WEEK 10 QUIZ 1

Large Scale Machine Learning

TOTAL POINTS 5

1. Suppose you are training a logistic regression classifier using stochastic gradient descent. You find that the cost (say, $cost(\theta, (x^{(i)}, y^{(i)}))$, averaged over the last 500 examples), plotted as a function of the number of iterations, is slowly increasing over time. Which of the following changes are likely to help?

   - (●) Try using a smaller learning rate α.
   - ( ) Try averaging the cost over a larger number of examples (say 1000 examples instead of 500) in the plot.
   - ( ) Try using a larger learning rate α.
   - ( ) This is not an issue, as we expect this to occur with stochastic gradient descent.

2. Which of the following statements about stochastic gradient descent are true? Check all that apply.

- [ ] In order to make sure stochastic gradient descent is converging, we typically compute $J_{\text{train}}(\theta)$ after each iteration (and plot it) in order to make sure that the cost function is generally decreasing.
- [x] Before running stochastic gradient descent, you should randomly shuffle (reorder) the training set.
- [x] You can use the method of numerical gradient checking to verify that your stochastic gradient descent implementation is bug-free. (One step of stochastic gradient descent computes the partial derivative $\frac{\partial}{\partial \theta_j} cost(\theta, (x^{(i)}, y^{(i)}))$.)
- [ ] Suppose you are using stochastic gradient descent to train a linear regression classifier. The cost function $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$ is guaranteed to decrease after every iteration of the stochastic gradient descent algorithm.

3. Which of the following statements about online learning are true? Check all that apply.

- [ ] One of the disadvantages of online learning is that it requires a large amount of computer memory/disk space to store all the training examples we have seen.

- [x] When using online learning, in each step we get a new example (x, y), perform one step of (essentially stochastic gradient descent) learning on that example, and then discard that example and move on to the next.

- [ ] One of the advantages of online learning is that there is no need to pick a learning rate α.

- [x] In the approach to online learning discussed in the lecture video, we repeatedly get a single training example, take one step of stochastic gradient descent using that example, and then move on to the next example.

4. Assuming that you have a very large training set, which of the following algorithms do you think can be parallelized using map-reduce and splitting the training set across different machines? Check all that apply.

- [x] Linear regression trained using batch gradient descent.
- [ ] An online learning setting, where you repeatedly get a single example (x, y), and want to learn from that single example before moving on.
- [x] A neural network trained using batch gradient descent.
- [ ] Logistic regression trained using stochastic gradient descent.

---

5. Which of the following statements about map-reduce are true? Check all that apply.

- [ ] Linear regression and logistic regression can be parallelized using map-reduce, but not neural network training.

- [x] Because of network latency and other overhead associated with map-reduce, if we run map-reduce using N computers, we might get less than an N-fold speedup compared to using 1 computer.

- [ ] When using map-reduce with gradient descent, we usually use a single machine that accumulates the gradients from each of the map-reduce machines, in order to compute the parameter update for that iteration.

- [x] If you have only 1 computer with 1 computing core, then map-reduce is unlikely to help.
