机器学习

书籍简介：

大数据时代，机器学习是人工智能领域中一个极其重要的研究方向。随着科学技术的发展，各个企业可能积累了大量的数据，如何从这些数据中挖掘出有商业价值的信息或模式，是目前企业所面临的大数据挑战之一。

本书内容涵盖了机器学习各个领域的算法，回归分析，有监督学习，半监督学习，无监督学习，自然语言处理等。内容结构上注重知识的实用性与可操作性，算法结合案例，提供完全的代码实现，帮助读者快速掌握知识并应用到实处。全书采用深入浅出、循序渐进的讲解方式，帮助初学者快速入门学习。

本书一共十九个章节，大致可以分为以下几部分。第一部分（第一章到第三章），介绍机器学习基础知识与数据的基本处理方法；第二部分（第四章到第十八章）介绍了机器学习中的常用算法（回归分析、决策树、随机森林、神经网络、kmeans聚类等），并结合案例讲解具体算法的代码实现，第三部分（第十九章）介绍了概率图模型算法以及机器学习在自然语言处理中的应用案例。

本课程的受众主要是没有经过专业训练的IT专业人员，他们可能是程序员，运维，IT系统架构师等等，也适合没有经过科班训练的数据分析师。数据分析是一个业务+算法+IT的交叉领域，同时熟悉这三方面知识的人，可以玩转大数据，产生无穷无尽的花样，产生巨大的价值。但很无奈的情况是这种人才太少，IT人员即使熟悉本公司的业务，但同时又熟悉算法的人极少。一般只能做一些简单的维度统计，指标计算等等，如果说到开发更高智能的系统，知识上鞭长莫及。本课程的目标，正是要打破知识的鸿沟，向IT人员普及算法知识，并把这些知识用于实际项目，把中国的机器学习应用能力提高一个台阶

书籍目录：

1. 机器学习概论
   1. 机器学习发展现状
   2. 监督学习、半监督学习、无监督学习简介
   3. 统计学基础
   4. 机器学习常用软件介绍
2. 特征工程

