Learning with Large Datasets

[If you look back at a recent 5 or 10-year history of machine learning.](https://www.coursera.org/learn/machine-learning/lecture/CipHf/learning-with-large-datasets) [One of the reasons that learning algorithms work so much better now than even say, 5-years ago,](https://www.coursera.org/learn/machine-learning/lecture/CipHf/learning-with-large-datasets) [is just the sheer amount of data that we have now and that we can train our algorithms on.](https://www.coursera.org/learn/machine-learning/lecture/CipHf/learning-with-large-datasets)

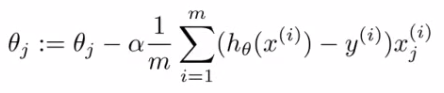
**Why large datasets?**

* One of best ways to get high performance is take a low bias algorithm and train it on a lot of data (if your algorithm is suffering from a bias problem, the training set error would be high and you may find that the cross validation error will also be high. It might be close, maybe just slightly higher than a training error. And so, if you see this combination, that's a sign that your algorithm may be suffering from high bias.)
  + e.g. Classification between confusable words
  + We saw that so long as you feed an algorithm lots of data they all perform pretty similarly

 So it's good to learn with large datasets

 But learning with large datasets comes with its own computational problems

How do we train a **logistic regression** model on such a big system?



 So you have to sum over 100,000,000 terms per step of gradient descent

 Because of the computational cost of this massive summation, we'll look at more efficient ways around this

**Randomly pick a small selection**

* Sometimes yes - if this is the case you can avoid a lot of the headaches associated with big data

To see if taking a smaller sample works, **you can sanity check by plotting error vs. training set size**

Looks like a **high variance problem** (if your algorithm is suffering from high variance; then the training error, is going to be low. That is, you're fitting the training set very well. Whereas, your cross validation error, assuming that this say the squared error which we're trying to minimize. Whereas in contrast; your error on a cross validation set or your cross function like cross validation set, will be much bigger than your training set error.), more examples should improve performance.

More examples may not actually help - save a lot of time and effort if we know this beforehand*.* One natural thing to do here might be to;

Add extra features

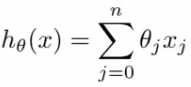
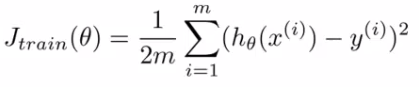
Add extra hidden units (if using neural networks)

**Stochastic Gradient Descent**

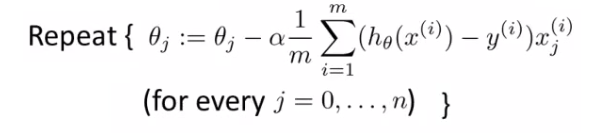
 For many learning algorithms, we derived them by coming up with an optimization objective (cost function) and using an algorithm to minimize that cost function

* When you have a large dataset, gradient descent becomes very expensive
* So here we'll define a different way to optimize for large data sets which will allow us to scale the algorithms

 Suppose you're training a **linear regression** model with gradient descent

* **Hypothesis**   
  
* **Cost function**   
  

**Quick reminder - how does gradient descent work?**



 n the inner loop we repeatedly update the parameters θ

 We will **use linear regression** for our algorithmic example here when talking about **stochastic gradient descent**, although the ideas apply to other algorithms too, such as

* Logistic regression
* Neural networks
* As mentioned, if *m* is large, gradient descent can be very expensive. Although so far we just referred to it as gradient descent, this kind of gradient descent is called **batch gradient descent**. **This just means we look at all the examples at the same time.** Batch gradient descent is not great for huge datasets. If you have 300,000,000 records you need to read in all the records into memory from disk because you can't store them all in memory.

