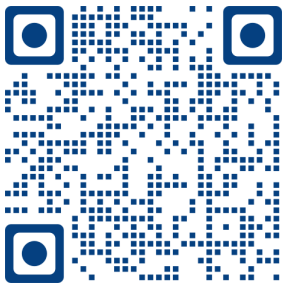
大数据存储和计算

陈一帅

yschen@bjtu.edu.cn

北京交通大学电子信息工程学院网络智能实验室



北京交通大学《大数据存储和计算》课程，源自斯坦福CS245大规模数据挖掘，基于 Spark 讲解大数据机器学习和数据挖掘算法的基本原理和算法，一路下来，带大家在动手中，走上大数据研发的职业道路。详细课程信息请访问：https://yishuai.github.io/bigalgo

目录

A、大数据

人类已经进入了大数据时代。数据就像空气、水、电力、能源一样，成为了最重要的生产要素。本章介绍大数据的特点和存储计算模型，为后面的大数据机器学习算法奠定基础。

一、大数据介绍

本节带大家了解大数据及其应用的特点，参观大数据中心

B站视频

课程PPT：PPT（7MB）

二、存储模型

本节详细介绍大数据系统的存储模型和各项性能指标，然后学习目前最流行的分布式文件系统HDFS的基础知识。这是了解大数据系统的基础，对理解大数据系统性能至关重要。快来看看吧。

B站视频

课程PPT：PPT（1MB）

三、计算模型

本节介绍Map-Reduce计算模型、框架、开销分析和优化。大数据计算就是通过Map-Reduce实现的，所以掌握这些内容非常重要。快来看看吧。

B站视频

课程PPT：PPT（1MB）

B、感知机

感知机模型，从人类大脑神经元得到启发，具有完美几何解释，一手开创了人工智能。这样的模型，不了解行吗？不行。那就让我们开始吧。

四、机器学习基本概念

机器学习是从已知数据中学习出一个函数，然后用这个函数对未知的数据进行预测。本节我们简单了解一下这个概念。

B站视频

课程PPT：PPT（90KB）

五、感知机

感知机模型是一个非常优美、容易理解的机器学习模型。让我们以它为例子，理解什么是机器学习模型吧。很好理解的。试试吧?

B站视频

课程PPT：PPT（1.8MB）

六、感知机的学习

感知机有着非常优美的几何描述。基于该几何描述，我们能够非常轻松地理解机器学习是如何从数据中学会一个模型的。这个过程非常有意思，就像人类一样，它能够从错误中改进自己，取得进步呢！所以犯错误真的是非常棒的，因为错误是最好的学习机会。快来看看吧。

B站视频

课程PPT：PPT（146KB）

七、感知机的优化

感知机模型也有一些不足，比如它只能模型能够线性分隔的数据。这个缺点曾经导致感知机被放弃了很多年，直到深度学习挽救了它。本节我们介绍当数据线性不可分时，如何训练感知机模型，以及多元感知机和非线性感知机。它们让我们理解现实世界中的机器学习任务是非常复杂的，我们需要对数据有清楚的认识，才来训练出好的机器学习模型。这就是成为一个机器学习高手的秘诀。快来看看吧。

B站视频

课程PPT：PPT（168KB）

八、Winnow分类算法

Winnow分类算法和感知机很像，但它使用乘法。当许多维度无关时，它性能更好。它很简单，因此很适合高维数据，在大数据中很常用。快来看看吧。

B站视频

课程PPT：PPT（129KB）

C、支持向量机

具有最优美数学形式的支持向量机分类模型，自从被提出以来，就震惊了整个学术界。人们无法想象，这样美的模型怎么可能被人类发明，然而它确实被发明出来了。叹为观止。本章介绍支持向量机的原理和其梯度下降、随机梯度下降的优化方法。

九、SVM支持向量机

和感知机一样，SVM支持向量机也是要找到一个线性分隔平面。但它比感知机厉害。感知机只要训练集没有错误了，就停止优化了，而SVM还会继续优化，直到找到最佳的分隔平面为止。这是什么意思呢？快来看看吧。

B站视频

课程PPT：PPT（690KB）

十、SVM的学习

本节介绍如何构建SVM的优化问题，找到最优线性分隔平面。这个过程非常有意思。快来看看吧。

B站视频

课程PPT：PPT（523KB）

十一、Hinge Loss

加入Hinge Loss，对越过分隔平面的样本点进行惩罚，这让SVM更能容忍噪声，反映数据的本质特征。Hinge Loss非常有趣，让我们看看吧。

B站视频

课程PPT：PPT（776KB）

十二、SVM Loss

本节我们综合考虑分割平面的距离Loss和样本的Hinge Loss，得到整个SVM模型的Loss函数。通过控制该函数中的C参数，我们可以调节模型对噪声的容忍度，及其泛化能力。该Loss函数是Convex的，所以可以用梯度下降法优化，这就太方便了。快来看看吧。

