ECE521 Midterm Examination

University of Toronto Faculty of Applied Science and Engineering
Department of Electrical and Computer Engineering

16 February 2017, corrected version

Name: _____

Student Number: _____ **Section:** L0101 or L0102 *(circle one)*

Instructions:

- Time allowed: 90 minutes
- Answer all questions. Page 8 has space for overflow
- Any questions completed in pencil rather than pen may not be eligible to be remarked even if there was a marking error
- Aids allowed: You are allowed to bring in one $8.5'' \times 11''$ aid sheet double-sided, and a non-programmable calculator
- Some useful formulas:

  Gaussian pdf: $P(\mathbf{x}) = (2\pi)^{-d/2}|\Sigma|^{-1/2}e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$

  $\tanh(x) = \frac{e^{2x}-1}{e^{2x}+1} = \frac{e^x-e^{-x}}{e^x+e^{-x}}$

Part	Value	Mark
I	16	
II	32	
III	14	
IV	22	
V	9	
Total	93	

This test should have 8 pages including this page

Part I: True or false

Circle one answer (2 marks each, 16 marks total).

1.	True	False	For $k = 1$, k-NN *classifiers* can achieve 100% accuracy on the training set, therefore implying that choosing $k = 1$ tends to produce the best model.
2.	True	False	In a regression dataset with extreme outliers in the training targets, it is better to use the ℓ_1 loss function than squared ℓ_2 loss.
3.	True	False	Linear regression assumes that the input features $\mathbf{x}$ follow a Gaussian distribution.
4.	True	False	Adding ℓ_2 regularization on the weights in our training loss function is equivalent to maximizing *a posteriori* under a Gaussian prior.
5.	True	False	For a 2-D multivariate normal with a diagonal covariance matrix, its PDF contours must be circles.
6.	True	False	When learning a logistic regression model using gradient descent, generally it is helpful to centre (zero-mean) the input features.
7.	True	False	When using a k-NN model based on squared ℓ_2 distance, generally it's helpful to centre the input features
8.	True	False	In a multi-layered neural network, if the activation of a hidden unit is zero, then the gradients of the weights of all of its *incoming* connections are zero.

Part II: Multiple Choice

For each question, circle one answer. (4 marks each, 32 marks total): If you select "I don't know", you will receive 2 marks, whereas if you select another answer and get it wrong, you will receive 0 marks.

1. Consider learning a k-NN classifier for a binary classification task using a training set of 100 examples: 80 of one class and 20 of the other class. Each data point has its own unique input vector. What is the classification accuracy on the *training set* when k = 3?:

(a) 100%

(b) 80%

(c) 20%

(d) It depends

(e) I don't know.

2. Consider learning a k-NN classifier for a binary classification task using a training set of 100 examples: 80 of one class and 20 of the other class. Each data point has its own unique input vector. What is the classification accuracy on the *averaged 5-fold cross validation* when k = 3?:

(a) 100%

(b) 80%

(c) 20%

(d) It depends

(e) I don't know.

3. Which of the following techniques do not improve a neural network's test performance?

(a) Early stopping

(b) Adding momentum to stochastic gradient descent (SGD)

(c) Dropout

(d) Tuning the weight-decay coefficient using k-fold cross-validation

(e) I don't know.

4. Training a convolutional neural network for object classification, you find that performance on the training set is good while the performance on the validation set is unacceptably low. A reasonable fix might be to:

(a) Decrease the weight decay

(b) Reduce the training set size

(c) Reduce the number of layers and neurons

(d) Increase the number of layers and neurons

(e) I don't know.

5. Why does logistic regression generally perform better than linear regression for classification?

(a) Logistic regression is easier to optimize since its cost function is convex

(b) Unlike logistic regression, linear regression is highly sensitive to outliers

(c) Logistic regression finds a nonlinear decision boundary and thus is more flexible

(d) All of the above

(e) I don't know.

6. You are given the following set of input-output pairs:

$$(\mathbf{x} = [1, 1], y = 1), \quad (\mathbf{x} = [-1, 1], y = -1), \quad (\mathbf{x} = [-1, -1], y = 1), \quad (\mathbf{x} = [1, -1], y = -1)$$

We want to find a function $f(\mathbf{x})$ to estimate y. Which of the following is true?

(a) A linear function, $f = \mathbf{w}^\top \mathbf{x}$, can be used to approximate y accurately

(b) No linear model will be accurate; a nonlinear basis function is required

(c) No linear or nonlinear models can be used to approximate $f(\mathbf{x})$ perfectly

(d) It depends

(e) I don't know.