2.1 特征类型介绍

2.2 特征的归一化、标准化、离散化等常见变换

2.3 文本特征提取

2.4 图像特征提取

2.5 样本增强

2.6 特征重要性评估与特征选择

1. 模型评估与模型选择

3.1 模型评估：量化模型质量

3.2 交叉检验：评估模型性能

3.3 验证曲线：绘制分数以评估模型

3.4 模型比较检验

3.5 超参数调节方法

3.6 偏差与方差，过拟与欠拟

1. 回归模型

4.1 变量相关性

4.2 一元线性回归与最小二乘法

4.3 多元线性回归

4.4 应用：电子商务业绩预测

1. 岭回归与LASSO

5.1 多重共线性问题

5.2 岭回归

5.3 LASSO

5.4 弹性网

5.5 最小角回归

1. 广义线性回归

6.1 非线性回归

6.2 逻辑回归

6.3 正则化处理

6.4 应用：信用卡违约率预测

1. k最近邻算法

7.1 knn算法

7.2 knn算法的实现：kd树

7.3 应用：Facebook登陆地点预测

1. 贝叶斯分类器

8.1 极大似然估计与贝叶斯估计

8.2 朴素贝叶斯分类器

8.3 贝叶斯网络

8.4 应用：使用贝叶斯分类器判断商品评价情感

1. 贝叶斯统计

9.1 先验分布与后验分布

9.2 贝叶斯推断

9.3 贝叶斯决策

9.4 高斯过程回归

9.5 应用：基于贝叶斯优化的超参数调优

1. 抽样方法

10.1 吉布斯抽样

10.2 蒙特卡洛抽样

10.3 应用：网站转化率A/B测试案例

1. 决策树

11.1 决策树模型

11.2 信息增益

11.3 基尼指数

11.4 ID3算法

11.5 C4.5算法

11.6 CART算法

11.7 预剪枝与后剪枝

11.8 应用：

1. 集成学习与随机森林

12.1 boosting算法

12.2 bagging算法

12.3 随机森林算法

12.4 结合策略

12.5 Stacking算法

12.6 Blending算法

12.7 应用：泰塔尼克号幸存者预测

1. 梯度提升树与XGBoost

13.1 Adaboost算法

13.2 提升树

13.3 梯度提升GBDT算法

13.4 XGoost 解读

13.5 深度森林

13.6 应用：预测道路交通流量，合理规划出行路线

1. 支持向量机

14.1 间隔与支持向量

14.2 线性可分支持向量机

14.3 非线性可分支持向量机与核函数

14.4 应用：基于SVM的股票预测

1. 人工神经网络

15.1 单层感知器

15.2 线性神经网络

15.3 BP神经网络

15.4 误差逆传播算法

15.5 梯度下降学习算法

15.6 应用：商品图片分类器

1. 聚类算法

16.1 层次聚类算法

16.2 k-means算法

16.3 密度聚类

16.4 高斯混合聚类

16.5 期望最大化算法

16.6 孤立点判别

16.7 应用：交通站点聚类分析

1. 半监督学习

17.1 未标记样本

17.2 基于EM算法的半监督学习生成式方法

17.3 半监督SVM

17.4 半监督聚类

17.5 应用：用label propagation进行手写体识别

1. 降维与度量学习

18.1 主成分分析

18.2 因子分析

18.3 基于协同过滤的推荐系统

18.4 应用：课程推荐系统实战

1. 概率图模型

19.1 隐马尔可夫的基本概念

19.2 隐马尔可夫算法

19.3 条件随机场

19.4 应用：概率图模型在自然语言处理中的应用

# 第一章 机器学习概论

## 机器学习发展现状

### 机器学习的定义

机器学习在最近几年越来越引起各行各业的关注，尤其是在Google的AlphaGo以4：1战绩击败围棋界霸主李世石之后，更是零封柯洁，不管是在围棋界还是在IT圈都引起了不小的轰动。

然而机器学习其实并不是最近突然发展起来，从最初被提出到现在已经有70多年，只是发展过程有些曲折。在发展过程中，出现了不少贡献卓越的大师，也分别对机器学习给出

了各自的理解，比如以下几位：

1. Langley：机器学习是一门人工智能的科学，该领域的主要研究对象是人工智能，特别是如何在经验学习中改善具体算法的性能。

2. Tom Mitchell：机器学习是对能通过经验自动改进的计算机算法的研究。

3. Alpaydin：机器学习是用数据或以往的经验，以此优化计算机程序的性能标准。

站在巨人的肩膀上去学习好比站在泰山之巅去欣赏朝霞，会更清晰、更纯粹。

踩着多位大师的脚印，我们可以很清晰的感觉到机器学习是让机器能够对过去的经验进行总结，并指导未来行为的一种思维方式或者过程，从而使机器可以对将来产生的新的信息进行自动识别。

举例说明：假如你是学医的，有一天老师给你一个任务，把全世界可以找到的骨科相关的文献都整理出来。当你听到这个任务的时候估计是要惊掉下巴了，因为全世界范围内

跟骨科研究相关的文献够你数一辈子的，让你一个人一份一份的去整理，估计你也没必要毕业了。但是，当你运用机器学习的思维去做这件事情的时候就要轻松得多。你可以先找出100份文献，将其中可以判定该份文档为骨科相关的信息标注出来，然后利用这些标签在剩下的文档中进行搜索，存在这些标签的文献即为骨科相关。这样一个分类的过程就是一个机器学习的过程。

### 机器学习的现状

类似华山论剑、各大门派尽显绝学，机器学习也是在各大门派的相互竞争中不断地被往前推进。其中，主要的流派有：符号学派、贝叶斯学派、连接学派、进化学派、类推学派。

各学派之间的相互竞争犹如春秋战国时期的诸子百家各显神通，不管在理论还是实践上都成绩显著。不过也因如此瓶颈渐渐凸显，风风火火的大踏步之势戛然而止。在艰难地度过低潮期之后，随着神经科学和概率论的方法广泛应用，机器学习便开始犹如神助，以星火燎原之势飞速发展。神经网络可以更精准地识别图像、语音，做好机器翻译乃至情感分析的工作，

