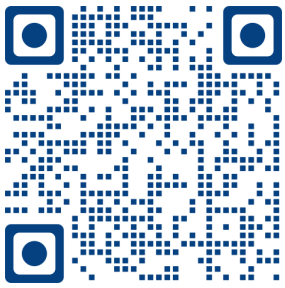
# 大数据存储和计算

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

## 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

## 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

### 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）