7. A single-hidden-layer neural network uses ReLU activation functions and all the bias units are zero. If the weight matrix between the input and the hidden layer is doubled (scaled up by a factor of two) and the weight matrix between the hidden layer and output is halved, what happens to the output of the neural network in general?

(a) Doubled

(b) Stays the same

(c) Scaled up by a factor between 1 and 2

(d) It depends

(e) I don't know.

8. At a certain electronics factory, in a typical day's output 10% percent of resistors are bad, and the rest are good. Good resistors have an 80% chance of passing the factory's test, and bad resistors have a 30% chance of passing the test. Suppose the factory tests each resistor three times. If a particular resistor passes the test 2 times out of 3, what are the chances it is good?

(a) ~ 0.902

(b) 0.8

(c) ~ 0.975

(d) None of the above

(e) I don't know.

Part III: Classification

1. (5 marks) In binary classification, suppose $t_m \in \{-1, 1\}$ instead of the usual $\{0, 1\}$. The cross-entropy loss function here is:

$$\mathcal{L} = -\frac{1}{2} \sum_{m=1}^{M} \left\{ \ln \left[\sigma(\mathbf{w}^T \mathbf{x}_m) \right] (t_m + 1) + \ln \left[1 - \sigma(\mathbf{w}^T \mathbf{x}_m) \right] (1 - t_m) \right\}$$

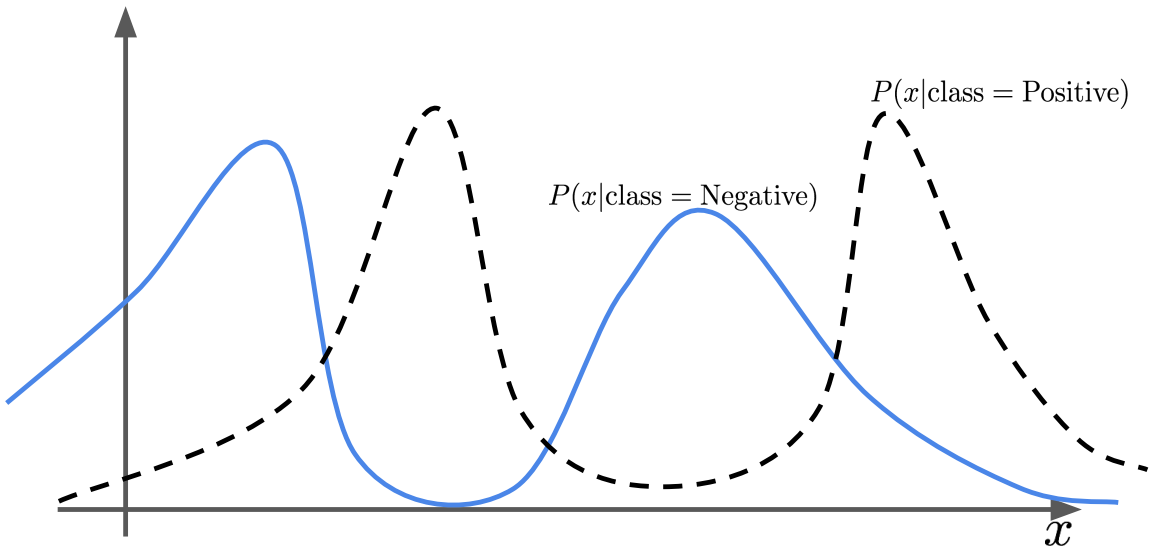
where M is the number of observations of input-target pairs $(\mathbf{x}_m, t_m)$, and $\sigma(\cdot)$ is the sigmoid function. Note that the first term is zero when $t_m = -1$ and the second term is zero when $t_m = 1$. Show that we can rewrite the above loss function as the equivalent $\mathcal{L}' = \mathcal{L}$, where:

$$\mathcal{L}' = \sum_{m=1}^{M} \ln \left[1 + \exp(-t_m \mathbf{w}^T \mathbf{x}_m) \right].$$

2. (3 marks) Consider the two formulations of cross-entropy loss stated in part [a]. The equivalent, $\mathcal{L}'$, is simpler to compute. Briefly comment on its numerical stability compared to that of $\mathcal{L}$.