By reading all the records, you can move one step (iteration) through the algorithm. Then repeat for EVERY step

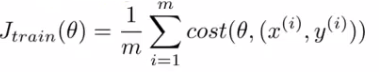
* + This means it takes a **LONG** time to converge.
  + Especially because disk I/O is typically a system bottleneck anyway, and this will inevitably require a *huge* number of reads.
* **Let us Come up with a different algorithm which only needs to look at single example at a time**

**Define our cost function slightly differently**, as

C:\Books\Machine Learning\Machine_learning_complete\Machine_learning_complete\17_Large_Scale_Machine_Learning_files\Image [9].png

*  So the function represents the cost of θ with respect to a specific example (xi,yi)
  + And we calculate this value as one half times the squared error on that example
* Measures how well the hypothesis works on a single example

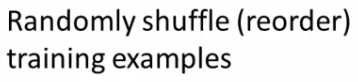
 The overall cost function can now be re-written in the following form;



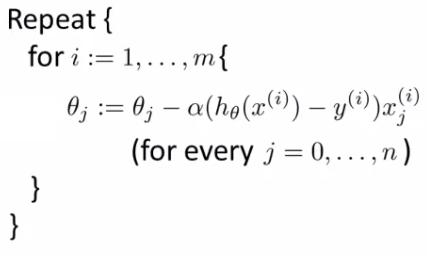
*  This is equivalent to the batch gradient descent cost function

 With this slightly modified (but equivalent) view of linear regression we can write out how stochastic gradient descent works.

* 1. **Randomly shuffle**



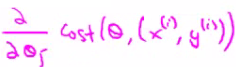
* 1. **- Algorithm body**



 The term   
C:\Books\Machine Learning\Machine_learning_complete\Machine_learning_complete\17_Large_Scale_Machine_Learning_files\Image [13].png

 Is the same as that found in the summation for batch gradient descent

 It's possible to show that this term is equal to the partial derivative with respect to the parameter θj of the cost (θ,(xi,yi))



 **What stochastic gradient descent algorithm is doing is scanning through each example**

* The inner ‘for loop’ does something like this...
  + Looking at example 1, take a step with respect to the cost of just the 1st training example
    - Having done this, we go on to the second training example
  + Now take a second step in parameter space to try and fit the second training example better
    - Now move onto the third training example
  + And so on...
  + Until it gets to the end of the data
* We may now repeat this who procedure and take multiple passes over the data

 The **randomly shuffling** at the start means we ensure the data is in a random order so we don't bias the movement

* **Randomization should speed up convergence a little bit**

 Although stochastic gradient descent is a lot like batch gradient descent, rather than waiting to sum up the gradient terms over all *m* examples, we take just one example and make progress in improving the parameters already

* **Means we update the parameters on EVERY step through data, instead of at the end of each loop through all the data**

 What does the algorithm do to the parameters?

* **As we saw, batch gradient descent does something like this to get to a global minimum**
* **With stochastic gradient descent every iteration is much faster, but every iteration is flitting a single example**
  + What you find is that you "generally" move in the direction of the global minimum, but not always
  + You never actually converge like batch gradient descent does, but end up wandering around some region close to the global minimum
    - In practice, this isn't a problem - as long as you're close to the minimum that's probably OK

 One final implementation note

* May need to loop over the entire dataset 1-10 times
* If you have a truly massive dataset it's possible that by the time you've taken a first pass through the dataset you may already have a perfectly good hypothesis
  + In which case the inner loop might only need to happen once if *m* is very very large

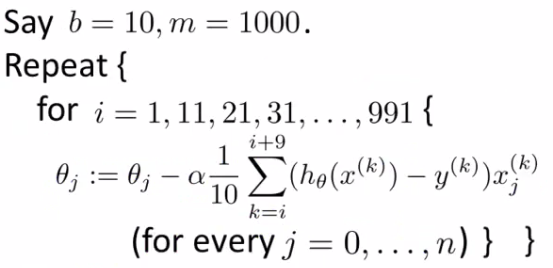
 If we contrast this to batch gradient descent

* We have to make k passes through the entire dataset, where k is the number of steps needed to move through the data.

**Mini Batch Gradient Descent**

* **Mini-batch gradient descent** is an additional approach which can work even faster than stochastic gradient descent
* To summarize our approaches so far
  + **Batch gradient descent: Use all *m* examples in each iteration**
  + **Stochastic gradient descent: Use 1 example in each iteration**
  + **Mini-batch gradient descent: Use *b* examples in each iteration**
    - *b* = mini-batch size
* So just like batch gradient descent, except we use tiny batches
  + Typical range for *b* = 2-100 (10 maybe)
* For example
  + *b* = 10
  + Get 10 examples from training set
  + Perform gradient descent update using the ten examples

**Mini-batch algorithm**



* We for-loop through *b-*size batches of *m*
* Compared to batch gradient descent this allows us to get through data in a much more efficient way.
  + After just *b* examples we begin to improve our parameters.
  + Don't have to update parameters after *every* example, and don't have to wait until you cycled through all the data.

