LAST (Family) NAME: ______________________________

FIRST (Given) NAME: ______________________________

STUDENT NUMBER: ______________________________

UNIVERSITY OF TORONTO
Faculty of Arts and Science
DECEMBER 2018 EXAMINATIONS
CSC411H1F / CSC2515H1F

Duration — 3 hours
No Aids Allowed

Exam Reminders:

- Fill out your name and student number on the top of this page.
- Do not begin writing the actual exam until the announcements have ended and the Exam Facilitator has started the exam.
- As a student, you help create a fair and inclusive writing environment. If you possess an unauthorized aid during an exam, you may be charged with an academic offence.
- Turn off and place all cell phones, smart watches, electronic devices, and unauthorized study materials in your bag under your desk. If it is left in your pocket, it may be an academic offence.
- When you are done your exam, raise your hand for someone to come and collect your exam. Do not collect your bag and jacket before your exam is handed in.
- If you are feeling ill and unable to finish your exam, please bring it to the attention of an Exam Facilitator so it can be recorded before leaving the exam hall.
- In the event of a fire alarm, do not check your cell phone when escorted outside.

Exam Format and Grading Scheme:

- There are 16 questions, some with multiple parts.
- Longer questions are near the end, but otherwise the questions are not ordered by difficulty. So you should attempt every question.
- The exam is marked out of 35 marks.
- Many questions have more than one right answer.

Students must hand in all examination materials at the end.

.

Q1: _______ / 1

Q2: _______ / 1

Q3: _______ / 2

Q4: _______ / 2

Q5: _______ / 2

Q6: _______ / 2

Q7: _______ / 1

Q8: _______ / 2

Q9: _______ / 2

Q10: _______ / 3

Q11: _______ / 3

Q12: _______ / 2

Q13: _______ / 1

Q14: _______ / 3

Q15: _______ / 2

Q16: _______ / 6

Final mark: _______ / 35

1. [**1pts**] A common preprocessing step of many learning algorithms is to normalize each feature to be zero mean and unit variance. Give the formula for the normalized feature $\tilde{x}_j$ as a function of the original feature x_j and the mean μ_j and standard deviation σ_j of that feature. You don't need to justify your answer.

2. [**1pt**] We showed that each step of K-means reduces a particular cost function. What is that cost function? You can give a formula or explain it in words. You don't need to justify your answer.

3. [**2pts**] Suppose your classifier achieves poor accuracy on both the training and test sets. Which would be a better choice to try to improve the performance: bagging or boosting? Justify your answer.

4. **[2pts]** Bayesian linear regression and Gaussian processes are two different approaches to Bayesian regression.

   (a) **[1pt]** Give one situation in which Gaussian process regression would be more *computationally efficient* than Bayesian linear regression.

   (b) **[1pt]** Give one other advantage of Gaussian process regression over Bayesian linear regression.

5. **[2pts]** We showed that AdaBoost can be viewed as minimizing the exponential loss.

   (a) **[1pt]** Give the definition of exponential loss. (You don't need to provide any justification.)

   $$\mathcal{L}_{\mathrm{E}}(y, t) =$$

   (b) **[1pt]** TRUE or FALSE: there is some value ε such that if the sum of the exponential loss on all the training examples is less than ε, then all the training examples are classified correctly. Justify your answer.

6. **[2pts]** Suppose you are running AdaBoost with 4 training examples. At the start of the current iteration, the four examples have the weights shown in the following table. Another column says if the weak classifier got them correct or incorrect. Determine the *new* weights for these four examples, and fill in the corresponding entries in the table.

   You do not need to justify your answer or explain your reasoning, although doing so may help you obtain partial credit.

	Old Weight	Correcct?	New Weight
Example 1	0.16	Correct	
Example 2	0.64	Correct	
Example 3	0.08	Incorrect	
Example 4	0.12	Incorrect	

Hint: this question doesn't require much calculation. Observe that:

- *the weights for certain pairs of examples will be updated by the same multiplicative factor*
- *you know something about the sum of weights for certain sets of examples.*

7. **[1pt]** Recall two linear classification methods we considered:

Model 1:

$$y = \mathbf{w}^\top \mathbf{x} + b$$
$$\mathcal{L}_{\mathrm{SE}}(y, t) = \tfrac{1}{2}(y - t)^2$$

Model 2:

$$z = \mathbf{w}^\top \mathbf{x} + b$$
$$y = \sigma(z)$$
$$\mathcal{L}_{\mathrm{SE}}(y, t) = \tfrac{1}{2}(y - t)^2$$

Here, σ denotes the logistic function, and the targets t take values in $\{0, 1\}$. Briefly explain our reason for preferring Model 2 to Model 1.

