ML Week 10 Large Scale Machine Learning

Close Large Scale Machine Learning

5 questions

1. Suppose you are training a logistic regression classifier using stochastic gradient descent. You find that the cost (say, cost(θ,(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?

- CORRECT 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.
- Try averaging the cost over a smaller number of examples (say 250 examples instead of 500) in the plot.
- This is not possible with stochastic gradient descent, as it is guaranteed to converge to the optimal parameters θ.
- Use fewer examples from your training set.

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

- WRONG Suppose you are using stochastic gradient descent to train a linear regression classifier. The cost function is guaranteed to decrease after every iteration of the stochastic gradient descent algorithm.
- SELECTED You can use the method of numerical gradient checking to verify that your stochastic gradient descent implementation is bug-free.
- Before running stochastic gradient descent, you should randomly shuffle (reorder) the training set.
- SELECTED In order to make sure stochastic gradient descent is converging, we typically compute Jtrain(θ) after each iteration (and plot it) in order to make sure that the cost function is generally decreasing.

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

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

- WRONG 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.
- CORRECT 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.
- WRONG 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.

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.

- CORRECT A neural network trained using batch gradient descent.
- WRONG Logistic regression trained using stochastic gradient descent.
- CORRECT Linear regression trained using batch gradient descent.
- WRONG An online learning setting, where you repeatedly get a single example (x,y), and want to learn from that single example before moving on.
- WRONG Computing the average of all the features in your training set

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

- WRONG Running map-reduce over N computers requires that we split the training set into N^2 pieces.
- WRONG In order to parallelize 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.
- CORRECT 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.
- CORRECT 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.

TRUE Because of network latency and other overheads associated with map-reduce, if we run map-reduce using N computers, we might get less than an N-fld speedup compared to using 1 computer

Stochastic Gradient Checking

- In each iteration of the stochastic gradient, the algorithm needs to examine/use only one training example.
- Each iteration updates the parameters based on the cost of only one example $Cost(\theta, (x^{(i)}, y^{(i)}))$.
- You can use numerical gradient checking to verify if your stochastic gradient descent implementation is bug-free.

Machine Learning System Design

Suppose a massive dataset is available for training a learning algorithm. Training on a lot of data is likely to give good performance when two of the following conditions hold true. Which are the two?

(i) We train a learning algorithm with a small number of parameters (that is thus unlikely to overfit).

No - If the model has a small number of parameters, then it will underfit the large training set and not make good use of all the data.

(ii) We train a learning algorithm with a large number of parameters (that is able to learn/represent fairly complex functions).

Yes - You should use a "***low bias***" algorithm with many parameters, as it will be able to make use of the large dataset provided. If the model has too few parameters, it will underfit the large training set.

(iii) The features x contain sufficient information to predict y accurately. (For example, one way to verify this is if a human expert on the domain can confidently predict y when given only x).

Yes - It is important that the features contain sufficient information, as otherwise no amount of data can solve a learning problem in which the features do not contain enough information to make an accurate prediction.

(iv) When we are willing to include high order polynomial features of x (such as x_1^2, x_2^2, $x_1 x_2$, etc.).

No - As we saw with neural networks, polynomial features can still be insufficient to capture the complexity of the data, especially if the features are very high-dimensional. Instead, you should use a complex model with many parameters to fit to the large training set.