

学科的最新动态，论文会议，期刊。

Weka，最新的动态，学术动态。

Q:收敛速度意味着什么?

A：趋于某个目标值的速度

Q:数据挖掘其实是数据学科。

K：学习并不是听

K:训练一个会思考的头脑，而不是简单即东西,清华的老师说话就是不一样。所以就是为什么要跟牛人在一起。

K:预测，分类，决策

K:向比自己优秀的人看齐

公开数据网址

核心知识点之：分类问题的解决：

有标记：有标记的数据很难找到，目前有标记数据还是需要人工

无标记：聚类

关联规则：

Q:我既然知道了用什么算法和解决什么问题，那么难题是什么呢？

A:其实知道用什么算法并对算法特性很了解就已经很厉害了，如果又能非常好的将其应用在合适的问题上，那么难就是调整参数吧。

Q:机器学习是如何解决分类问题？那么多解决分类问题的方法，有什么区别么？

A：先回归然后逻辑化，模型基本是回归和逻辑化的综合起来的结果

Q:loss value是什么？

A:损失函数，一般指误差值，注意既然有误差值，基本都是指监督学习

Q：神经网络跟传统的分类有什么不一样？

A:就目前的理解，神经网络也是一种算法，是一种可以分类，回归，预测的算法。我的理解基本一致，按照机器学习的演讲是可以分为传统算法，和神经网路，在神经网络的基础就有了深度学习的区别，可以这么说，用到了神经网络的的机器学习就是深度学习。从数学的角度来看，是众里寻他千百度，确定未知数系数来确定函数。

<https://yq.aliyun.com/articles/277391?utm_content=m_35973> 这篇文章虽然没有直接回答这个问题，但是说出一个非常好的想法，我们所做的一切都是在提取特征，神经网络提供一种更好提供特征的方式

Q:为什么叫线性回归？

A:参数和未知数的关系是线性的

K:人要会使用工具

K:最**麻烦**的在于数据的预处理，这不是最难的。数据的质量是最容易忽视的

Q:模型是什么意思？

A:模型就是你和出来函数，模型训练过程就是对函数未知数参数确认的过程，一旦函数未知数参数训练后找到合适的值，模型创建完毕。

Q:GPU为什么更擅长计算？

难点之一：计算速度

难点之二：数据的数量与质量，数据预处理可能占到大部分精力

难点之三：世界不是直接坐标系，情况很多是多个维度的

难点之四：直观的数据，也存在哈姆雷特现象，完全相反也是有可能的（常见谬论常识，很有意思，当老师者可以拿来启蒙）

K:算法要选最合适的，简单的分类用神经网络明显就大材小用了

幸存者偏差，有意识

知识改变命运还是学历改变命运

已经看到P10

K:不要小看简单的东西，所有的复杂与困难来源简单，解决也来源简单，很哲学。困难生于简单，没有简单就没有困难。

Q:统计学对数据的描述有哪些？(网上收集一下)

1.分布的集中趋势，反映各数据（众数、中位数、分位数、均值、几何平均数、切尾均值）向其中心值靠拢或聚集的程度。

2.分布的离散程度，反映各数据（极差、内距、方差和标准差、离散系数）远离其中心值的趋势。

3.分布的形状，反映数据分布的偏态和峰态

Q:数据可视化有哪些图？（网上收集一下）

Q:老师说到了很多数据可是化工具在12，等下把他整理一下？

A:一维（matlab），二维（matlab），三维（matlab）==主要是坐标图标，折线，柱状，饼状等等常规统计

三维之上（Box plots，Parallel coordinates）（三维之外的数据目前无法在显示世界描述和展现）

Citespace(软件：比如论文引用图)