而与此同时，神经网络需要的大量的计算能力也随着硬件平台的飞速发展得到解决，使得大家熟知的天猫精灵、亚马逊的Alexa、Google的Home进入寻常百姓的生活。

### 机器学习的前景

从 2040 年以后，根据普华永道的预测，主流学派将成为 Algorithmic convergence，也即各种算法融合在一起，也成终极算法，届时机器自主学习，也即元学习（Meta-learning）得以实现，计算服务将无处不在。如今，在机器学习的帮助下，无人机可以实时近距离地拍摄例如桥梁之类的地方，然后快速、准确地评估重建项目的范围。同时能够通过“学习”大量的数据，在不需要人为编程的情况下，生成以及识别特定的对象，比如人脸。目前，机器学习也是商业应用中最常用的算法

## 监督学习、半监督学习、无监督学习简介

机器学习从学习方式上分类可以分为以下几类：

1. 监督学习

2. 半监督学习

3. 无监督学习

4. 强化学习

在机器学习的领域中前三种研究的比较多的，也是运用的比较广泛的技术。

### 监督学习

监督学习是机器学习中的一种方法，它是通过对已有数据（这些数据通常是有标签的、也就是有确定的输入和输出的数据）的观察，使机器学习到一个固定的模式（即一个确定的函数或逻辑），然后给机器输入一个新的数据，这样机器就可以得到对这个新的数据进行判断所产生的结果。最常见的应用就是回归分析（输出数据是连续值）和分类（输出数据是离散值）。

1. 回归分析：比如我们想要去预测一下某个城市A未来半年的房价，现在我们手头上有城市A过去十年的房价数据，以及这个城市A相关的一些数据，比如人口、GDP等。这时我们可以猜测这样一个关系：城市A的平均房价（Y）与该城市的人口（X）和GDP（Z）存在某种线性关系，即Y=aX+bZ。然后我们就可以通过对已有的数据代入房价函数进行训练，从而得出a、b的值，这样我就得到了城市A的房价的预测模型，也就可以通过未来半年可预测的人口与GDP数据进行预测未来半年的房价。

2. 分类：比如常见的垃圾邮件分类。当有一部分邮件被我们归入了垃圾邮件一类（这部分数据就是样本），在这些垃圾邮件中我们可以找到一些关键词，然后可以计算出这些关键字在这些邮件中出现的频率，有些关键字出现的频率很高，比如：推销等，这样我们就可以将这些关键词作为标签，当我们再收到一封邮件的时候，就可以扫描这封邮件的内容里面是否存在“推销”等关键词，如果存在就可以把它归入垃圾邮件一类。

常见的监督学习算法有：

1. 回归（线性回归、逻辑回归等）

2. 朴素贝叶斯

3. 决策树

4. K-近邻算法等

通常情况下，不同的场景适合使用不同的算法。在机器学习中很多算法都是基于概率学与统计学，所以不同的算法在不同的场景中的优势不同，最终得出的效果也就千差万别。比如上面提到的对垃圾邮件的自动识别，这显然是属于一个分类问题，那我们就会采用分类算法，比如朴素贝叶斯，这样可以达到最佳效果，给客户提供最好的服务。

### 无监督学习

在监督学习中，我们通过有标签的数据进行分类或者回归。但在现实中已知的存在标签的数据很少，往往不带标签的数据是常态，通过人工去进行标签效率太低、成本也太高，所以我们也希望对没有标签的数据进行归纳总结，形成固定的模式解决通用的问题，比如对新的数据进行分类，这就是无监督学习。

