Week 1

# Supervised Learning

## Regression Analysis

回归分析Regression Analysis是一种预测性的建模技术，它研究的是因变量（目标）和自变量（预测器）之间的关系。这种技术通常用于预测分析，时间序列模型以及发现变量之间的[因果关系](http://www.analyticsvidhya.com/blog/2015/06/establish-causality-events/)。

回归分析（regression analysis)是确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。

## Supervised Learning

Right answer is given.

## Regression problem

The goal is to output continuous value output

## Classification problem

Discrete value output

## Support Vector Machine

To have a neat mathematical trick that will allow a computer to deal with an infinite number of features.

# Unsupervised Learning

Approach problems with little or no idea what our results should look like.

## Clustering algorithm

To group individuals into different categories.

## Cocktail party problem

## Development tool – Octave

# Linear Regression

## Notation:

m - the number of training examples

x - the input variables or the input features

y -  the output variables or the target variable

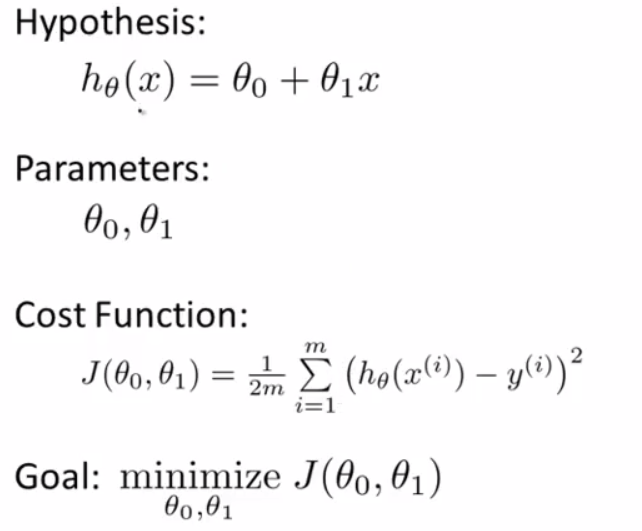
(x, y) - a single training example

(x(i), y(i)) - ith training example

h – hypothesis

## Cost Function

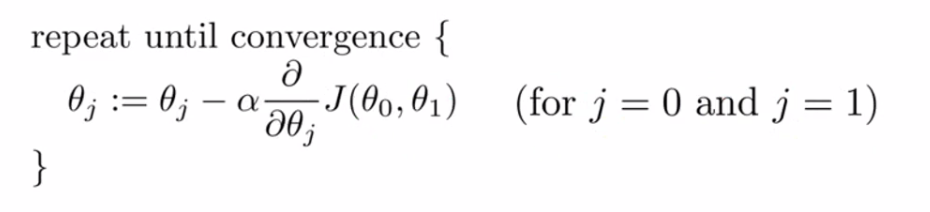
Minimize the cost function (also called squared error function)



### Contour plots (figures)

Contour plots/figures to show the cost function result

# Gradient descent



Property: could end up to a local minimum, because of different starting points. And will take smaller steps to converge to local minimum with fixed-value learning rate.

## Notation:

a:=b – assign b’s value to a, := assignment

a = b – assert a and b are the same, = truth assertion

α - learning rate, how big step to take when updating the parameter theory j.

Derivative term

## Apply gradient descent to linear regression model

## Batch Gradient Descent

Each step of gradient descent uses all the training examples.

# Linear Algebra

## Matrix and Vectors

Vector: an n x 1 matrix. We are using 1-indexed vector.

Lower-case refers to vector; upper-case refers to matrix.

"Scalar" means that an object is a single value, not a vector or matrix.

ℝ refers to the set of scalar real numbers.

ℝn refers to the set of n-dimensional vectors of real numbers.

## Matrix Multiplication Properties

Matrices are not commutative:

Matrices are associative:

The identity matrix, denoted I, as below

## Inverse and Transpose

### Matrix Inverse

If A is an m x m matrix, and if it has an inverse,

Compute inverse matrix in Octave.

