分享内容

# Python基础

## 1.1 Python语言的特点

Python的特点：

* 简单易学

有时候感觉读python代码跟读英语段落一样，比如：for <变量> in <循环序列>。

* 开源

Python 是 FLOSS（自由/开源源码软件）之一，简单地理解就是，用户使用 Python 进行开发和发布自己编写的程序，不需要支付任何费用，也不用担心版权问题，即使作为商业用途。

* 高级语言

Python 是高级语言，因此当使用 Python 语言编写程序时，我们无需再考虑一些底层细节方面的问题。

* 可移植性

所有Python程序都无需修改就可以在好多平台上运行，包括 Linux 、Windows、FreeBSD、Solaris 等。

* 强大的功能和可扩展性

从字符串处理到复杂的 3D 图形编程，Python 借助扩展模块都可以轻松完成，Python 具有脚本语言中最丰富和强大的类库，这些类库覆盖了文件 I/O、GUI、网络编程、数据库访问、文本操作等绝大部分应用场景。

Python 可扩展性一个最好的体现是，当我们需要一段关键代码运行的更快时，可以将其用 C 或 C++ 语言编写，然后在 Python 程序中使用它们即可。

Python可以调用C/C++的库。

Python中有句名言：**不要重复造轮子，知名的库就有400多个。**许多有趣的功能都可以通过调用**已有的python库**配合简短的代码就可以实现。下面举几个例子：

* 抓取网站上的图片（Selenium爬虫库、request库）。
* 分析诗的作者是李白还是杜甫（jieba中文分词库、NLTK自然语言处理的库）。
* 自动写检讨书（输入主题和字数）。

## 1.2 Python能做什么

1. **Web开发**

Django和flask, 可以快速地开发功能强大的Web应用。

1. **网络爬虫**

Python自带的urllib库，第三方的requests库和Scrappy框架让开发爬虫变得非常容易。可以爬取网络上的各种数据和信息。

1. **计算与数据分析**

随着NumPy，pandas，SciPy，Matplotlib等众多程序库的开发和完善，Python越来越适合于做科学计算和数据分析了。它不仅支持各种数学运算，还可以绘制高质量的2D和3D图像。和科学计算领域最流行的商业软件Matlab相比，Python比Matlab所采用的脚本语言的应用范围更广泛，可以处理更多类型的文件和数据。

1. **人工智能**

Python在人工智能大范畴领域内的机器学习、神经网络、深度学习等方面都是主流的编程语言，得到广泛的支持和应用。最流行的神经网络框架如Facebook的PyTorch和Google的TensorFlow都采用了Python语言，另外python还提供了机器学习库scikit-learn。

1. **自动化运维**

Python可以说是运维工程师首选的编程语言。在很多操作系统里，Python是标准的系统组件。大多数Linux发行版和MacOSX都集成了Python，可以在终端下直接运行Python。Python标准库包含了多个调用操作系统功能的库。Python编写的系统管理脚本在可读性、性能、代码重用度、扩展性几方面都优于普通的shell脚本。

1. **云计算**

Python的最强大之处在于模块化和灵活性，而构建云计算平台服务的OpenStack（可以说是虚拟化管理平台）就是采用Python的。

1. **网络编程**

Python提供了丰富的模块支持sockets编程，能方便快速地开发分布式应用程序

1. **游戏开发**

很多游戏使用C++编写图形显示等高性能模块，而使用Python或者Lua编写游戏的逻辑等。Python的PyGame库也可用于直接开发一些简单游戏。

# 计算与数据分析常用的库

主要介绍numpy、pandas、Matplotlib、这三个常用的库。

1）NumPy(Numerical Python) 是 Python 语言的一个扩展程序库，支持大量的维度数组与矩阵运算，此外也针对数组运算提供大量的数学函数库。

NumPy 是一个运行速度非常快的数学库，主要用于数组计算，包含：

* 一个强大的N维数组对象 ndarray
* 广播功能函数
* 整合 C/C++/Fortran 代码的工具
* 线性代数、傅里叶变换、随机数生成等功能

NumPy 通常与 SciPy（Scientific Python）和 Matplotlib（绘图库）一起使用， 这种组合广泛用于替代 MatLab，是一个强大的科学计算环境，有助于我们通过 Python 学习数据科学或者机器学习。