Gephi（软件：比如社交工具）

Q:什么是特征选择？

K：我觉得人类无法理解的逻辑可以说是奇淫技巧了，但就是秒不可言呀

Q:降维的方法？或者问多维数据的分析方法？多维数据的分类

Q:决策树能不能通俗的说给我听？常用的决策树有哪些？解决哪些问题？

Q:神经网络能不能通俗的给我讲一下？常用的神经网络有哪些？解决哪些问题？

Q:支持向量机能不能通俗的讲给我听？有哪些？解决哪些问题？不能解决哪些问题？

Q:贝叶斯以及朴素贝叶斯通俗的讲给我听？有哪些？在什么情况下用？解决什么问题？或者说哪些情况不能用？能不能总结一下！（或者用这个算法的解决问题的思路）

Q:算法之多，如果让你挑选几个必须有要学的几个，你会推荐哪些？（考查对整个算法的核心理解，体系理解，掌握少部分）

Q:向量机的核函数的意义

Q:为什么称SVM是线性分类器，什么是线性分类器？为什么是线性？还有其它的线性分类器？有哪些非线性分类器？

K:应用层面就是对工具非常熟悉，知道在最合适的时间，最合适的地方使用，最搞笑的解决问题！

K:数据没有规律，数据间没有关系，何来挖掘呢，何来识别呢

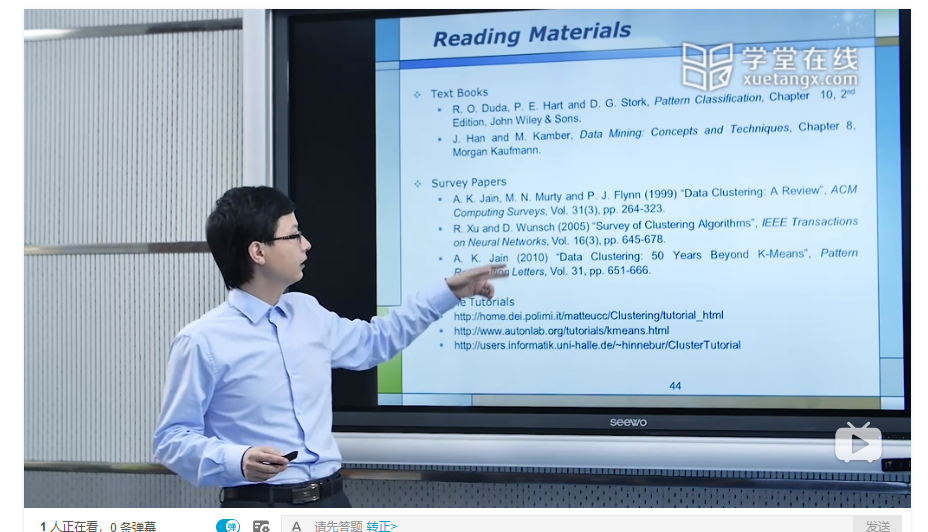
Q:首先你需要知道是，这些函数，数学公式，开源库里都有，那么你要做的是什么？

A：正确的确定模型，确定最终的函数

Q:K-means

Q：期望最大法？高斯模型？密度与层次？都是聚类的算法。

书籍推荐：



P33

Q:数据挖掘十大算法是什么？

K:学算法，要弄清楚概念

Q:Apriori

Q：突然想知道云架构，如何构建一个云服务

Q:推荐的两种方式：

A:基于内容，协同过滤

Q：推荐系统有哪几种思想，比如频次，怎么处理隐含语义分析呢？

Q:处理文本？处理网页？（PageRank）

Q:其实有一种比较学的快的方法，想办法理解统计思路，然后知道用哪些函数就行，因为 函数目前不需要你去设计。

Q:一个不行就多用几个么？

Q:不管是机器学习还是数据挖掘，第一件事就是自己要可以通俗易懂的讲给别人听。（基本上两本书就可以概括）

P44

接下来再讲算法的时候，就自己用通俗的方式在复述一遍

集成学习：把多种解决思路聚合起来以便灵活的应对的问题

群策群议：

决策树：一个属性一个属性的分，比如就拿人来说，根据性别分成男女，根据身高又可以分成低中高，根据学历，又可以分成专科，本科，硕士，博士，博士后等等。

集成学习》Bagging：多个算法分类算法进行投票，举个例子对于一个输入，看他是不是猫，100个分类算法，超过50个算法都说是猫，那么就说是猫，这种算法的好处不用一种算法解决，用最适合的算法的解决一个问题的最适合部分，很少有算法能做到一劳永逸。

集成学习》随机深林：

集成学习》Stacking：迭代分类，选10个分类器进行一次分类，对分类结果进行权重加权，在拿几个分类器在训练

集成学习》Boosting：串行训练，你做错的我来做，你不行我来做

协同过滤：从你的身边人的喜好来推断你的喜好

集成学习》AdaBoost：十大算法之一，参数较少，阿尔法值可以推导出来，老师讲到这个算法时很激动，说明还是比较重要的。

集成学习》RegionBoost：相对于AdaBoost权重会动态改变

学术这种东西，千万不要闭关锁国，老师说研究了很长时间，发现别人已经做过了。

最后一个单元：进化计算（？？？分单元的？？？？）

P48

不是所有的多维问题都可以拆成一维问题