Feedback — XVII. Large Scale Machine Learning

Help

You submitted this quiz on **Sat 4 Jan 2014 9:01 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
Try using a smaller learning rate α.	✔	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.
Try averaging the cost over a larger number of examples (say 1000 examples instead of 500) in the plot.			
Use fewer examples from your training set.			
This is not possible with stochastic gradient descent, as it is guaranteed to converge to the optimal parameters θ.			
Total		1.00 / 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
☐ 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.
☑ 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.
☑ 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, $cost(\theta, (x^{(i)}, y^{(i)}))$.
☐ 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.	✔	0.25	We want to plot $cost(\theta, (x^{(i)}, y^{(i)}))$ at each iteration, as computing the full summation $J_{\text{train}}(\theta)$ is too expensive.
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.
☑ 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 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.
☐ When using online learning, you must save every new training example you	✔	0.25	Online learning algorithms throw away old examples, incorporating them only once when they are first seen.

training example you get, as you will need to reuse past examples to re-train the model even after you get new training examples in the future.

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 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.
☐ An online learning setting, where you repeatedly get a single example (x, y), and want to learn from that single example before	✔	0.25	Since you process one example at a time, this algorithm cannot be easily parallelized.

example before
moving on.

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
☑ 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 CPUs or multiple cores, then map-reduce might be a viable way to parallelize your learning algorithm.	✔	0.25	Treating each core as a separate computer makes map-reduce just as useful with multiple cores as with multiple computers.
☑ Linear regression and logistic regression can be parallelized using map-reduce, but not neural network training.	✘	0.00	All three can be parallelized using map-reduce.
☑ 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.
Total		0.75 / 1.00	