SciPy 是一个开源的 Python 算法库和数学工具包。

SciPy 包含的模块有最优化、线性代数、积分、插值、特殊函数、快速傅里叶变换、信号处理和图像处理、常微分方程求解和其他科学与工程中常用的计算。

2）Pandas是一个数据结构和数据分析工具。

Pandas是一个强大的分析结构化数据的工具集，基础是 Numpy（提供高性能的矩阵运算）。

Pandas 可以从各种文件格式比如 CSV、JSON、SQL、Microsoft Excel 导入数据。

Pandas 可以对各种数据进行运算操作，比如归并、再成形、选择，还有数据清洗和数据加工特征。

Pandas 广泛应用在学术、金融、统计学等各个数据分析领域。

3）Matplotlib 是一个数据可视化库。

## 2.1 numpy 库

Numpy属性、创建array、运算、随机数生成矩阵运算、numpy索引、array合并、array分隔。

## pandas库

pandas基础、pandas处理丢失数据、读写文件、合并

## matplotlib库

实现数据可视化。

# 人工智能常用的库

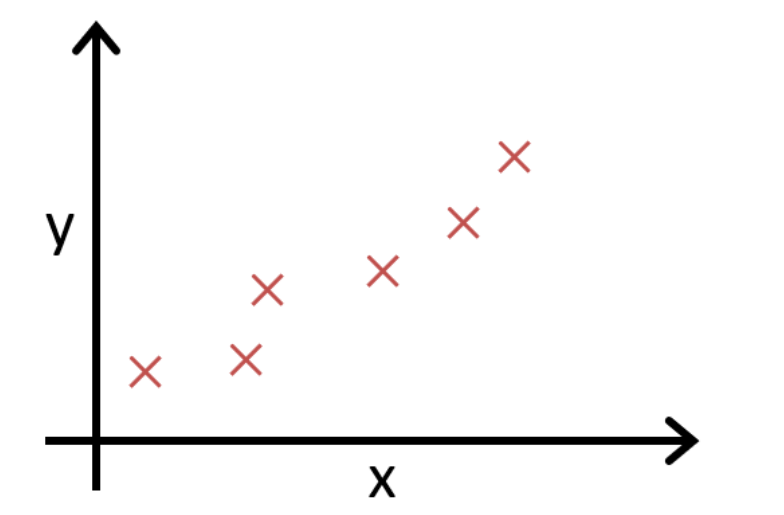
框架（Facebook的PyTorch、 Google的TensorFlow）、机器学习库scikit-learn。本次分享主要讲scikit-learn机器学习库。另外在大数据领域 spark mllib 机器学习库也实现了传统的算法。可以实现分布式上的计算，而tensorflow 不适合用在分布式中。