**Mini-batch gradient descent vs. stochastic gradient descent**

* Why should we use mini-batch?
  + Allows you to have a vectorized implementation
  + Means implementation is much more efficient
  + Can partially parallelize your computation (i.e. do 10 at once)
* A **disadvantage** of mini-batch gradient descent is the optimization of the parameter *b*
  + But this is often worth it!
* **To be honest, stochastic gradient descent and batch gradient descent are just specific forms of batch-gradient descent**
  + For mini-batch gradient descent, b is somewhere in between 1 and m and you can try to optimize for it!
* **Stochastic gradient descent convergence**

We now know about stochastic gradient descent

* But how do you know when it's done!?
* How do you tune learning rate alpha (α)?

**Checking for convergence**

* **With batch gradient descent, we could plot cost function vs number of iterations**
  + **Should decrease on every iteration**
  + **This works when the training set size was small because we could sum over all examples**
    - **Doesn't work when you have a massive dataset**
  + **With stochastic gradient descent**
    - **We don't want to have to pause the algorithm periodically to do a summation over all data**
    - **Moreover, the whole point of stochastic gradient descent is to *avoid* those whole-data summations**
* **For stochastic gradient descent, we have to** **do something different**
  + Take cost function definition

C:\Books\Machine Learning\Machine_learning_complete\Machine_learning_complete\17_Large_Scale_Machine_Learning_files\Image [18].png

* + One half the squared error on a single example
* While the algorithm is looking at the example (xi,yi) but *before* it has updated θ we can compute the cost of the example (cost(θ,(xi,yi)
  + i.e. we compute how well the hypothesis is working on the training example
    - Need to do this before we update θ because if we did it after θ was updated the algorithm would be performing a bit better (because we'd have just used (xi,yi) to improve θ)
* To check for the convergence, every 1000 iterations we can plot the costs averaged over the last 1000 examples
  + Gives a running estimate of how well we've done on the last 1000 estimates
  + By looking at the plots we should be able to check convergence is happening

 What do these plots look like

* In general
  + Might be a bit noisy (1000 examples isn't that much)
  + If you use a smaller learning rate you may get an even better final solution
* This is because the parameter oscillates around the global minimum
* A smaller learning rate means smaller oscillations

If you average over 1000 examples and 5000 examples you may get a smoother curve

* This disadvantage of a larger average means you get less frequent feedback.
* That noise is ironed out by taking a greater number of entries per average, it may not decrease, even with a large number. If you see a curve the looks like its increasing, then the algorithm may be displaying divergence. **Should use a smaller learning rate.**

**Learning rate**

 We saw that with stochastic gradient descent we get this wandering around the minimum

* **In most implementations the learning rate is held constant**

 However, if you want to converge to a minimum you can slowly decrease the learning rate over time

* A classic way of doing this is to calculate α as follows  
  α = const1/(iterationNumber + const2)
* Which means you're guaranteed to converge somewhere
  + You also need to determine const1 and const2

**Online learning**

New setting

* Allows us to model problems where you have a continuous stream of data you want an algorithm to learn from
* Similar idea of stochastic gradient descent, in that you do slow updates
* Web companies use various types of online learning algorithms to learn from traffic
  + Can (for example) learn about user preferences and hence optimize your website

 Build this model with something like

* + Logistic regression
  + Neural network

 If you have a website that runs continuously an online learning algorithm would do something like this

* User comes - is represented as an (x,y) pair where
  + x - feature vector including price we offer, origin, destination
  + y - if they chose to use our service or not
* The algorithm updates θ using just the (x,y) pair

C:\Books\Machine Learning\Machine_learning_complete\Machine_learning_complete\17_Large_Scale_Machine_Learning_files\Image [25].png

So we basically update all the θ parameters every time we get some new data

While in previous examples we might have described the data example as (xi,yi), for an online learning problem we discard this idea of a data "set" - instead we have a continuous stream of data so indexing is largely irrelevant as you're not storing the data (although presumably you could store it) /font.

