathcal{D} = (k_1, k_2, \ldots, k_m)$, of a defective protein taken from patients with lung cancer—these measurements are assumed to be independent. We hypothesise that there are in fact two variants of the disease $d \in \{0, 1\}$. We model the counts, k_i, by a Poisson distribution $\mathbb{P}(k_i \mid \mu_d) = \mu_d^{k_i}\,\mathrm{e}^{-\mu_d}/k_i!$ where μ_d takes different values for the two variants of the disease. As we do not know which variant the different patients have we introduce a latent variable $z_i \in \{0, 1\}$ to signify if count i was from patient with disease $d = z_i$.

(a) Write down the likelihood $\mathbb{P}(k_i \,|\, \mu_0, \mu_1, z_i)$. [3 marks]

(This is a problem the students have not seen. Test ability to specify and manipulate probabilities.)

$$\mathbb{P}(k_i \,|\, \mu_0, \mu_1, z_i) = (1 - z_i)\,\frac{\mu_0^{k_i}\,\mathrm{e}^{-\mu_0}}{k_i!} + z_i\,\frac{\mu_1^{k_i}\,\mathrm{e}^{-\mu_1}}{k_i!}$$

(b) Let p_d be the prior probability of a patient with lung cancer to have variant $d \in \{0, 1\}$ (we are assuming $p_0 + p_1 = 1$). Write down the joint probability $\mathbb{P}(k_i, z_i \,|\, \mu_0, \mu_1)$. [2 marks]

$$\mathbb{P}(k_i, z_i \,|\, \mu_0, \mu_1) = \mathbb{P}(k_i \,|\, \mu_0, \mu_1, z_i)\ \mathbb{P}(z_i) = p_0\,(1 - z_i)\,\frac{\mu_0^{k_i}\,\mathrm{e}^{-\mu_0}}{k_i!} + p_1\,z_i\,\frac{\mu_1^{k_i}\,\mathrm{e}^{-\mu_1}}{k_i!}$$

(c) Use Bayes rule to compute $\mathbb{P}(z_i = d \,|\, \mu_0, \mu_1, p_d, k_i)$. [5 marks]

$$\mathbb{P}(z_i = d \,|\, \mu_0, \mu_1, p_d, k_i) = \frac{\mathbb{P}(k_i \,|\, \mu_0, \mu_1, z_i)\ \mathbb{P}(z_i|p_d)}{\mathbb{P}(k_i \,|\, \mu_0, \mu_1, p_0, p_1)} = \frac{p_d\, \mu_d^{k_i}\, \mathrm{e}^{-\mu_d}}{p_0\, \mu_0^{k_i}\, \mathrm{e}^{-\mu_0} + p_1\, \mu_1^{k_i}\, \mathrm{e}^{-\mu_1}}$$

(d) We wish to estimate the set of parameters $\boldsymbol{\theta} = (\mu_0, \mu_1, p_1)$ by maximising the likelihood. Direct maximisation is difficult because the latent parameters will depend on the parameters $\boldsymbol{\theta}$, but their maximum likelihood values will depend on the latent variables. To solve this we use the EM algorithm where we iteratively maximise

$$Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{m} \sum_{z_i \in \{0,1\}} \mathbb{P}\Big(z_i|\boldsymbol{\theta}^{(t)}\Big) \log(\mathbb{P}(k_i, z_i|\mu_0, \mu_1))\,.$$

Let $P_i^{(t)} = \mathbb{P}\big(z_i|\boldsymbol{\theta}^{(t)}\big)$ write down $Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(t)})$ explicitly and compute the new set parameters

$$\boldsymbol{\theta}^{t+1} = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(t)})$$

(that is, compute $\mu_0^{(t+1)}$, $\mu_1^{(t+1)}$ and $p_1^{(t+1)}$). [10 marks]

(Test understanding of EM algorithm)

$$Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{m} \Big(P_0^{(k)} \left(\log(1 - p_1) + k_i\, \log(\mu_0) - \mu_0 - \log(k_i!) \right) + P_1^{(k)} \left(\log(p_1) + k_i\, \log(\mu_1) - \mu_1 - \log(k_i!) \right) \Big).$$

So that the maximum with respect to μ_d is

$$\frac{\partial Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(t)})}{\partial \mu_d} = \sum_{i=1}^{m} P_d^{(k)} \left(\frac{k_i}{\mu_d} - 1 \right) = 0$$

or

$$\mu_d^{(t+1)} = \frac{\sum_{i=1}^{m} P_d^{(k)}\, k_i}{\sum_{i=1}^{m} P_d^{(k)}}$$

and

$$\frac{\partial Q(\boldsymbol{\theta} \,|\, \boldsymbol{\theta}^{(t)})}{\partial p_1} = \sum_{i=1}^{m} \left(-\frac{P_0^{(k)}}{1 - p_1} + \frac{P_1^{(k)}}{p_1} \right) = 0$$

Or

$$p_1^{(t+1)} = \frac{\sum_{i=1}^{m} P_1^{(t)}}{\sum_{i=1}^{m} \left(P_0^{(t)} + P_1^{(t)} \right)}$$

End of question B4

END OF PAPER