# 算法

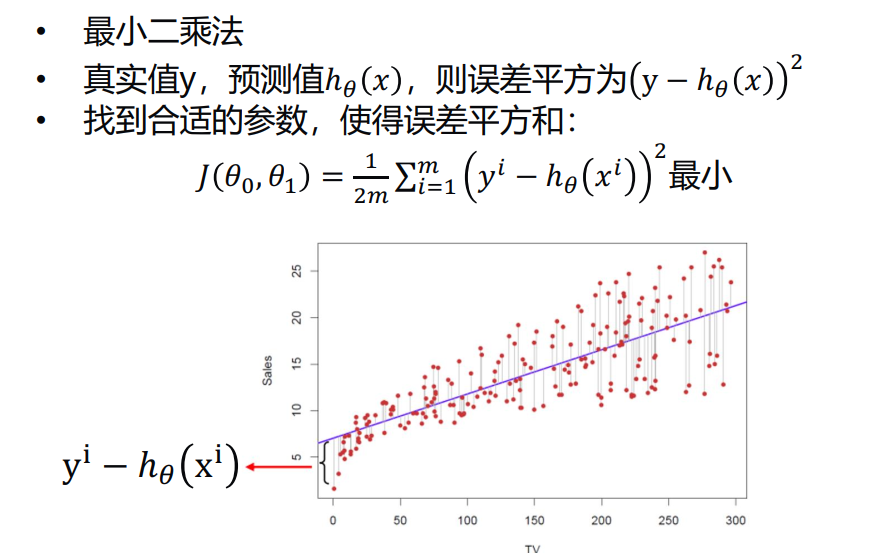
## 4.1 回归分析算法

1）一元线性回归

* 一元线性回归包含一个自变量和一个因变量。
* 以上两个变量的关系用一条直线来模拟。

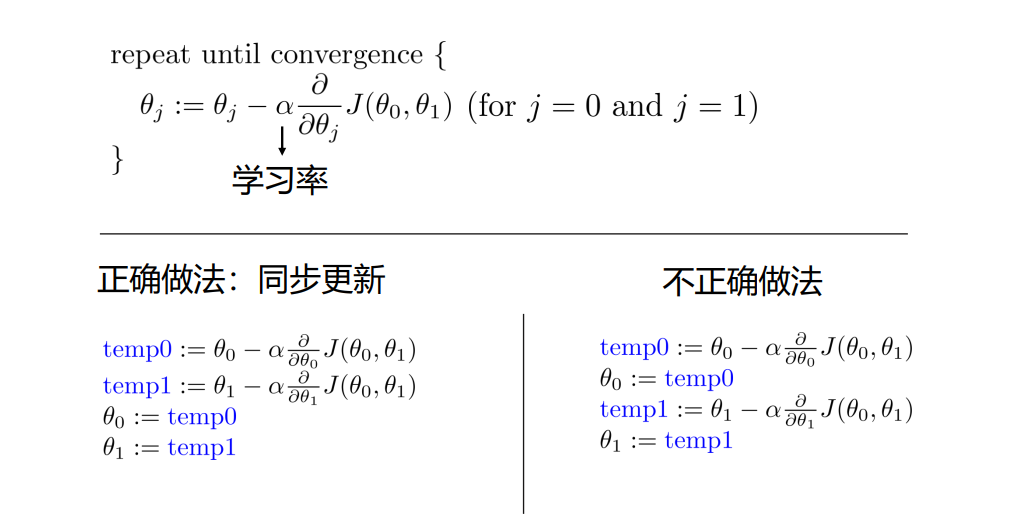


**代价函数：**



**算法步骤：**

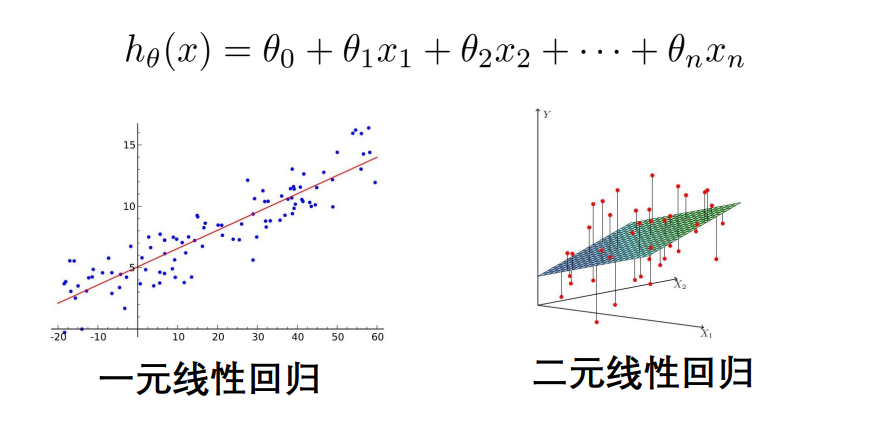
1. 初始化和；
2. 不断改变和，直到误差平方和达到一个全局最小值（即收敛），或局部最小值。



**存在的问题：**可能会出现局部最小值，但是做工程主要以能用为主。

2）多元线性回归

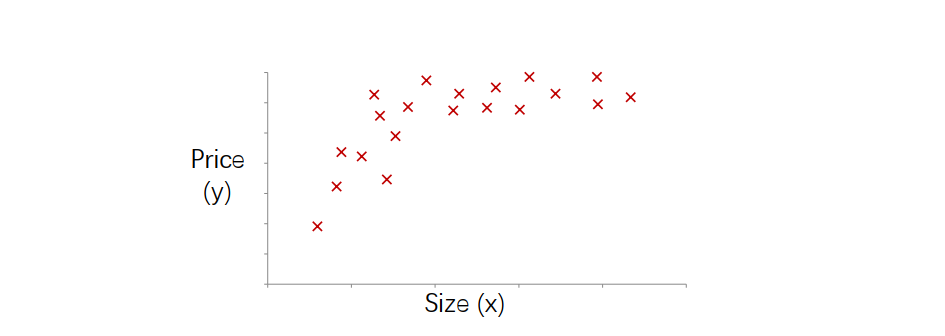
当y值的影响因素不是唯一时，采用多元线性回归模型



3）多项式回归

有些回归不能用线性表示（直线或者超平面），而是一个曲线或者超曲面，就要用到多项式回归。

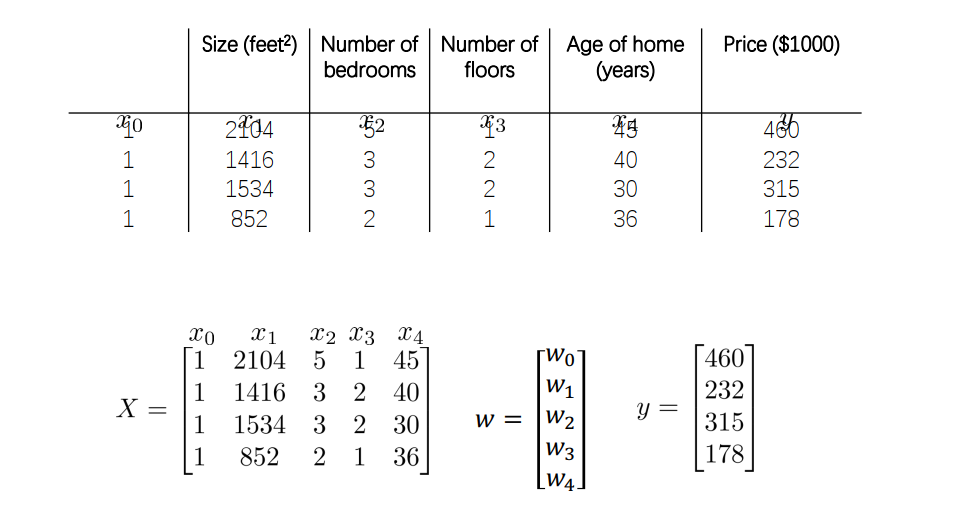
而对于连续函数而言，有一条维尔斯特拉斯（逼近）定理：闭区间上的连续函数可用多项式一致逼近。

