Feedback — XVII. Large Scale Machine Learning

Help

You submitted this quiz on **Sat 4 Jan 2014 9:46 PM PST**. You got a score of **5.00** out of **5.00**.

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
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.
Use fewer examples from your training set.			
Try using a larger learning rate α.			
This is not possible with stochastic gradient descent, as it is guaranteed to converge to the optimal parameters θ.			
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
☑ 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 $\frac{\partial}{\partial \theta_j} cost(\theta, (x^{(i)}, y^{(i)}))$.)	✔	0.25	Just as with batch gradient descent, you can compute the derivative numerically and compare it to your computed value to check for correctness.
☐ 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.
☐ Stochastic gradient descent is particularly well suited to problems with small training set sizes; in these problems, stochastic gradient descent is often preferred to batch gradient descent.	✔	0.25	Stochastic gradient descent is preferred when you have a large training set size; if the data set is small, then the summation over examples in batch gradient descent is not an issue.
☑ One of the advantages of stochastic gradient descent is that it can start progress in improving the parameters θ after looking at just a single training example; in contrast, batch gradient descent needs to take a pass over the entire training set before it starts to make progress in improving the parameters' values.	✔	0.25	This is true, since stochastic gradient descent updates the parameters for every training example, but batch gradient descent updates them based on an average over the entire training set.

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
☑ One of the advantages of online learning is that if the function we're modeling changes over time (such as if we are modeling the probability of users clicking on different URLs, and user tastes/preferences are changing over time), the online learning algorithm will automatically adapt to these changes.	✔	0.25	Online learning algorithms move toward correctly classifying the most recent examples, so as user tastes change and we receive new, different data, the algorithm will automatically take those into account.
☐ One of the advantages of online learning is that there is no need to pick a learning rate α.	✔	0.25	One still must choose a learning rate to use online learning.
☑ 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	✔	0.25	This is one good approach to online learning discussed in the lecture video.

using that example, and then move on to the next example.			
☐ 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 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.
☐ A neural network 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	✔	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.

(say in order to perform mean normalization).			
☐ 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
☑ 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.	✔	0.25	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.
☐ If we run map-reduce using N computers, then we will always get at least 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.
☑ If you have only 1 computer with 1 computing core, then map-reduce is unlikely to help.	✔	0.25	Map-reduce is a useful model for parallel computation.
☑ If you are have just 1 computer, but your computer has multiple	✔	0.25	Treating each core as a separate computer makes map-reduce just as useful with multiple cores as with multiple computers.

CPUs or multiple cores, then map-reduce might be a viable way to parallelize your learning algorithm.	
Total	1.00 / 1.00