B站视频

十三、SVM梯度下降优化

本节介绍如何计算SVM Loss函数的梯度，特别是Hinge Loss的梯度。得到了梯度后，我们就可以用梯度下降方法，从数据中学习SVM模型了！快来看看吧。

B站视频

十四、Hinge Loss导数表

本节介绍Hinge Loss导数表。我们将利用这个表，计算所有样本的Hinge Loss的导数。在大数据中，这个表会非常大，所以用Map-Reduce来实现它。了解这个对理解大数据下的SVM模型非常重要。让我们来看看吧。

B站视频

十五、随机和Batch梯度下降

本节介绍随机梯度下降和Batch梯度下降方法的原理、实现和效果。这些方法能够极大地提高模型训练的速度（上万倍），所以是目前机器学习和深度学习中的主流方法，请一定好好理解它们。我们然后为你准备了一个斯坦福大学的SVM三种梯度下降方法的作业。请一定要完成它，这样你才会真正懂得梯度下降和SVM模型。记住，一定完成它！

B站视频

练习：PDF（274KB），Zip（4.4MB）

D. 贝叶斯推断

伟大的贝叶斯定理，一直在人类探索世界的过程中处于绝对核心的位置，在此基础上，人们还提出了贝叶斯网络。它们都是我们探索世界的底层逻辑，做出最优决策的准绳。本章介绍它们及其在文本分类中的应用，即朴素贝叶斯分类器。

PPT（14.4MB）

十六、贝叶斯推断

贝叶斯推断能基于收集到的证据，对特定假设的概率进行估计，比如“昨天是不是下雨了？”。它是统计机器学习的基石，因此是人工智能和机器学习的核心概念。本节通过讲解和举例，带大家理解贝叶斯推断的内涵。快来看看吧。

B站视频

十七、条件独立

本节介绍如何综合考虑多个证据，对特定假设的概率进行估计。为此，朴素贝叶斯分类器引入了条件独立。条件独立让贝叶斯分类变得简单、可扩展，性能还特别好。快来看看吧。

B站视频

十八、文本分类

朴素贝叶斯分类特别适合文本分类。本节通过示例，带大家完成自己的第一个贝叶斯文本分类器。这一方法非常实用，请一定要掌握哦。

B站视频

十九、朴素贝叶斯模型

本节首先介绍朴素贝叶斯文本分类的数学模型，然后介绍机器学习的生成模型和判别模型基本概念，指出朴素贝叶斯模型是一个生成模型，这是它不同于感知机、支持向量机的地方。我们然后给出完整的朴素贝叶斯文本分类模型，包括对零概率的处理。这是我们第一次接触统计机器学习模型。模型的魅力是无穷的。快来看看吧。

B站视频

二十、维数诅咒

本节基于客户流失分类的例子，讲解我们在机器学习中经常遇到的一个非常重要的问题：维数诅咒，即：特征使用越多，数据越稀疏，导致分类器参数的精确估计变得更加困难。然后我们说明朴素贝叶斯是如何解决这个问题。这是一个理解维数诅咒的特别好的例子，快来看看吧。

B站视频

二十一、应用技巧

本节介绍在实际中运用朴素贝叶斯分类方法中可能遇到的两个问题：1）两种类别的先验概率极不平衡；2）连续变量，时的处理方法。这些方法在实际中非常有用。快来看看吧。

B站视频

二十二、贝叶斯网络

贝叶斯网络能够将我们对世界的理解，特别是对各种关系的理解，引入机器学习模型。这个优点非常重要，因为我们特别希望我们的机器学习模型是能够解释，是符合我们理解的世界的规律的。我们前面学过的朴素贝叶斯分类器就是贝叶斯网络中的一种。本节通过讲解和举例，带大家理解并掌握贝叶斯网络。本节内容十分重要，请一定要掌握哦。

B站视频

二十三、马尔科夫毯

本节分析几种贝叶斯网络中常见的元素关系的独立和条件独立，然后给出马尔科夫毯的概念。马尔科夫毯能帮助我们在一个贝叶斯网络中，定位和我们想要推断的元素的相关元素，因此展开测量和模型。本节内容十分重要。快来看看吧。

B站视频

E、降维

降维是机器学习改进性能的重要手段，同时隐含了重要的深度学习概念：表征学习，应用非常广泛。本章学习PCA（主元素分析）和SVD（奇异值分解）两种降维方法，非常有意思。

PPT（571KB）

二十四、降维

机器学习是从数据中进行学习。如果数据包含冗余或无关变量，模型性能会下降。降维能够消除这些变量，提高模型性能。本节通过具体示例，解释为什么应该降维。快来看看吧。

B站视频

二十五、特征值和特征向量

本节介绍降维的数学基础：矩阵的特征值和特征向量。通过它们，我们就可以实现各种神奇的降维效果。本节特别介绍幂迭代（power iteration）计算特征向量的方法。该方法特别适合大数据场景。快来看看吧！

B站视频

二十六、主元素分析

主元素分析（PCA）可以对原始输入数据进行变换，得到原始数据的正交基描述，而且输入数据在这些正交基上的能量还是依次递减的，第一个基就是数据的主元素。所以，我们就可以实现降维。当我们发现原始输入数据中的特征相关时，就应该做主元素分析。这非常重要，快来看看吧。