### Matrix Transpose

Week 2

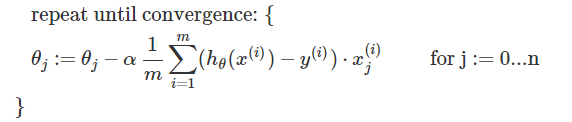
# Multivariate Linear Regression

Linear Regulation with multiple variables (features):

-- the input (features) of the training example

-- value of feature in the  training example

## Gradient Descent for Multiple Variables



m – the number of training examples

n – the number of variables (features)

## Features scaling

Trick to make sure features are on a similar scale, so that gradient descent can run faster.

#### Features scaling

Get ever feature into approximately a range.

#### Mean normalization

Replace with to make features have approximately zero mean (do not apply to )

– average value of in training set

– range value () of in training set

For example, if xi represents housing prices with a range of 100 to 2000 and a mean value of 1000, then,

## Learning Rate

Making sure gradient descent is working correctly. Plot the cost function, over the number of iterations of gradient descent.

Use smaller but efficient value of α.

## Features and Polynomial Regression

Our hypothesis function need not be linear (a straight line) if that does not fit the data well. We can change choice of features or the behavior or curve of our hypothesis function by making it **quadratic**, **cubic** or square root function (or any other form).

Applying feature scaling is very important in this way.

# Computing Parameters Analytically

## Normal Equation

(1)另一种线性回归方法：Normal Equation，正规化方程

(2)Gradient Descent与Normal Equation的优缺点；

前面我们通过Gradient Descent的方法进行了线性回归，但是梯度下降有如下特点：

(1)需要预先选定Learning rate；

(2)需要多次iteration；

(3)需要Feature Scaling；

因此可能会比较麻烦，这里介绍一种适用于Feature数量较少时使用的方法：Normal Equation；

当Feature数量小于100000时使用Normal Equation；

当Feature数量大于100000时使用Gradient Descent；

When n is small number (maybe <1000), calculate directly which makes minimize:

No need for feature scaling.

## Normal Equation and non-invertibility不可逆性

The **'pinv'** function in Octave will give you a value of θ even if  is **not invertible**.

If   is noninvertible, the common causes might be having :

* Redundant features, where two features are very closely related (i.e. they are linearly dependent)

e.g: m=3,62 feet, – m, – feet. Solution: delete or

* Too many features (e.g. m ≤ n). In this case, delete some features or use "regularization" (to be explained in a later lesson).

Solutions to the above problems include deleting a feature that is linearly dependent with another or deleting one or more features when there are too many features.

# Octave/Matlab Tutorial

Check tutorial on net: <https://en.wikibooks.org/wiki/Octave_Programming_Tutorial>

# Programming Assignment: Linear Regression

(./) operator to perform element by element division; (.\*) operator to perform element by element multiply, etc..

‘Markersize’: (e.g) '+' marker on the graph for each datapoint, ‘Markersize’ is the size of ‘+’ markers.

‘rx’: set the markers to red ‘x’

Function bsxfun(): the function applies an element-by-element binary operation to arrays a and b, with singleton expansion enabled.”

x’: Complex conjugate transpose. 在assignment里，就是转置的意思，因为没有conjugate transpose中的实数虚数。

Function logspace (a, pi, n): Return a row vector with n elements logarithmically spaced from 10^a to 10^b.

Function mean(A,2): For matrices, mean(A,2) is a column vector containing the mean value平均数 of each row.

Funciton std (x, opt, dim): compute the standard deviation标准差of the elements of the vector x.

标准差可以描述样本中的数据分布:

1. 找出平均数。
2. 找出方差。方差是数据偏离平均数的程度. 得到方差首先要计算单个样本数据和平均数的差，然后平方，再求平均数。
3. 方差开方即得到标准差。标准差会告诉你数据域平均数的离散程度，约68%的样本数据在一个标准差范围内，

Use standard deviation instead of taking the range of value (max-min): The standard deviation is a way of measuring how much variation there is in the range of values of a particular feature (most data points will lie within \_2 standard deviations of the mean); this is an alternative to taking the range of values (max-min).