8. **[2pts]** Consider a discounted Markov decision process (MDP) with discount parameter γ. It has a transition distribution $\mathcal{P}(\cdot \mid s, a)$ and deterministic reward function $r(s, a)$. The agent's policy is a deterministic function $\pi : \mathcal{S} \to \mathcal{A}$.

(a) **[1pt]** Give the definition of the state-action value function Q^π for a policy π. It should be given in terms of γ and the immediate rewards $R_t = r(S_t, A_t)$ for $t = 0, \ldots, \infty$. You don't need to justify your answer.

$$Q^\pi(s, a) =$$

(b) **[1pt]** Give the Bellman recurrence for Q^π, i.e. the formula expressing $Q^\pi(s, a)$ in terms of an expectation over successor states. You don't need to justify your answer.

$$Q^\pi(s, a) =$$

9. [**2pts**] Consider the following NumPy code for computing cross-entropy loss.

```
def cross_entropy_loss(z, t):
    y = 1 / (1 + np.exp(-z))
    return -t * np.log(y) - (1-t) * np.log(1-y)
```

The formulas for y and $\mathcal{L}$ are correct, but there's something wrong with this code.

(a) [**1pt**] What is wrong with the code? *Hint: what happens when z is large?*

(b) [**1pt**] Provide NumPy code implementing `cross_entropy_loss` which doesn't have this problem. You may want to use the function `np.logaddexp`, which takes two arguments a and b and returns $\log(e^a + e^b)$.

10. **[3pts]** We showed that the Support Vector Machine (SVM) can be viewed as minimizing hinge loss:

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} \mathcal{L}_{\mathrm{H}}(y,t) + \frac{1}{2\gamma}\|\mathbf{w}\|_2^2$$

where hinge loss is defined as:

$$\mathcal{L}_{\mathrm{H}}(y,t) = \max(0, 1 - ty)$$

(a) **[1pt]** TRUE or FALSE: if the total hinge loss is zero, then every training example must be classified correctly. Justify your answer.

(b) **[1pt]** TRUE or FALSE: if the dataset is linearly separable, then the optimal soft-margin SVM weights (according to the above objective) must classify every training example correctly. Justify your answer.

(c) **[1pt]** Suppose we replace the hinge loss with the following:

$$\mathcal{L}(y,t) = \max(0, -ty)$$

and otherwise keep the soft-margin SVM objective the same. What would go wrong?

11. [**3pts**] The Laplace distribution, parameterized by μ and β, is defined as follows:

$$\text{Laplace}(w; \mu, \beta) = \frac{1}{2\beta} \exp\left(-\frac{|w - \mu|}{\beta}\right).$$

Consider a variant of Bayesian linear regression where we assume the prior over the weights $\mathbf{w}$ consists of an independent zero-centered Laplace distribution for each dimension, with shared parameter β:

$$\begin{aligned} w_j &\sim \text{Laplace}(0, \beta) \\ t \,|\, \mathbf{w} &\sim \mathcal{N}(t; \mathbf{w}^\top \boldsymbol{\psi}(\mathbf{x}), \sigma) \end{aligned}$$

For reference, the Gaussian PDF is:

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

(a) [**2pts**] Suppose you have a labeled training set $\{(\mathbf{x}^{(i)}, t^{(i)})\}_{i=1}^N$. Give the cost function you would minimize to find the MAP estimate of $\mathbf{w}$. (It should be expressed in terms of mathematical operations.) You don't need to justify your answer, but doing so may help you earn partial credit.

(b) [**1pt**] Based on your answer to part (a), how might the MAP solution for a Laplace prior differ from the MAP solution if you use a Gaussian prior?

12. **[2pts]** Consider one layer of a multilayer perceptron (MLP), whose computations are defined as follows:

$$z_i = \sum_j w_{ij} h_j + b_i$$
$$y_i = \phi(z_i),$$

where ϕ is a nonlinear activation function, h_j denotes the input to this layer (i.e. the previous layer's hidden units), and y_i denotes the output of this layer.