举个简单的例子：我们在学校会有期末考试，我们希望在期末考试中取得好成绩，那我们需要怎么做呢？对，好好学习，天天向上。说具体点，就是上课认真听讲，课后认真做题目。考试成绩取决于做对题目的多少，但是期末考试的题目我们肯定是不知道的，这样就没有了固定的标签供我们有针对性的去学习。但是我们在平时的学习当中就通过了大量的做题渐渐学习到了一些题目的解题方法，这样就形成了我们对这些题目（特征）的标签（就是这个解题方法），所以在我们期末考试的时候，我们就知道了对于某些特征的题目该用什么样的解题方法去应对了。其实在平时的学习过程中总结解题方法就是一个无监督学习的过程。

常见的无监督学习算法有：

1. 聚类（k-means，基于密度的聚类，高斯混合型的最大期望聚类，层次聚类等）

2. 自编码器

3. 主成分分析

### 半监督学习

从上面监督学习与无监督学习就可以推理出，半监督学习就是整合了监督学习与无监督学习的优点。先利用少量带标签的数据进行监督学习，此时的形成的模式会因为样本量少而误差较大；接着我们在利用大量的不带标签的数据，不断往这少量带标签数据的样本中加入不带标签的数据形成新样本，再进行模式训练，用以优化模式来使误差逐渐变小，最终求得最优解。

在现实中，半监督学习应用的场景是非常多的，因为我们接触到的数据大部分都是不带标签的，而只有少部分的数据是带标签的。举个常见的例子：我们在做网页推荐的时候，需要用户自己标注感兴趣的网页，这样就可以有针对性的给用户推荐用户喜欢的内容，但很少有用户愿意花时间去标注，这时我们得到的带标签的样本数据就非常少。如果这时我们仅仅用这些带标签的数据进行监督学习用以预测，往往就会偏差很大，因为训练集的不充分导致机器学习到的模式无法完美地刻画用户的特征。既然这样，我又该如何优化这个问题呢？对，我们还有大量不带标签的数据，如何利用这些不带标签的数据使预测的误差减小、充分地刻画用户行为是我们需要攻克的难题。基于带标签的样本数据，我们已经得到了一个比较完善的模式，然后我们使用不带标签的数据带入前面的模式进行计算并得到误差，这样就得到了一个关于误差的集合，然后我们选取其中误差较小的数据放入前面的带标签的样本数据，形成新的样本再进行训练。之后重复这样的过程，便可以逐步利用全部的数据得到了相对靠谱的模式。这样一个过程就是一个典型的半监督学习过程。

半监督学习常见算法有：

1. self-training（自训练算法）

2. generative methods（生成式方法）

3. SVMs半监督支持向量机

4. graph-based methods（图论方法）

5. multiview learning（多视角算法）

## 统计学基础

### 基本的统计量

以下为鸢尾花数据集中的部分数据，我们将以上面的数据为例介绍统计学中基础的几个概念。其中有五列数据，分别为：Species（种类）、Sepal. Length（花萼长度）、Sepal.Width（花萼宽度）、Petal.Length（花瓣长度）、Petal.Width（花瓣宽度）。