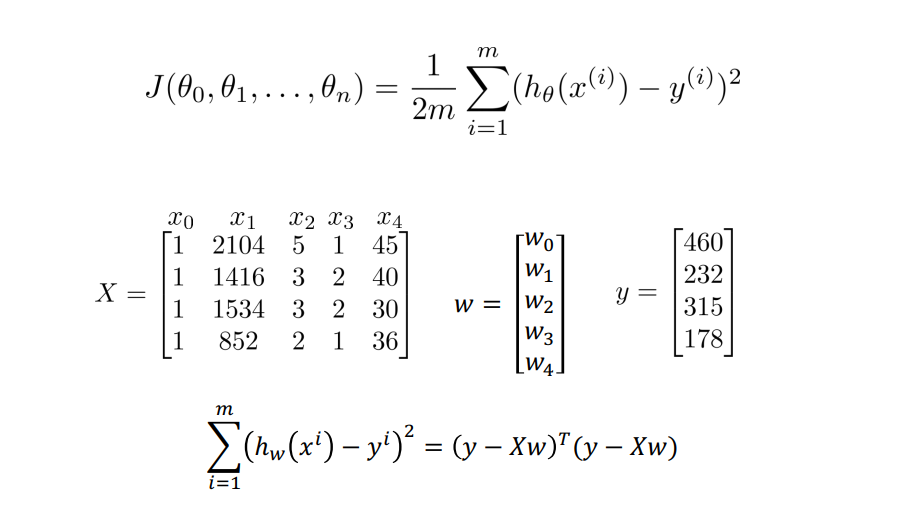


多项式回归可以写成如下形式：

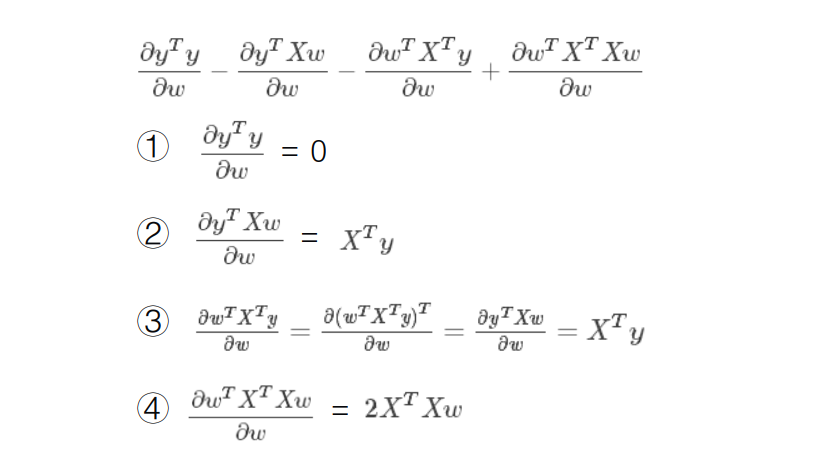


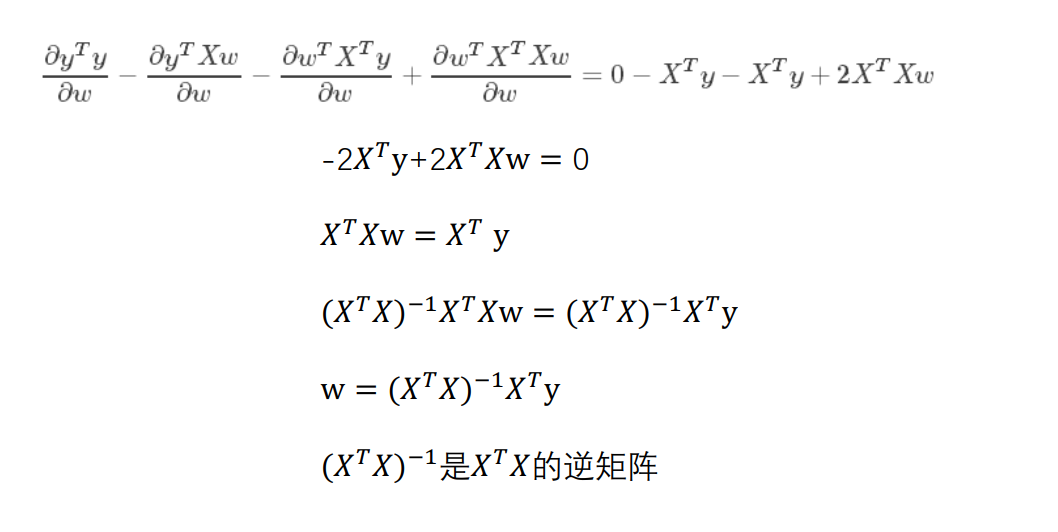
标准方程法求解多项式回归





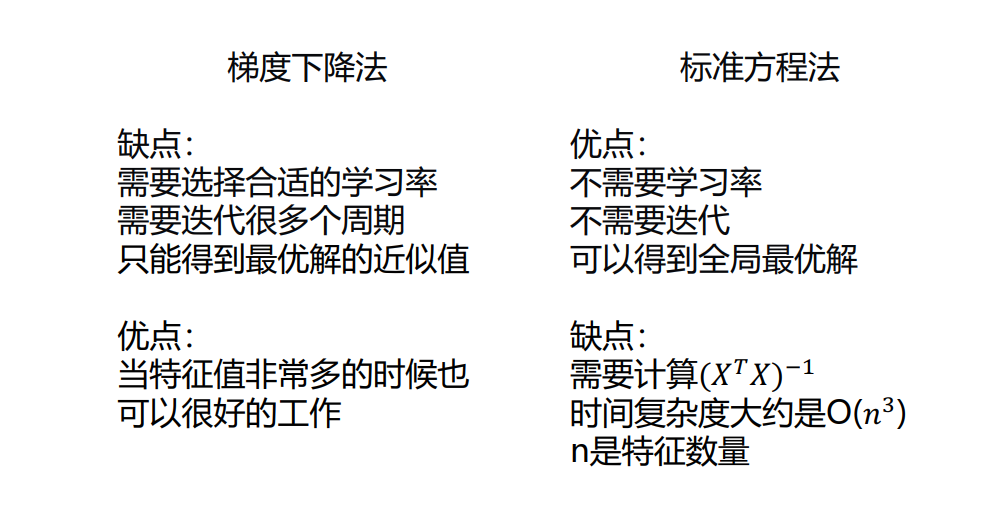