Give the backprop rules for $\overline{z_i}$, $\overline{h_j}$ and $\overline{w_{ij}}$ in terms of the error signal $\overline{y_i}$. You can use ϕ' to denote the derivative of ϕ. You don't need to show your work.

$\overline{z_i} =$

$\overline{h_j} =$

$\overline{w_{ij}} =$

13. **[1pt]** Recall that the beta distribution is defined by

$$\text{Beta}(\theta; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \theta^{a-1}(1-\theta)^{b-1},$$

where Γ is the gamma function. Give values of a and b such that the distribution is highly concentrated around $\theta = 0.75$. You don't need to justify your answer.

Hint: If you've forgotten the shape of the distribution, you can find the mode as a function of a and b by differentiating the log density.

$a =$

$b =$

14. **[3pts]** Recall that the optimal PCA subspace can be determined from the eigendecomposition of the empirical covariance matrix $\mathbf{\Sigma} = \mathrm{Cov}(\mathbf{x})$. Also recall that the eigendecomposition can be expressed in terms of the following spectral decomposition of $\mathbf{\Sigma}$:

$$\mathbf{\Sigma} = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top,$$

where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{\Lambda}$ is a digonal matrix. Assume the eigenalues are sorted from largest to smallest. You may assume all of the eigenvalues are distinct.

(a) **[1pt]** If you've already computed the eigendecomposition (i.e. $\mathbf{Q}$ and $\mathbf{\Lambda}$), how do you obtain the orthogonal basis $\mathbf{U}$ for the optimal PCA subspace? (You do not need to justify your answer.)

(b) **[2pts]** The PCA code vector for a data point $\mathbf{x}$ is given by $\mathbf{z} = \mathbf{U}^\top(\mathbf{x} - \boldsymbol{\mu})$. Show that the dimensions of $\mathbf{z}$ are uncorrelated. (Hint: start by finding a formula for $\mathrm{Cov}(\mathbf{z})$.)

15. **[2pts]** Recall that Gaussian discriminant analysis (GDA) can have very different decision boundary shapes depending on the precise model assumptions. Consider a GDA model with two classes, and where the covariance is shared between both classes and is spherical. Show mathematically that the decision boundary is linear.

    For reference, the multivariate Gaussian PDF is given by:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

16. **[6pts]** In this question, you will derive the E-M update rules for a univariate Gaussian mixture model (GMM) with two mixture components. Unlike the GMMs we covered in the course, the mean μ will be shared between the two mixture components, but each component will have its own standard deviation σ_k. The mixture component is indicated by a latent variable $z \in \{0, 1\}$. The model is defined as follows:

$$\begin{aligned} z &\sim \text{Bernoulli}(\theta) \\ x \,|\, z = k &\sim \mathcal{N}(\mu, \sigma_k) \qquad \text{for } k \in \{0, 1\} \end{aligned}$$

The parameters of the model are θ, μ, σ_0, and σ_1.

For reference, the PDF of the Gaussian distribution is as follows:

$$\mathcal{N}(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right)$$

(a) **[1pt]** Write the density defined by this model (i.e. the probability of x, with z marginalized out):

$p(x) =$

(b) **[1pt]** In the E-step, for each data point $x^{(i)}$, we need to compute the posterior probability $r^{(i)} = \Pr(z^{(i)} = 1 \,|\, x^{(i)})$. Give the formula for $r^{(i)}$. In your formula, you may use $\mathcal{N}(x^{(i)}; \mu, \sigma)$ to denote the Gaussian PDF, rather than writing it out explicitly. You do not need to justify your answer, but doing so may help you earn partial credit.

$r^{(i)} =$

Question 16 continued on next page ⟶

(Question 16, cont'd)

(c) **[1pt]** Write out the objective function that is to be maximized in the M-step. It should be expressed in terms of the $r^{(i)}$ and the Gaussian PDF $\mathcal{N}(x^{(i)}; \mu, \sigma)$. You do not need to jutsify your answer.

(d) **[2pts]** Derive the M-step update rule for μ by maximizing this objective with respect to μ. (In this step, the σ_k are fixed to their previous values.)

Question 16 continued on next page ⟶

(Question 16, cont'd)

(e) **[1pt]** Derive the M-step update rule for σ_1 by maximizing the objective with respect to σ_1. (In this step, assume μ is fixed to its previous value.)

Note: Even though this part has several steps, it's only worth one point. So you may want to come back to it once you're done with the rest of the exam.

(Scratch work or continued answers)

(Scratch work or continued answers)