 If you have a major website where you have a massive stream of data then this kind of algorithm is pretty reasonable

* You're free of the need to deal with all your training data

 If you had a small number of users you could save their data and then run a normal algorithm on a dataset

 An online algorithm can adapt to changing user preferences

* So over time users may become more price sensitive
* The algorithm adapts and learns to this
* So your system is dynamic

**Another example - product search**

How do we do this

* For each phone and given a specific user query, we create a feature vector (x) which has data like features of the phone, how many words in the user query match the name of the phone,
  + Basically how well does the phone match the user query
* We want to estimate the probability of a user selecting a phone
* So define
  + y = 1 if a user clicks on a link
  + y = 0 otherwise
* So we want to learn
  + p(y = 1|x ; θ) <- this is the problem of learning the predicted **click through rate** (CTR)
* If you can estimate the CTR for any phone we can use this to show the highest probability phones first
* If we display 10 phones per search, it means for each search we generate 10 training examples of data
  + i.e. user can click through one or more, or none of them, which defines how well the prediction performed

 Other things you can do

* Special offers to show the user
* Show news articles - learn what users like
* Product recommendation

 These problems could have been formulated using standard techniques, but they are the kinds of problems where you have so much data that this is a better way to do things

**Map reduce and data parallelism**

 Previously spoke about stochastic gradient descent and other algorithms

* These could be run on one machine
* Some problems are just too big for one computer
* Talk here about a different approach called Map Reduce

 Map reduce example

* We want to do batch gradient descent

C:\Books\Machine Learning\Machine_learning_complete\Machine_learning_complete\17_Large_Scale_Machine_Learning_files\Image [26].png

Split training sets into different subsets

* So split training set into 4 pieces

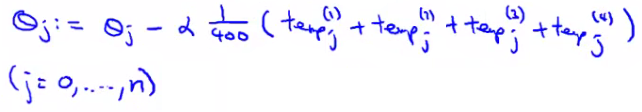
**Machine 1:** use (x1, y1), ..., (x100, y100)

* Uses first quarter of training set
* Just does the summation for the first 100

C:\Books\Machine Learning\Machine_learning_complete\Machine_learning_complete\17_Large_Scale_Machine_Learning_files\Image [27].png

Once we've got our temp variables

* Send to a centralized master server
* Put them back together
* Update θ using



 The bulk of the work in gradient descent is the summation

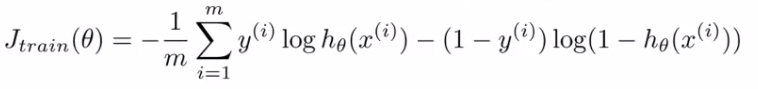
* Now, because each of the computers does a quarter of the work at the same time, you get a 4x speedup
* Of course, in practice, because of network latency, combining results, it's slightly less than 4x, but still good!

 Important thing to ask is

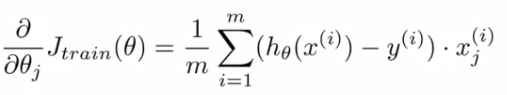
* "Can algorithm be expressed as computing sums of functions of the training set?"
  + Many algorithms can!

 Another example

* **Using an advanced optimization algorithm with logistic regression**



* Need to calculate cost function - see we sum over training set
* So split training set into x machines, have x machines compute the sum of the value over 1/xth of the data



* + - These terms are also a sum over the training set
    - So use same approach
* So with these results send temps to central server to deal with combining everything
* More broadly, by taking algorithms which compute sums you can scale them to very large datasets through parallelization
  + Parallelization can come from
    - Multiple machines
    - Multiple CPUs
    - Multiple cores in each CPU
  + So even on a single compute can implement parallelization
* The advantage of thinking about Map Reduce here is because you don't need to worry about network issues
  + It's all internal to the same machine
* Finally caveat/thought
  + Depending on implementation detail, certain numerical linear algebra libraries can automatically parallelize your calculations across multiple cores
  + So, if this is the case and you have a good vectorization implementation you can not worry about local parallelization and the local libraries sort optimization out for you