现在就转换成了矩阵的形式，求极值需要对w求偏导。



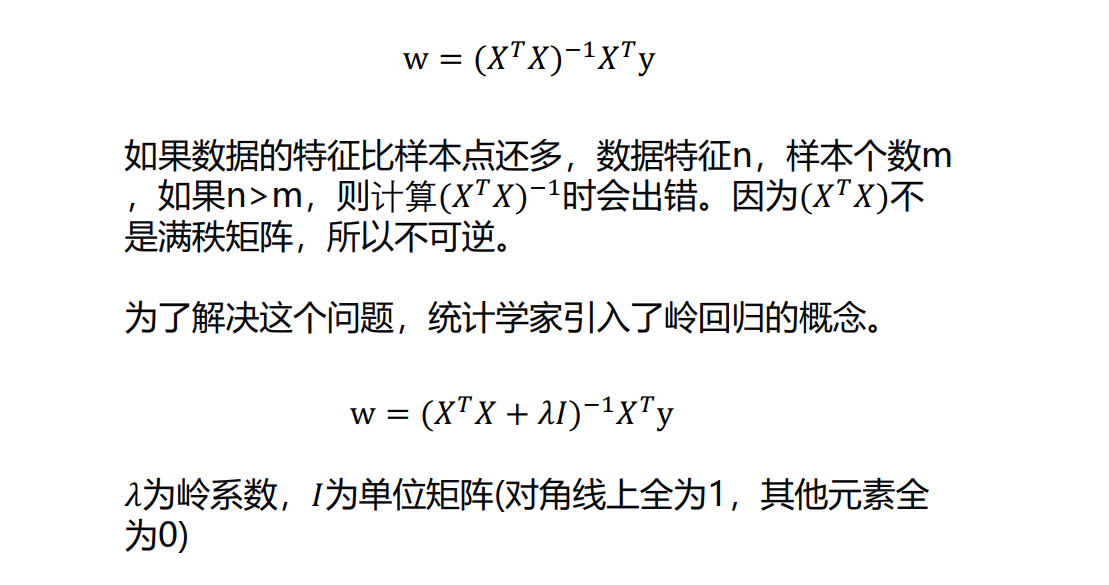


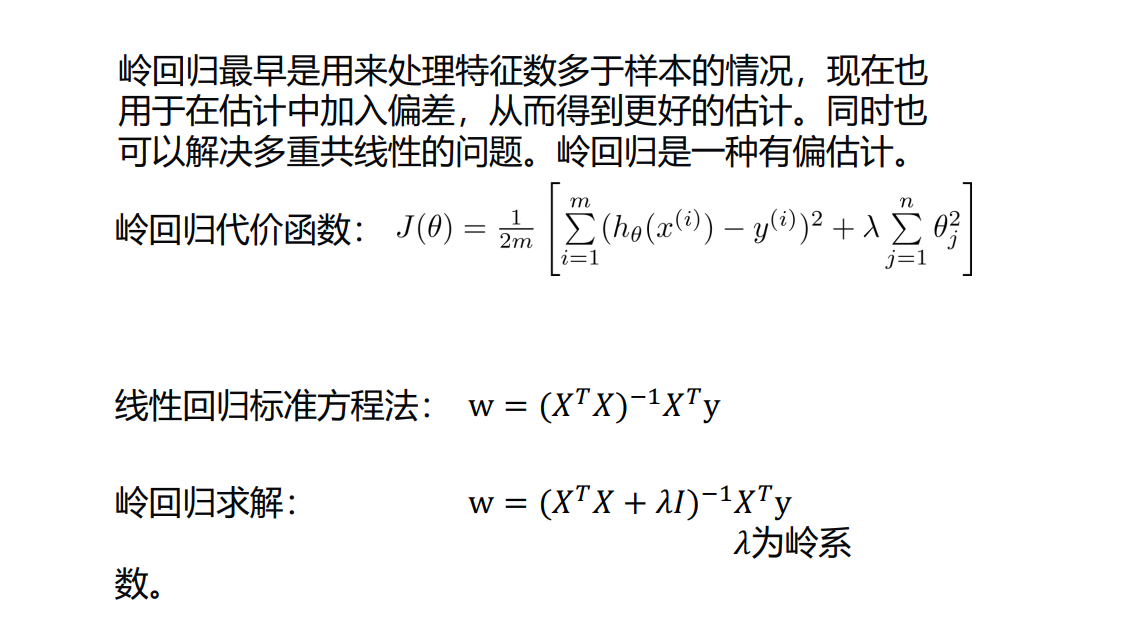
要想求w,这里就涉及到对矩阵求逆，但是并不是所有矩阵都可逆（满足行满秩可逆）。这就造成了标准方程法的使用限制。