Week 3

# Classification and Representation

## Classification

Classification:

|  |  |
| --- | --- |
|  | 0: Negative Class |
| 1: Positive Class |

Threshold: (e.g.) 0.5

Multiclass classification:

Linear regression is not suitable for classification problem:

* irregular training data may dramatically affect
* for some data, could be >1 or <0

Solution: Logistic Regression

## Hypothesis Representation

### Logistic/Sigmoid Function

To satisfy, use Logistic/Sigmoid Function:

Sigmoid function image:



Probability that y=1, given x parameterized by :

## Decision boundary

, when, that’s when

Based on given parameter, decision boundary is. The given parameter is calculated based on training data set.

### Non-linear decision boundary

Higher order polynomial terms, for example:

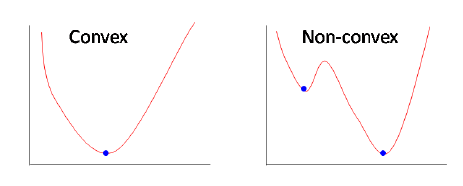
By given parameter, the non-linear decision boundary could be non-linear, for example

# Logistic Regression Model

## Cost Function

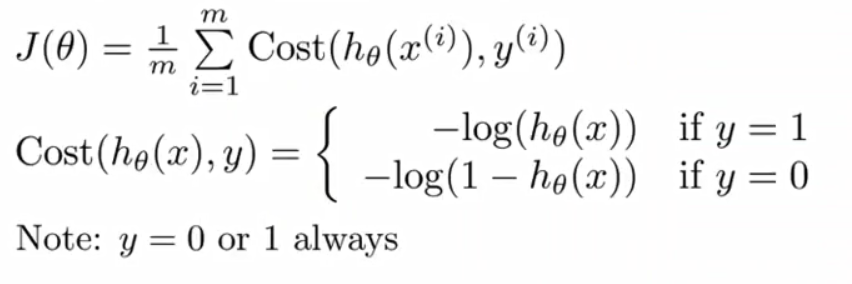
### The chosen of cost function:

Cost function in linear regression is not fit for logistic regression; it will cause the output to be wavy, causing many local optima. In other words, it will not be a convex function for gradient descent.



### Logistic Regression Cost Function

The cost function in this way guarantees that is convex for logistic regression.

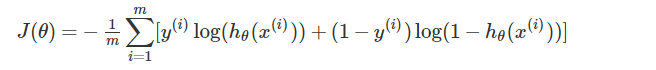


|  |  |
| --- | --- |
|  |  |

## Simplified cost function and Gradient Descent

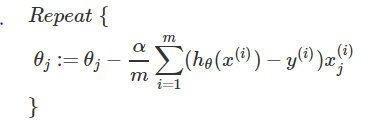
### Simplified cost function

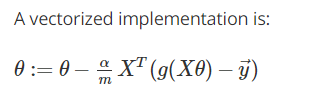




### Gradient Descent

This algorithm is **identical** to the one we used in linear regression, but





**Feature scaling** can be also applied to Logistic Regression

### Supplement: Maximum Likelihood Estimation最大似然估计

最大似然估计是利用已知的样本的结果，在使用某个模型的基础上，反推最有可能导致这样结果的模型参数值。

求解最大似然估计的一般过程为:

1. 写出似然函数
2. 如果无法直接求导的话，对似然函数取对数
3. 求导数
4. 求解模型中参数的最优值。

## Advanced Optimization

* Gradient Descent
* Conjugate gradient
* BFGS
* L-BFGS

"Conjugate gradient", "BFGS", and "L-BFGS" are more sophisticated, complex, faster ways to optimize θ that can be used instead of gradient descent.

Use library in Octave to call these optimization.