B站视频

二十七、奇异值分解

奇异值分解可以将一个矩阵分解为三个子矩阵的乘积，其中两个子矩阵分别反映了原始矩阵的行和列的基本信息，而另一矩阵的数值反映了它们在原始矩阵中的重要程度。基于它们，我们可以获得对样本和样本特征的降维描述，方便后续数据的处理和模型的学习，也可以据此进行推荐，非常有趣。快来看看吧。本节我们练习斯坦福的另一个作业。一定要完成哦。

B站视频

PPT（762KB）

PCA和SVD练习：PDF（318KB）

F、推荐系统

哪个机器学习算法最值钱？推荐。我们现在买的大部分东西都不是我们主动寻找的，而是被推荐的。其它呢？读的书、观的影、看的文、交的友、走的路、爱的人、....，一切的一切、都是被推荐的。所以，本节我们学习推荐。快来看看吧。

PPT（3.9MB）

二十八、推荐系统模型

本节介绍推荐系统的基础模型：一个非常稀疏的矩阵描述。推荐系统的作用就是基于有限的数据样本，推断出用户可能最感兴趣的商品。快来看看吧。

B站视频

二十九、基于内容的推荐

本节介绍如何利用TF-IDF等方法，提取、构建用户和商品的内容向量，然后匹配它们，为用户提供推荐，比如我们发现一篇文章是关于贝多芬的，而一位用户的历史表明他很喜欢贝多芬的文章，那么就为他推荐这篇文章。这就是基于内容的推荐。快来看看吧。

B站视频

三十、协同过滤

本节介绍另一种流行的推荐算法：协同过滤（CF：Collaborative Filter）。它通过用户-商品矩阵，发现相似用户或相似商品，进行推荐。该方法在实际中效果非常好。快来看看吧。本节也简短地总结两种推荐系统，讨论推荐中常遇到的另一个问题：冷启动问题。

B站视频

三十一、Netflix推荐大赛

在推荐系统的发展史中，奖金高达100万美元的Netflix推荐大赛占据着史诗般的地位。正是这个大赛，极大地提高了推荐系统的知名度，让推荐系统研究成为显学。本节介绍该大赛的问题设置、基于SGD的协同过滤算法、矩阵分解方法、时变模型、集成方法，并回顾当年百万美元花落谁家的惊险一幕。快来看看吧。

B站视频

推荐系统练习：第3，4题，PDF（318KB），数据（1.1MB）

G. 通过试验学习

朋友，你有没有走进游戏厅，面对满屋老虎机，不知道如何下手？你知不知道，人生也有一点像老虎机，你需要在尝试中学习？你又想不想知道，要玩好人生老虎机，是要乐观呢，还是悲观？没错，这是一个机器学习问题。本节我们就来系统学习这个问题，找到老虎机的最优玩法。

PPT（2.1MB）

三十二、通过试验学习

当你走进一个游戏厅，面对一排老虎机，是不是有一丝茫然：怎么玩才是最优的呢？这是一个通过实验进行学习的问题。本节介绍这一问题的基本概念和应用场景，你会发现原来它这么有用啊。快来看看吧。

B站视频

三十三、多臂老虎机模型

本节介绍多臂老虎机（MAB）问题的模型。它是“通过实验学习”这一问题的经典模型。这个模型非常有趣，快来看看吧。

B站视频

三十四、老虎机的悔恨

本节首先介绍多臂老虎机的性能评估指标：悔恨。这是通过试验学习特有的一个性能指标，非常有趣，也非常有道理。你一定会喜欢的。本节然后定义什么是多臂老虎机问题的最优策略。它非常深刻，很有意思，快来看看吧。

B站视频

三十五、探索与利用

通过实验学习的核心问题是：如何平衡“探索未知”和“利用已知”。这是决策制定中的经典难题。对这个问题的讨论，可能会给你带来极大的启发，改变你的人生！快来看看吧。

B站视频

三十六、epsilon-贪心算法

本节介绍基于实验的学习的一个简单解法方法：epsilon-贪心算法。它会随着时间的增长，慢慢减少探索，增加利用。它可是一种最优算法哦：采用该算法，你最终会找到最优的老虎机。快来看看吧。

B站视频

三十七、UCB 算法

UCB（置信区间上界）算法会选择置信区间上界最大的老虎机。这一方法既探索了未知，又利用了已知，让人拍案叫绝，成为目前应用最广泛的通过试验学习的算法。它也是一种最优算法：采用该算法，你最终会找到最优的老虎机。快来看看吧。

B站视频

三十八、上下文老虎机

上下文老虎机（Contextual Bandit）允许你在做决策时，考虑此时情境。比如在推荐新闻时，考虑读者的喜好。这实在是太强大了，快来看看吧。

B站视频

三十九、LinUCB 算法

本节介绍经典的 Contextual Bandit 算法：LinUCB。它假设回报是情境变量的线性加权和，即线性回归模型。它会按UCB算法选出最好的老虎机，也会根据实验结果不断更新每个老虎机的回报模型。本节也介绍Yahoo是如何用该方法实现新闻推荐的，快来看看吧。

B站视频

四十、结课

恭喜大家完成了大数据机器学习的内容。致敬。再会！

B站视频