3. (6 marks) Consider a binary classification task in which the training data are equally split between a positive class and a negative class. We have estimated the class conditional distribution of the input feature x and two conditional pdfs are shown in the figure below. Draw the decision boundaries on the figure to partition the x-axis and annotate the partitions to indicate the assigned class labels in the figure such that the expected loss is minimized under the following loss matrix:

$$L = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$



Part IV: Neural networks

1. (4 marks) When designing neural networks, a problem with the sigmoid function's derivative is that it is very small when its input $z \gg 0$ or $z \ll 0$. Suppose we decide to reparameterize a neural network using $\tanh(z)$ instead of $\sigma(z)$ as the activation functions.

$$\tanh(z) = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Write down the expression for $\frac{\partial \tanh(z)}{\partial z}$ in terms of $\tanh(z)$ itself (similar to how we express the partial derivative of the sigmoid function) and comment on whether it exhibits a similar problem to the sigmoid function.

2. (8 marks) Consider the binary threshold neuron, $h = \text{sgn}(\mathbf{w}^T\mathbf{x})$, defined such that $h \in \{0, 1\}$, with no bias b or w_0. Consider the following set of four input features, $\mathbf{x}$:

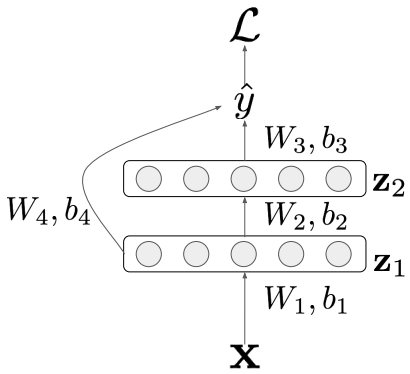
$$(1, 0, 0)^T, \quad (0, 1, 0)^T, \quad (0, 0, 1)^T, \quad (1, 1, 1)^T$$

(a) (2 marks) Find a three-dimensional parameter vector $\mathbf{w}$ such that the neuron will have the output pattern $\{h\} = \{1, 1, 1, 1\}$ for the given four input features.

(b) (2 marks) Find a three-dimensional parameter vector $\mathbf{w}$ such that the neuron will have the output pattern $\{h\} = \{1, 1, 0, 0\}$ for the given four input features.

(c) (4 marks) Find an unrealizable output pattern $\{h\}$.

3. (10 marks) Consider the following figure illustrating a fully connected multi-layer neural network that has a linear output $\hat{y}$ and some additional skip-layer connections with the weight matrix W_4 and the bias b_4. Assume the network uses the hidden activation function, $\phi(\cdot)$. The column vector $\mathbf{z}$ represents the weighted sum of the inputs at a hidden layer. Also, the weight matrices follow the convention that the number of rows in a weight matrix corresponds to the number of output units, so $W \in \mathbb{R}^{(\#\text{output\_units}) \times (\#\text{input\_units})}$



(a) (2 marks) For a single training case, express $\hat{y}$ in terms of the symbols shown in the above figure. Do so using vector notation; i.e. your answer should not contain explicit summations.

(b) (4 marks) For a single training case, derive the gradient of the loss function $\mathcal{L}$ w.r.t. W_4 using vector notation in terms of the symbols shown in the above figure and their partial derivatives:

(c) (4 marks) For a single training case, derive the expression for the partial derivative $\frac{\partial \mathcal{L}}{\partial \mathbf{z}_1}$ using vector notation in terms of the symbols shown in the above figure and their partial derivatives. You may use $\frac{\partial \phi(\mathbf{z}_1)}{\partial \mathbf{z}_1}$ to denote a diagonal matrix whose ith diagonal term is $\frac{\partial \phi(z_{1i})}{\partial z_{1i}}$.

Part V: Linear basis function models

Consider the regression function

$$f(x, \mathbf{w}) = \mathbf{w}^T \phi(x) = w_0 + w_1\phi_1(x) + w_2\phi_2(x) + ... + w_M\phi_M(x)$$

where x is a scalar and f a scalar-valued function. We want to apply this to two possible data sources:

(i) A polynomial of second degree.

(ii) A periodic source which oscillates with a known period L.

1. (6 marks) Without limiting yourself to functions discussed in class necessarily, what might comprise suitable basis functions, for (i)? For (ii)?

2. (3 marks) Explain whether or not it's possible to save time and design a single set of basis functions $\phi_i(x)$ that allows you to model observations from either source.

If you use the below space to answer an earlier question, you must indicate so near the question itself.