Midterm for CSC321, Intro to Neural Networks
Winter 2017, afternoon section
Tuesday, Feb. 28, 1:10-2pm

Name: ______________________________

Student number: ____________________

This is a closed-book test. It is marked out of 15 marks. Please answer ALL of the questions. Here is some advice:

- The questions are NOT arranged in order of difficulty, so you should attempt every question.
- Questions that ask you to "briefly explain" something only require short (1-3 sentence) explanations. Don't write a full page of text. We're just looking for the main idea.
- None of the questions require long derivations. If you find yourself plugging through lots of equations, consider giving less detail or moving on to the next question.
- Many questions have more than one right answer.

Final mark: ______ / 15

1. **[1pt]** Carla tells you, “Overfitting is bad, so you want to make sure your model is simple enough that the test error is no higher than the training error.” Is she right or wrong? Justify your answer.

   Note that there are good arguments for either side, so you will receive full credit as long as you justify your answer well.

2. **[1pt]** Suppose you are training a neural net using stochastic gradient descent (SGD), and you compute the cost function on the entire training set after each update. TRUE or FALSE: if you ever see the training cost increase after an SGD update, that means your learning rate is too large. Justify your answer.

3. Suppose you are given the following two-dimensional dataset for a binary classification task:

x_1	x_2	t
0	0	1
1	1	1
1	0	0

You use a linear model with a hard threshold activation function, and a dummy dimension x_0 so that w_0 functions as a bias. You would like to find a weight vector in the strictly feasible region, i.e. none of the training examples should lie on the decision boundary.

(a) **[1pt]** Each of the training examples gives a constraint on w_0, w_1, and w_2. Write down all three constraints.

(b) **[1pt]** Find a set of weights which satisfies all the constraints. You do not need to show your work or justify your answer. *Hint: pick w_0 first, then w_1, then w_2.*

4. **[2pts]** In Homework 5, we analyzed dropout for a linear regression model; the predictions had the form

$$y = \sum_j m_j w_j x_j,$$

where the m_j's were i.i.d. Bernoulli random variables. Let's modify the dropout algorithm so that the m_j's take on continuous values, and they are i.i.d. Gaussian random variables with mean 1 and variance σ^2.

Under this model, determine the variance of the predictions, $\text{Var}[y]$, as a function of the x_j's and w_j's. Show your work. *Hint: use the properties of variance.*

5. **[2pts]** Briefly explain the difference between invariant and equivariant feature detectors. Give an example of an equivariant operation.

6. Recall that in the domain of binary classification, the logistic activation function combined with cross-entropy loss does not suffer from saturated units. Now let's suppose we don't like making overly confident predictions, so we transform the predictions y to lie in the interval $[0.1, 0.9]$. In other words, we take

$$y = 0.8\,\sigma(z) + 0.1,$$

where σ denotes the logistic function. We still use cross-entropy loss.

(a) **[1pt]** For a positive training example, sketch the cross-entropy loss as a function of z. Your sketch doesn't need to be precise, but it should make clear the asymptotic behavior as $z \to \pm\infty$. (You should label any asymptote lines.)

(b) **[1pt]** Based on your answer to Part (a), will gradient descent on this model suffer from saturation when the predictions are very wrong? Why or why not?

7. Suppose we somehow know that the weights for a two-dimensional regression problem should lie on the unit circle. We can parameterize the weights in terms of the angle θ, i.e. let $(w_1, w_2) = (\cos\theta, \sin\theta)$. The model and loss function are as follows:

$$\begin{aligned} w_1 &= \cos\theta \\ w_2 &= \sin\theta \\ y &= w_1 x_1 + w_2 x_2 \\ \mathcal{L} &= \frac{1}{2}(y - t)^2 \end{aligned}$$

   (a) **[1pt]** Draw the computation graph relating θ, w_1, w_2, y, and $\mathcal{L}$.

   (b) **[2pts]** Determine the backprop update rules which let you compute the derivative $\mathrm{d}\mathcal{L}/\mathrm{d}\theta$.

   Your equations should refer to previously computed values (e.g. your formula for $\overline{z}$ should be a function of $\overline{y}$). You do not need to show your work, but it may help you get partial credit. The first two steps have been filled in for you.

$$\overline{\mathcal{L}} = 1$$

$$\overline{y} = \overline{\mathcal{L}} \cdot (y - t)$$

$$\overline{w_1} =$$

$$\overline{w_2} =$$

$$\overline{\theta} =$$

8. **[2pts]** Suppose you are given a two-dimensional linear regression problem (with no bias parameter), using the following dataset:

x_1	x_2	t
83.1	-82.4	3.3
83.2	-82.8	1.5
83.5	-82.1	2.0
$\vdots$	$\vdots$	$\vdots$

Circle the contour plot which best represents the cost function for this regression problem. Justify your answer.