两种算法求解对比：



4）岭回归





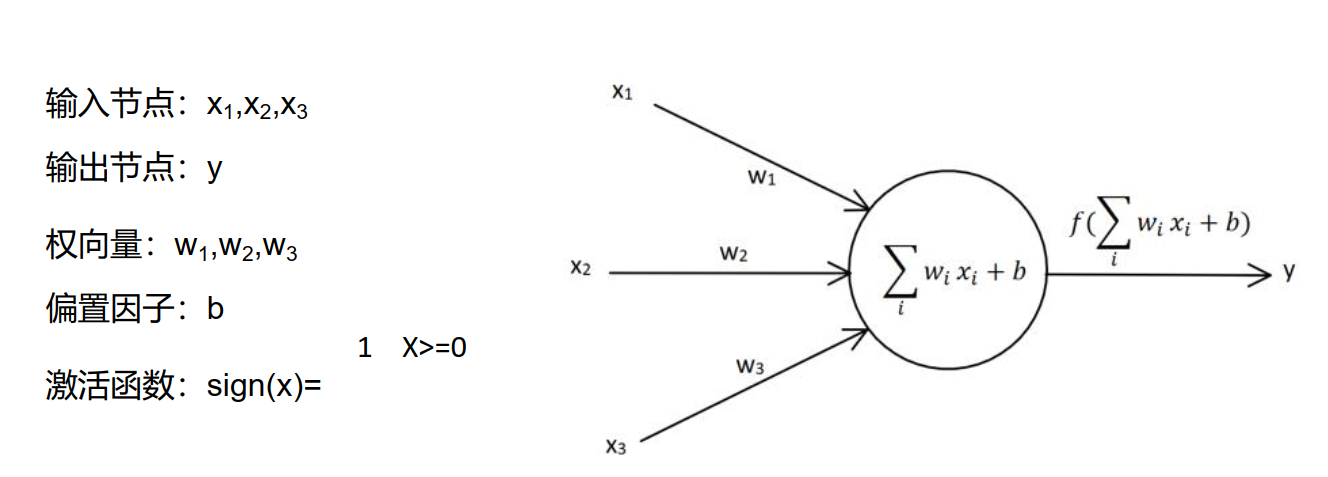
5）LASSO回归

与岭回归类似，都是对标准方程法缺陷的改进，擅长处理具有多重共线性的数据，与岭回归一样是有偏估计。

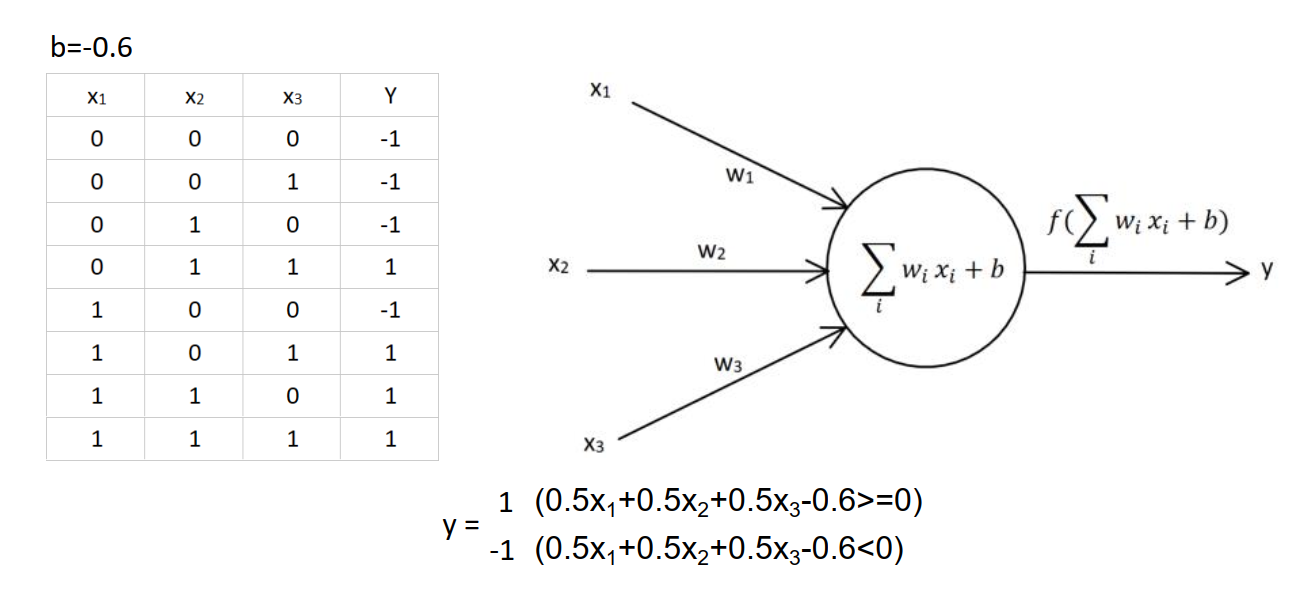
## 神经网络算法

神经网络算法属于监督学习。

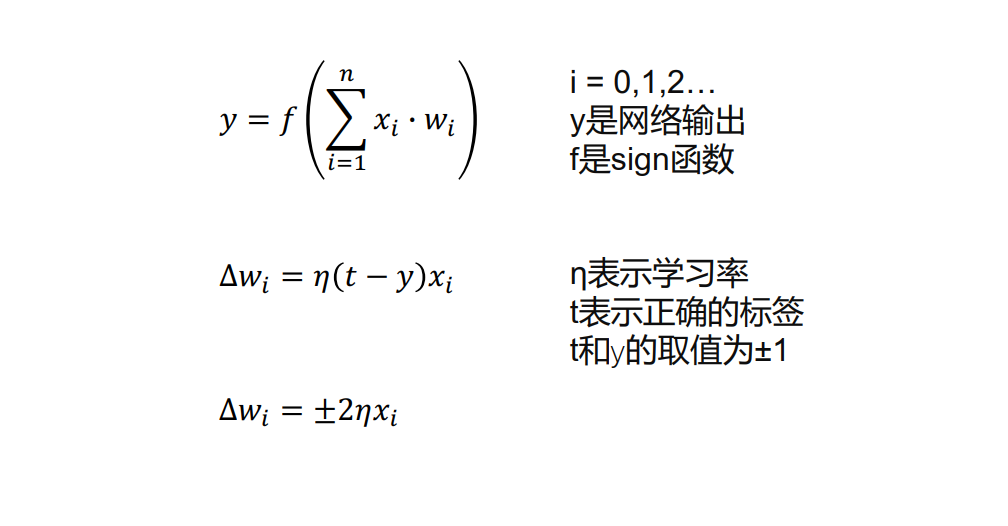
1）单层感知器

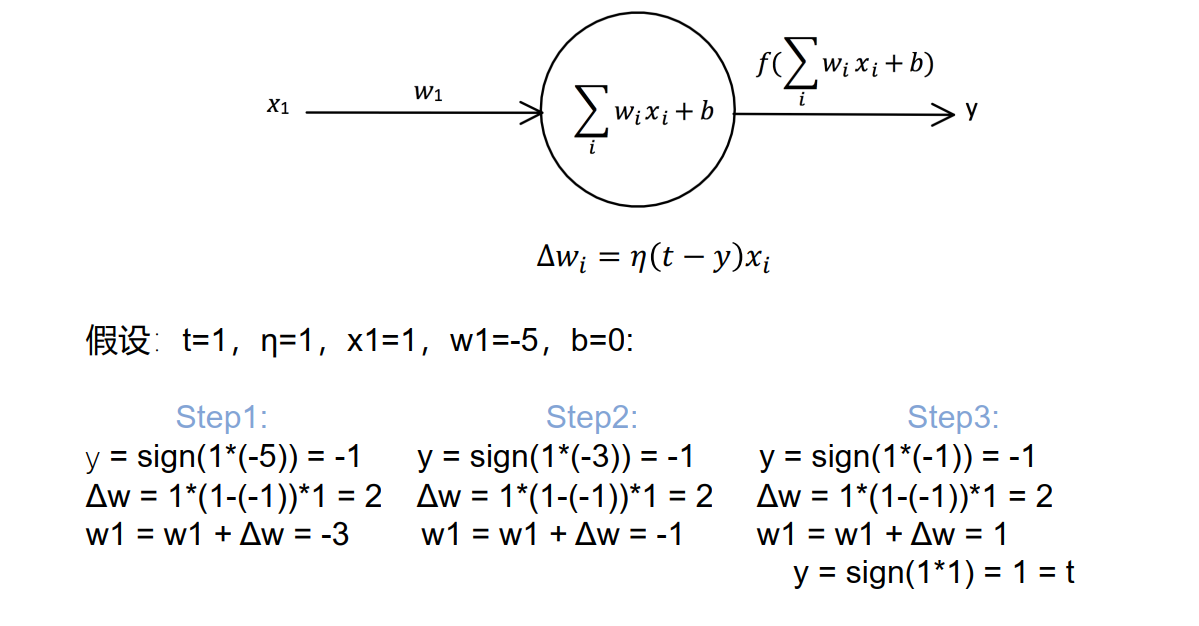


单层感知器举例



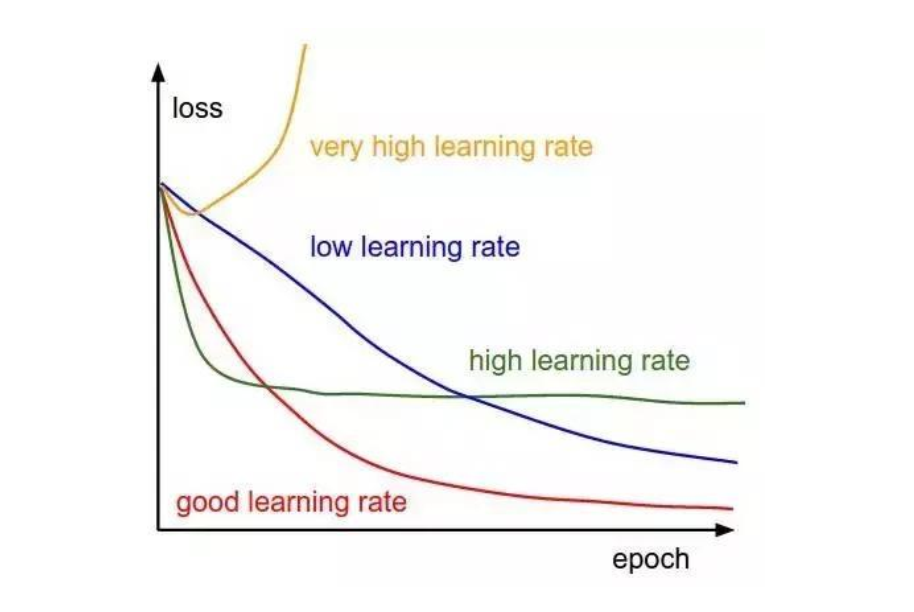
感知器的学习规则





学习率

* η取值一般在0-1之间
* 学习率太大容易造成权值调整不稳定
* 学习率太小、权值调整太慢，迭代次数太多



模型收敛条件

* 误差小于某个预先设定的较小的值
* 两次迭代之间的权值变化已经很小