Sepal.Length Sepal.Width Petal.Length Petal.Width Species

5.1 3.5 1.4 0.2 setosa

4.9 3.0 1.4 0.2 setosa

4.7 3.2 1.3 0.2 setosa

4.6 3.1 1.5 0.2 setosa

5.0 3.6 1.4 0.2 setosa

5.4 3.9 1.7 0.4 setosa

4.6 3.4 1.4 0.3 setosa

5.0 3.4 1.5 0.2 setosa

4.4 2.9 1.4 0.2 setosa

4.8 3.1 1.4 0.3 setosa

7.0 3.2 4.7 1.4 versicolor

6.4 3.2 4.5 1.5 versicolor

6.9 3.1 4.9 1.5 versicolor

5.5 2.3 4.0 1.3 versicolor

6.5 2.8 4.6 1.5 versicolor

5.7 2.8 4.5 1.3 versicolor

6.3 3.3 4.7 1.6 versicolor

4.9 2.4 3.3 1.0 versicolor

6.6 2.9 4.6 1.3 versicolor

5.2 2.7 3.9 1.4 versicolor

6.1 3.0 4.9 1.8 virginica

6.4 2.8 5.6 2.1 virginica

7.2 3.0 5.8 1.6 virginica

7.4 2.8 6.1 1.9 virginica

7.9 3.8 6.4 2.0 virginica

6.4 2.8 5.6 2.2 virginica

6.3 2.8 5.1 1.5 virginica

6.1 2.6 5.6 1.4 virginica

7.7 3.0 6.1 2.3 virginica

6.3 3.4 5.6 2.4 virginica

6.4 3.1 5.5 1.8 virginica

#### 期望

在[概率论](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E8%AE%BA)和统计学中，数学期望(mean)（或[均值](https://baike.baidu.com/item/%E5%9D%87%E5%80%BC/5922988)，亦简称期望）是试验中每次可能结果的[概率](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87/828845)乘以其结果的总和，是最基本的数学特征之一。它反映随机变量平均取值的大小。然而随着随机变量取值的不同，期望的计算方式又有所区别。

1. 离散型

如果随机变量只有有限个值或可一定次序一一列出的无限值，这样的随机变量称为离散型随机变量。

即

https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D111/sign=cdc6e7624ded2e73f8e9822db603a16d/08f790529822720ef74fea1c7ccb0a46f31fab57.jpg

以上面的鸢尾花数据为例，计算种属为“setosa”的花萼长度L的期望为：

E(L)=(5.1+4.9+4.7+4.6+5.0+5.4+4.6+5.0+4.4+4.8)/10=4.85

1. 连续型

设连续性随机变量X的概率密度函数为f(x)，若积分绝对收敛，则称积分的值https://gss2.bdstatic.com/9fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D94/sign=6f9509c105087bf479ec5bedf3d387ef/7dd98d1001e93901533fbbb47cec54e736d1967e.jpg为随机变量的数学期望。

即

https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D149/sign=14ad9bc8afd3fd1f3209a63e094f25ce/dcc451da81cb39db20db6fb0d7160924ab183038.jpg

#### 方差

方差是对随机变量或者一组数据离散程度的度量。从上面的例子我们知道，一组数据存在一个均值，离散程度的意思就是这组数据中的每个数据偏离这个均值的程度，这个程度就被定义为方差。

公式：

https://gss2.bdstatic.com/9fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D113/sign=c388d5738013632711edc632a28ea056/023b5bb5c9ea15cee484a9a6bc003af33a87b233.jpg

其中，https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D18/sign=cbff73bb48a7d933bba8e07bac4b41a2/f7246b600c3387442fb466d35b0fd9f9d72aa028.jpg为总体方差，https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D12/sign=19a04804093b5bb5bad724fc37d3f460/4034970a304e251fa45ead57ad86c9177e3e53f7.jpg为变量，https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D13/sign=18b6d1d4db1373f0f13f6b9ca60f5cc6/1b4c510fd9f9d72aa9ac59b2de2a2834359bbb51.jpg为总体均值，https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D14/sign=198a4804093b5bb5bad724fa37d3f40c/4034970a304e251fa474ad57ad86c9177e3e5399.jpg为总体个数。

但在实际应用中，总体数据通常难以得到，通常使用样本来进行计算，使用样本统计量代替总体参数，经校正后，样本方差计算公式：

其中，表示样本方差，X为变量，https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D15/sign=1b37698372f40ad111e4c3e6562c8689/2e2eb9389b504fc2c42eeebeecdde71190ef6d13.jpg为样本均值，n为样本量，n-1表示自由度。

以上面的鸢尾花数据为例，计算种属为“setosa”的花萼长度L的方差为：

=(5.1+4.9+4.7+4.6+5.0+5.4+4.6+5.0+4.4+4.8)/10=4.85

=/(10-1)=0.85

#### 协方差

协方差用来描述两个变量之间的总体误差，是方差的一般形式，方差是协方差的特殊形式，即两个变量相同。协方差可以理解为两个变量在变化过程中，变化的方向以及程度之间的关系。比如有两个变量A和B，当A变大的时候，B是变大还是变小，变大即正相关、变小即负相关；当A变动一个单位的时候，B变动越大，协方差越大 ，即相向程度越大，反之亦然。

