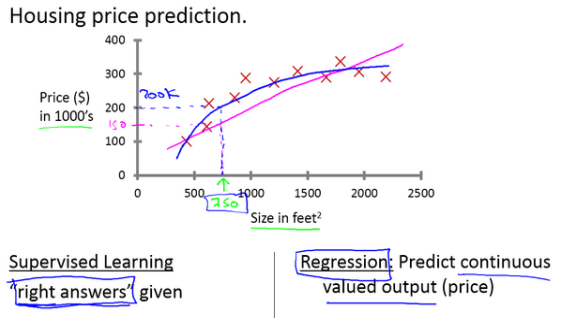
# 吴恩达机器学习--纯理论研究

1-3 Supervised Learning

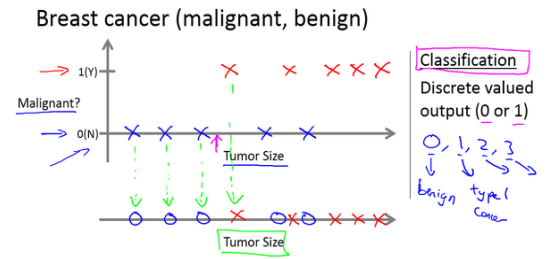
监督学习：

实例：预测房价。买房的人都知道买房的时候有很多考虑因素，面积，单价，地段，户型等。 这里仅采用面积来衡量。机器学习怎么帮助你预测房价呢？



一组数据在xy轴上打点，我们在这组数据中心画一条直线，或者说画一条近似直的线。从而可以预测房价接近150000。当然，直线并不是唯一的，可以用曲线(二次方程)，曲线预测能卖接近200000.肯定是越接近真实卖价越好。这就是典型的监督学习的例子。

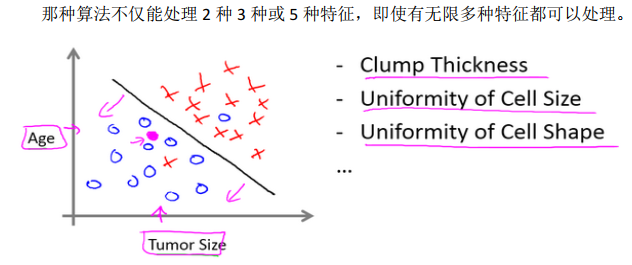
实例2： 通过病历来推测乳腺癌良性与否。



数据：横轴标识肿瘤的大小，纵轴1和0表示是或者不是恶性肿瘤，如果恶性则为1，良性则为0.

通过肿瘤样本能估算出肿瘤是恶性的还是良性的，这就是一个监督学习中的分类问题。

说明：上述两个例子中特征都有很多，



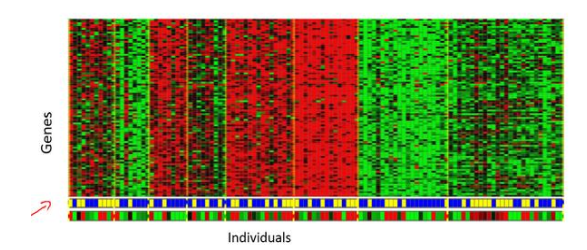
比如上图就有5个特征，但是有的场景中特征非常多，存储或者处理都成了问题。 以后学习的一个算法叫做支持向量机，里面有一个巧妙的数学技巧，能让计算机处理无限多个特征。

1.4 Unsupervised Learning

无监督学习

无监督学习没有任何的标签，已经知道了数据集但是不知道怎么处理。 别的什么都不知道，就是一个数据集。能从这些数据中找出某种结构吗？ 针对数据集，无监督学习就能判断出数据有两个不同的聚集簇。 无监督学习会把这些数据分成两个不同的簇，所以叫做聚类算法。

实例：谷歌新闻，news.google.com , 每天收集非常多的新闻内容。 将这些新闻分组，组成有关联的新闻。

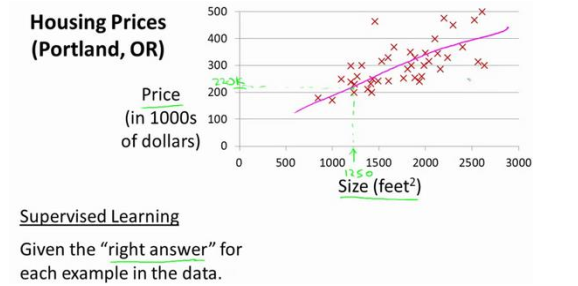


其中就有基因学的理解应用。一个DNA微观数据的例子。就是采用聚类算法，把个体聚类到不同的类或者不同类型的组。

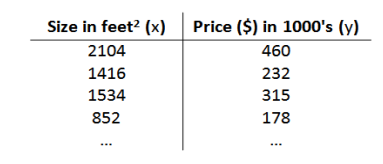
2.1 单变量线性回归(Linear Regression with one Variable)

第一个算法： 线性回归算法

实例： 预测房价



原始带有标记的数据集称为训练集



m 代表训练集中实例的数量

X 代表特征、输入变量

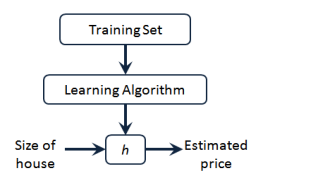
Y 代表目标变量/输出变量

(x,y)代表训练集中的实例

代表第i个观察实例

h代表学习算法的解决方案或函数也称为假设

流程如下:



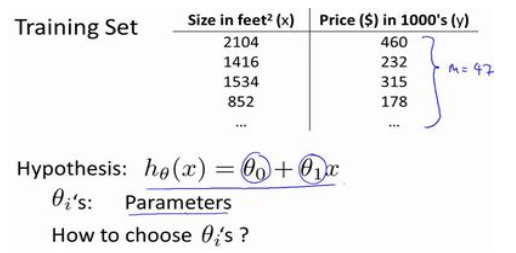
但是对于h，如何表达呢？

m代表训练样本的数量，如m=47

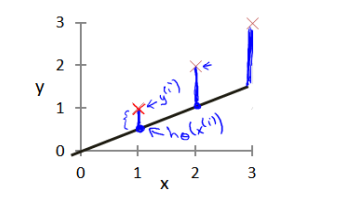
一种表示方式是，因为只含有一个特征/输入变量，因此这样的问题就叫做单变量线性回归问题。

2.2 代价函数 cost-function

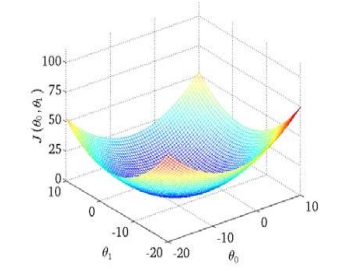
到底什么是代价函数？



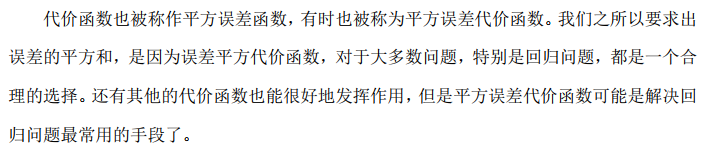
我们认为参数就是直线的斜率和在y轴上的截距。我们选择的参数决定了我们得到的直线对于我们训练集的准确程度，模型所预测值和训练集中实际值的差距就是建模误差



目标便是选择出可以使得建模误差的平方和能够最小的模型参数。即使得代价函数。



可以看到在三维空间中存在一个使得最小的点。



所以说，回归问题首选的误差函数是平方误差函数



后续会讲解其他代价函数，但是平方误差函数对于大多数的回归是非常合理的。