* 设定最大迭代次数，当迭代超过最大次数就停止

2）线性神经网络

线性神经网络在结构上与感知器非常相似，只是激活函数不同。在模型训练时把原来的sign函数改成了purelin函数：x=y。

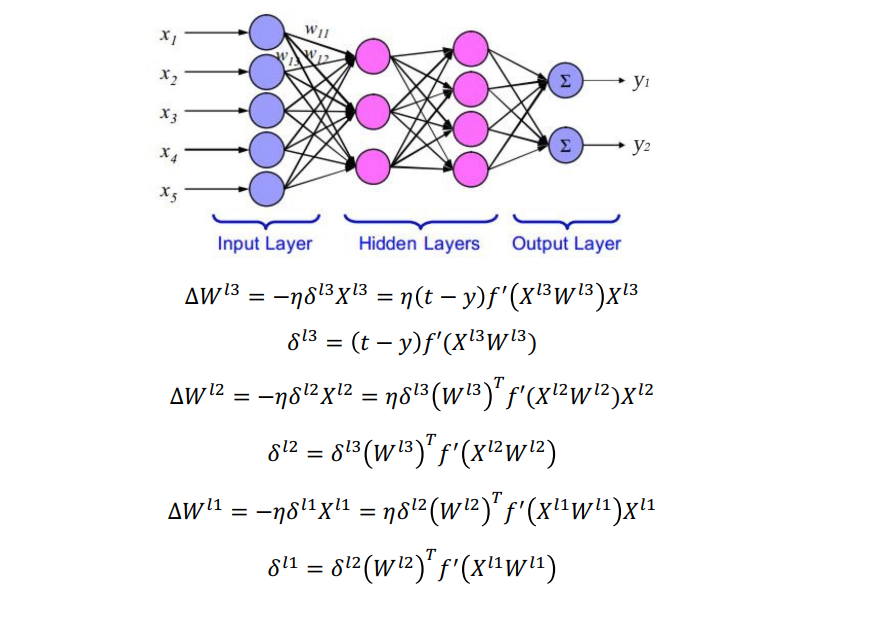
线性神经网络异或问题（线性不可分问题）：

https://www.jianshu.com/p/32c40d407d3e

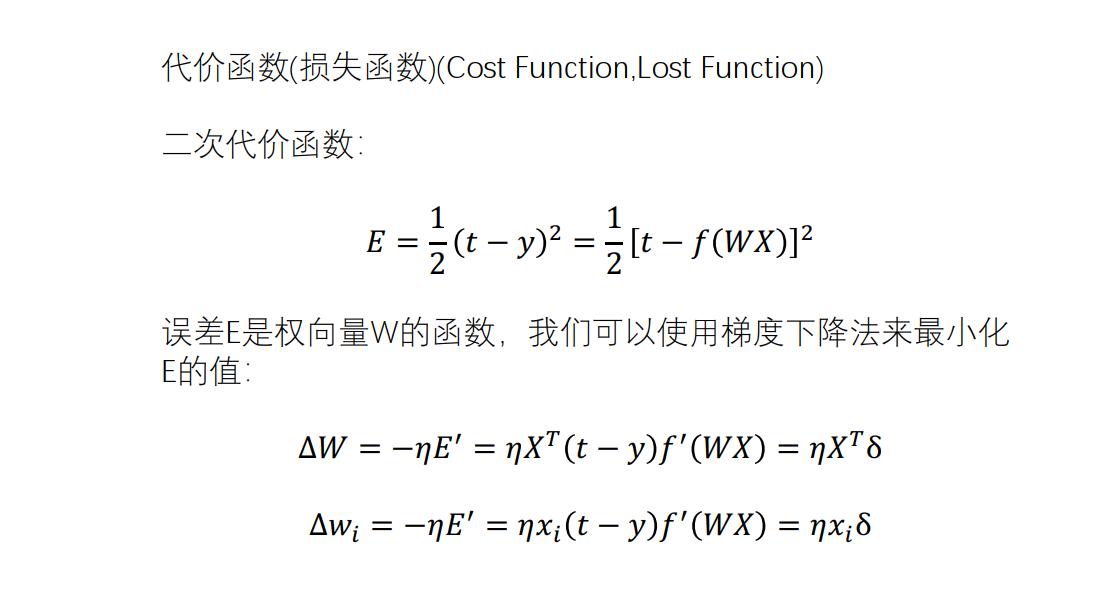
3）BP神经网络

BP（Back Propagation）网络是1986年由Rumelhart和McCelland为首的科学家小组提出，是一种按误差逆传播算法训练的多层前馈网络，是目前应用最广泛的神经网络模型之一。

它的学习规则是使用梯度下降法，通过反向传播来不断调整网络的权值和阈值，使网络的误差平方和最小。BP神经网络模型拓扑结构包括输入层（input）、隐层(hide layer)和输出层(output layer)。



Delta学习规则



# 5 建模

1. 在项目中该如何建模？建模时样本特征如何选择？

用回归中的房屋价格案例举例。

1. 房屋价格受多种因素影响（基本影响因素：大小、卧室数、厅数、房屋年龄；其它影响因素：周边学校数、地铁数等等）。这时候就涉及到特征缩放，选取几个主要特征进行训练。
2. 另外就是不同评价指标有不同的量纲，为了避免量纲大的数据占优势一般采用数据归一化的方式将数据的处理范围处理为0-1、 -1-1。
3. 每个人构建的算法模型都不一样，因为每个人评判标准不一样，比如决策树中根节点的选择，这时候就需要具体问题具体分析了。
4. 建立好模型之后，算法该如何选择？选定算法后算法中的参数该如何选择？

算法的选择根算法的特性有关，比如说采用两种算法都可以实现，就需要对比训练结果的好坏了，这个就得靠工程实践了。

1. 训练的时候训练集和测试集比例该如何确定？

在真正工程应用中，有效数据集比较宝贵一般将数据集分成10份，9份训练，1份测试，另外就是采用交叉验证的方式进行训练。