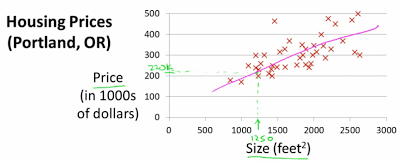
**[Machine-learning: gradient descent](http://digitheadslabnotebook.blogspot.com/2011/10/machine-learning-gradient-descent.html)**

The first section of [Andrew Ng's Machine Learning class](http://ml-class.org) is about applying gradient descent to linear regression problems.

[](http://4.bp.blogspot.com/-S11k6vJwxdc/TpDINkQ_bDI/AAAAAAAADGY/Q8t09bzT4y0/s1600/house_prices_portland.png)

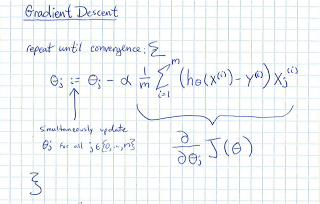
*Andrew Ng*

Our input data is an *m-by-n* matrix *X*, where we have *m* training examples with *n* features each. For these training examples, we know the expected outputs *y* where *y* is the variable we're trying to predict. We want to find a line defined by the parameter vector ϴ that minimizes the squared error between the line and our data points.

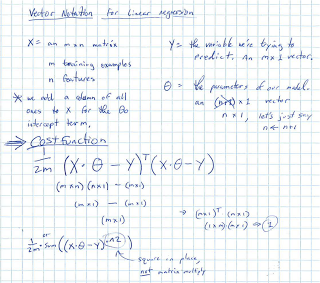
Gradient descent takes a cost function, which is the squared error of the prediction vs. the training data. I think the 2 in the denominator is there so that it cancels out when we take the derivative, leaving us with a simpler gradient function.

[http://3.bp.blogspot.com/-7vLgOxMhaBg/Tqjb34B9A9I/AAAAAAAADIg/w5E6cXN6x1c/s320/cost_function.png](http://3.bp.blogspot.com/-7vLgOxMhaBg/Tqjb34B9A9I/AAAAAAAADIg/w5E6cXN6x1c/s1600/cost_function.png)

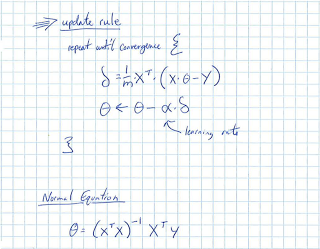
The update rule for each ϴj is the partial derivative of the cost function with respect to ϴj.

[](http://1.bp.blogspot.com/-ysYz2pUSNXc/TqdJwZqGs9I/AAAAAAAADH8/0HROxzQ49HI/s1600/ml-class-notes-gradient-descent-1.png)

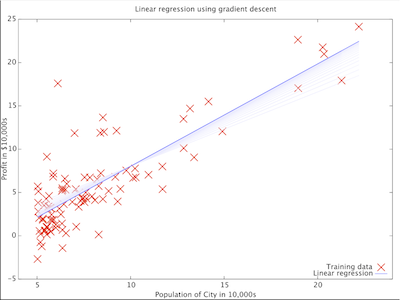
Part of the challenge is converting this to matrix notation, to take advantage of fast matrix arithmetic algorithms.

[](http://4.bp.blogspot.com/-XkseSMwpOaU/TqdJwu2QxII/AAAAAAAADIM/SWaFvyVtMdk/s1600/ml-class-notes-gradient-descent-2.png)

Next we vectorize the update rule and show how to compute least squared error directly with the normal equation.

[](http://3.bp.blogspot.com/-trhkyiBiaTs/TqdJxHBlVvI/AAAAAAAADIU/qFztMw_Ysys/s1600/ml-class-notes-gradient-descent-3.png)

In action, gradient descent gradually approaches optimal values for ϴ. How gradual depends on the learning rate, α.

[](http://1.bp.blogspot.com/-Jq0vhkuzgfU/TqjfI_mXhLI/AAAAAAAADIs/juXvuk4qvYs/s1600/gradient_descent_linear_regression_screenshot.png)

While the classwork was done in [Octave](http://digitheadslabnotebook.blogspot.com/2011/10/octave-cheat-sheet.html), I also did a simple [gradient descent implementation in R](http://digitheadslabnotebook.blogspot.com/2012/07/linear-regression-by-gradient-descent.html).