CSCE 5063 Fall 2018 Final exam

1. (10 points) What is your name?

2. (5 points) Which of these is the most reasonable way to detect convergence?
   - Perform *m*-repetitions of *n*-fold cross-validation at periodic intervals during training. When it stops improving, convergence is detected.
   - When the magnitude of the gradient vector becomes negative, it has converged.
   - Detect when L^1 regularization begins adding instead of subtracting from the weights.
   - Measure accuracy with a validation set at periodic intervals during training. Stop when it fails to improve by a certain amount.
   - Measure accuracy on both training and test data. When they separate, overfit has started, which indicates convergence.

3. (5 points) What would happen if you only train the last two layers (on the output end) of a deep neural network? (Circle one)
   - It would probably not learn unless all the weights in the neural network were trained.
   - It would probably learn, but not as well as a model that had only two layers.
   - It would probably work better than a model with just two layers, but probably not as well as a deep network in which all weights were trained.
   - It would probably work better than training a whole deep neural network.
   - It would probably become self-aware and take over the world! (Hint: don't pick this one.)

4. (5 points) According to machine learning theory, in a universe where all possible hypotheses are equally probable,
   - learning is possible.
   - learning is impossible.

5. (6 points) Machine learning works best with
   - under-constrained problems
   - over-constrained problems

6. (5 points) Circle the TWO correct statements:
   - Batch gradient descent guarantees to reduce loss (error) with every step.
   - Stochastic gradient descent guarantees to reduce loss (error) with every step.
   - Stochastic gradient descent is usually faster than batch gradient descent.
   - Batch gradient descent is usually faster than stochastic gradient descent.

7. (6 points) Circle all statements that are true regarding autoencoders:
   - They implement unsupervised learning.
   - They produce an internal representation of observed values.
   - They provide a way to approximate momentum without having a momentum term.
   - They can be trained with stochastic gradient descent.
   - They are neural networks with no weights.
   - They always have convolutional layers.

8. (5 points) If you have two different prediction tasks in related domains, which is likely to make better generalizing predictions? (assume you can tune meta-parameters as needed.)
   - Train separate learning models for each of the two prediction tasks.
   - Train one model to perform both tasks simultaneously.

9. (15 points) Given a univariate non-linear model, $h(x) = f(mx + b)$, and the objective function $e = (y_i - h(x_i))^2$, if you present the training pattern $x_1 \rightarrow y_1$ to update this model by stochastic gradient descent with a learning rate of η (and no momentum), what will be the updated values for the parameters m and b? (Note that the derivative of $f(x)$ can be expressed as $f'(x)$.)

10. (11 points) Continuing with the previous problem, give an expression for updating x by stochastic gradient descent.

11. (15 points) Consider a model with just one convolutional layer. This layer has just one 1-dimensional filter with two elements initialized with the values $\langle 0.1, 0.2 \rangle$, and a bias term initialized to 0. If the training pattern $\langle 0, 1, 2 \rangle \rightarrow \langle 0.3, 0.6 \rangle$ is presented to this model for one iteration of training by stochastic gradient descent with a learning rate of 0.1, what will the updated weights and bias values be?

12. (12 points) Please repeat the previous problem, but this time also apply L^1 regularization with a regularization term of $\lambda = 0.2$.