公式：

以上面的鸢尾花数据为例，当我们想知道鸢尾花种属为“setosa”的花萼的长度变化之后，花萼的宽度会有什么样的变化呢？这时，我们就需要用到协方差这一统计量。计算结果为：

#### 相关系数

相关系数用来描述两个变量之间的相关关系，最早是由统计学家[卡尔·皮尔逊](https://baike.baidu.com/item/%E5%8D%A1%E5%B0%94%C2%B7%E7%9A%AE%E5%B0%94%E9%80%8A)设计的统计指标，着重用来研究两个变量之间的线性相关关系，用两个变量偏离各自平均值的离差相乘来描述两个变量之间的相关程度，是最为常见的一种相关系数。

上一小节我们介绍的方差，同样是描述两个变量之间的相关关系，为什么这里还要研究相关系数呢？从方差的公式可以看出，最后求得的结果是一个绝对量，比如变量X描述的是股票的波动率，变量Y描述的是股票的收益，这样最后的结果就是波动率乘以收益，这样的结果显然不容易理解，所以就有了相对量的表示方法，即相关系数。

相关系数通过公式表示如下：

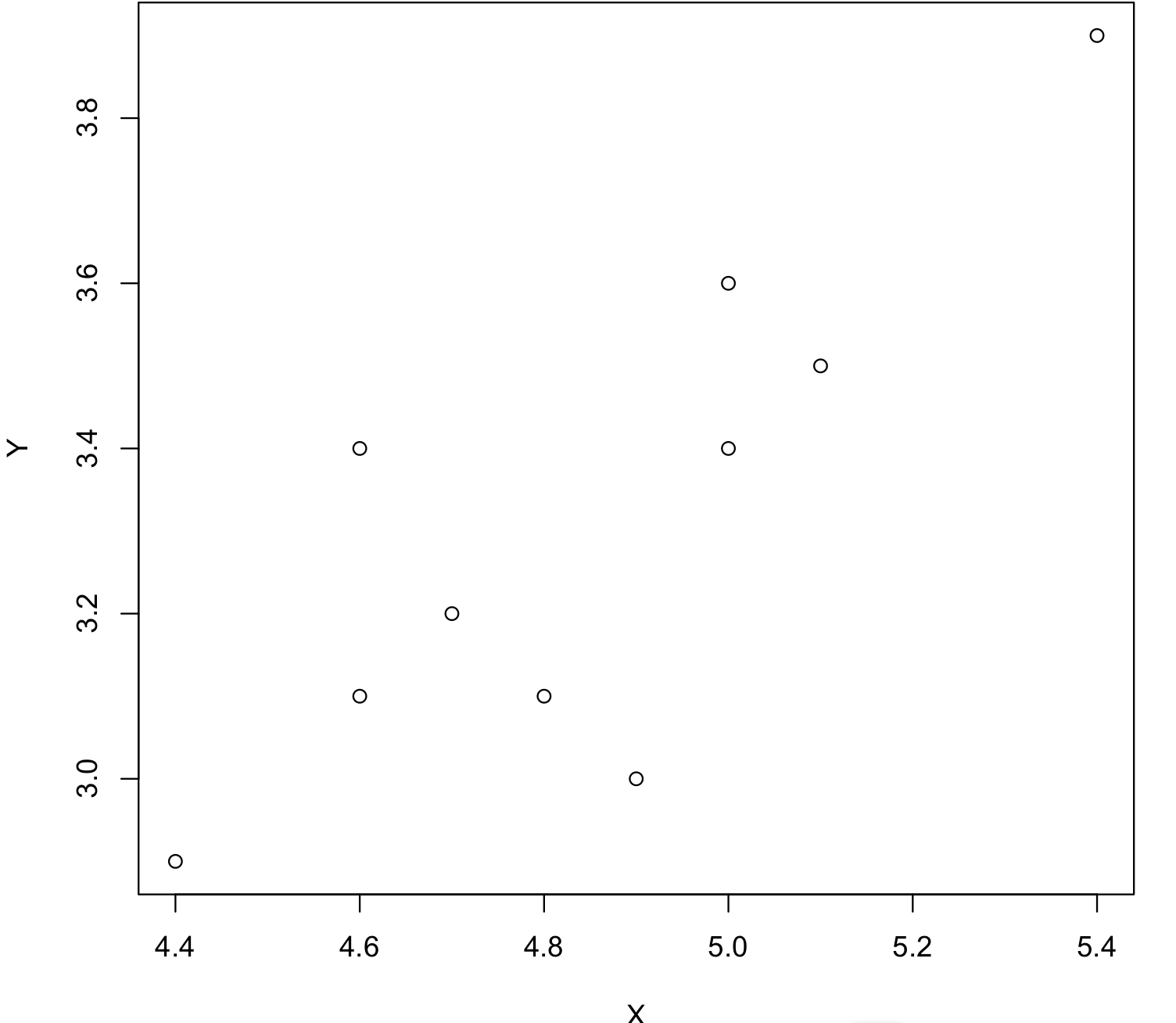
https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D196/sign=0ffe13cf02fa513d55aa68d70b6c554c/14ce36d3d539b600a48289a4e350352ac65cb751.jpg

