Feedback — XVII. Large Scale Machine Learning

Help

You submitted this quiz on **Sat 4 Jan 2014 9:24 PM PST**. You got a score of **4.75** out of **5.00**. You can attempt again in 10 minutes.

Question 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?

Your Answer		Score	Explanation
This is not an issue, as we expect this to occur with stochastic gradient descent.			
Try using a larger learning rate α.			
Try halving (decreasing) the learning rate α, and see if that causes the cost to now consistently go down; and if not, keep halving it until it does.	✔	1.00	Such a plot indicates that the algorithm is diverging. Decreasing the learning rate α means that each iteration of stochastic gradient descent will take a smaller step, thus it will likely converge instead of diverging.
This is not possible with stochastic gradient descent, as it is guaranteed to converge to the optimal parameters θ.			
Total		1.00 /	

Total	1.00 / 1.00	

Question 2

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

Your Answer		Score	Explanation
☑ Before running stochastic gradient descent, you should randomly shuffle (reorder) the training set.	✔	0.25	It is a good idea to shuffle your data so that gradient descent does not take a long sequence of steps based on a biased subset of the data (such as a long run of $y = 0$ examples in logistic regression).
☐ One of the advantages of stochastic gradient descent is that it uses parallelization and thus runs much faster than batch gradient descent.	✔	0.25	Stochastic gradient descent still runs in series, one example at a time.
☐ 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.	✔	0.25	Since each iteration of stochastic gradient descent takes into account only one training example, it is not guaranteed that every update lowers the cost function over the entire training set.
☑ In each iteration of stochastic gradient descent, the algorithm needs to examine/use only one training example.	✔	0.25	Every iteration updates the parameters based on the cost of only one example, $cost(\theta, (x^{(i)}, y^{(i)}))$.
Total		1.00 / 1.00	

Question 3

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

Your Answer		Score	Explanation
☐ Online learning algorithms are most appropriate when we have a fixed training set of size m that we want to train on.	✔	0.25	It is the opposite: they are most appropriate when we have a stream of training data of unbounded size.
☑ 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.	✔	0.25	This is one good approach to online learning discussed in the lecture video.
☑ Online learning algorithms are usually best suited to problems were we have a continuous/non-stop stream of data that we want to learn from.	✔	0.25	Such a stream of data is well-suited to online learning because online learning does not save old training examples, but instead uses them once and then throws them out.
☐ 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.	✔	0.25	Since online learning algorithms do not save old examples, they can be very efficent in terms of computer memory and disk space.
Total		1.00 / 1.00	

Question 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.

Your Answer		Score	Explanation
☐ Logistic regression trained using stochastic gradient descent.	✔	0.25	Since stochastic gradient descent processes one example at a time and updates the parameter values after each, it cannot be easily parallelized.
☑ Computing the average of all the features in your training set $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$ (say in order to perform mean normalization).	✔	0.25	You can split the dataset into N smaller batches, compute the feature average of each smaller batch on one of N separate computers, and then average those results on a central computer to get the final result.
☑ A neural network trained using batch gradient descent.	✔	0.25	You can split the dataset into N smaller batches, compute the gradient for each smaller batch on one of N separate computers, and then average those gradients on a central computer to use for the gradient update.
☐ Linear regression trained using stochastic gradient descent.	✔	0.25	Since stochastic gradient descent processes one example at a time and updates the parameter values after each, it cannot be easily parallelized.
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☐ Running map-reduce over N computers requires that we split the training set into N^2 pieces.	✔	0.25	Usually, you will split the data into N pieces, but map-reduce does not require a specific division of the data.
☑ In order to parellelize a learning algorithm using map-reduce, the first step is to figure out how to express the main work done by the algorithm as computing sums of functions of training examples.	✔	0.25	In the reduce step of map-reduce, we sum together the results computed by many computers on the training data.
☐ 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.	✘	0.00	Such a setup allows us to use many computers to do the hard work of gradient computation while making the parameter update simple, as it occurs in one place.
☑ 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.	✔	0.25	The maximum speedup possible is N-fold, and it is unlikely you will get an N-fold speedup because of the overhead.
Total		0.75 / 1.00	