当然，除了研究线性相关关系的相关系数，同样也有研究曲线相关关系的非线性相关系数，以及研究多元线性相关关系的复相关系数，这里就不做一一介绍。

同样，以上面的鸢尾花数据集为例，我们想知道种属“setosa”花萼长度与其宽度之间是否存在相关关系，通过计算可以得到：

r(X,Y)= 0.8127

因为相关系数的取值范围为「-1，1」，其绝对值越大说明两个变量之间的相关性越高，0.8127对于趋近于1来说已经相对较大，所以我们可以确定花萼的长度与宽度之间存在线性相关关系，这也和我们日常看到的现象相一致。通过如下图标我们可以凭肉眼观察确实如此：



### 常见分布

#### 正态分布

正态分布是自然界最为普遍的一种分布。想要理解正态分布，首先要理解两个概念，一个大数定律，一个中心极限定理。

* 1. 大数定律

大数定律通俗讲，就是在样本量很大时，样本均值与总体的真实平均值充分接近。它告诉我们，在大量随机事件的重复出现中，往往伴随着几乎必然的规律，随机事件的出现的概率近似等于它出现的频率。

举个最常见的例子就是抛硬币，一枚硬币有正反两面，抛硬币一次的结果必然是50%概率正面朝上50%概率反面朝上，这是大家所熟知的现象。但是当我们抛硬币10次，情况却不一样了，有的时候4次正面6次反面，有的时候7次正面3次反面，没有规律而且不稳定。但是当我们抛硬币的次数足够多，出现正面朝上和反面朝上的概率都会无限趋近于50%。

* 1. 中心极限定理

通过大数定律，我们知道在一个总体中随机抽取样本，该样本的均值会无限趋近于该总体的均值，而中心极限定理在此基础上证明了所有样本的均值的分布趋近于正态分布。这里有一个假设是大样本量，即样本量大于等于30。

举个常见的例子，我们要统计全国人民的平均身高，把所有人的身高量一遍显然不现实，所以我们随机选取1000人作为样本，计算这1000人的平均身高。继续，我们重复这样的工作100次，也就是选取了100个这样的样本，最后我们计算这100个样本均值的平均值作为最后的结果。这样的结果会显著的接近于真实的全国人民的平均身高。这时，我们再把这100个样本均值在图上画出来，凭肉眼观察就可以很明显的发现，它的形态与正态分布完全一致。

通常，我们把以均值为、方差为的正态分布描述为

当均值为0，方差为1时，我们将该正态分布成为标准正态分布，即

#### t分布

t-分布又称学生t-分布，是威廉

### 假设检验

## 机器学习常用软件介绍

### Shogun

### Keras

### scikit-learn

### Pattern

### Theano

# 第十章 抽样方法

## 吉布斯抽样

## 吉布斯抽样

## 应用：使用贝叶斯分类器判断商品评价情感