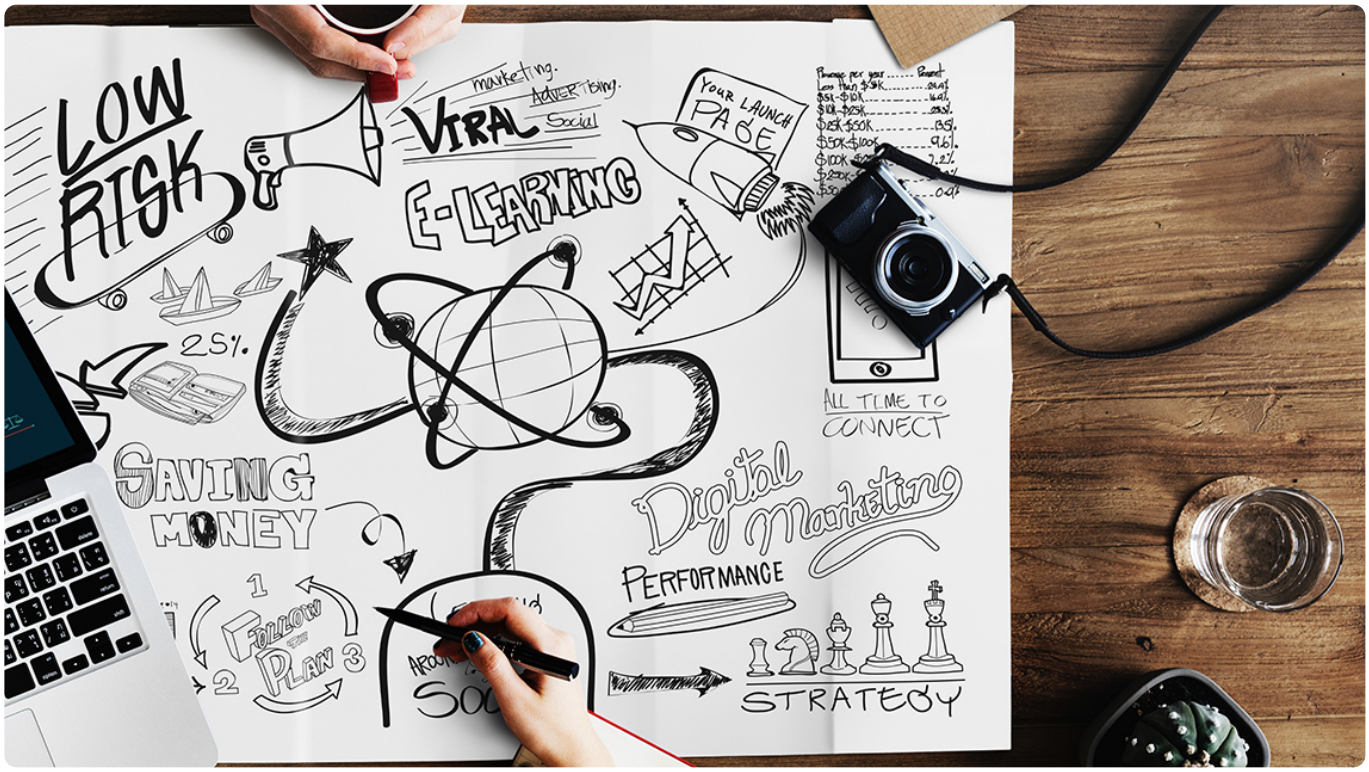
02丨学习数据挖掘的最佳路径是什么?

2018-12-19 陈旸

数据分析实战45讲 进入课程 >



讲述：陈旸
时长 09:33 大小 8.75M

上一节中，我给你分享了数据分析的全景图，其中最关键的部分就是数据挖掘，那什么是数据挖掘呢?

想象一下，茫茫的大海上，孤零零地屹立着钻井，想要从大海中开采出宝贵的石油。

对于普通人来说，大海是很难感知的，就更不用说找到宝藏了。但对于熟练的石油开采人员来说，大海是有坐标的。他们对地质做勘探，分析地质构造，从而发现哪些地方更可能有石油。然后用开采工具，进行深度挖掘，直到打到石油为止。

大海、地质信息、石油对开采人员来说就是数据源、地理位置、以及分析得到的结果。

而我们要做的数据挖掘工作，就好像这个钻井一样，通过分析这些数据，从庞大的数据中发现规律，找到宝藏。

数据挖掘，从知识清单开始

我们第一天学开车的时候一定不会直接上路，而是要你先学习基本的知识，然后再进行上车模拟。

只有对知识有全面的认知，才能确保在以后的工作中即使遇到了问题，也可以快速定位问题所在，然后找方法去对应和解决。

所以我列了一个数据挖掘的知识清单，分别是数据挖掘的基本流程、十大算法和数学原理，以此来开启我们的学习之旅。

数据挖掘的基本流程

在正式讲数据挖掘知识清单之前，我先和你聊聊数据挖掘的基本流程。

数据挖掘的过程可以分成以下 6 个步骤。

1. **商业理解**：数据挖掘不是我们的目的，我们的目的是更好地帮助业务，所以第一步我们要从商业的角度理解项目需求，在这个基础上，再对数据挖掘的目标进行定义。
2. **数据理解**：尝试收集部分数据，然后对数据进行探索，包括数据描述、数据质量验证等。这有助于你对收集的数据有个初步的认知。
3. **数据准备**：开始收集数据，并对数据进行清洗、数据集成等操作，完成数据挖掘前的准备工作。
4. **模型建立**：选择和应用各种数据挖掘模型，并进行优化，以便得到更好的分类结果。
5. **模型评估**：对模型进行评价，并检查构建模型的每个步骤，确认模型是否实现了预定的商业目标。
6. **上线发布**：模型的作用是从数据中找到金矿，也就是我们所说的“知识”，获得的知识需要转化成用户可以使用的方式，呈现的形式可以是一份报告，也可以是实现一个比较复杂的、可重复的数据挖掘过程。数据挖掘结果如果是日常运营的一部分，那么后续的监控和维护就会变得重要。

数据挖掘的十大算法

为了进行数据挖掘任务，数据科学家们提出了各种模型，在众多的数据挖掘模型中，国际权威的学术组织 ICDM （the IEEE International Conference on Data Mining）评选出了十大经典的算法。

按照不同的目的，我可以将这些算法分成四类，以便你更好的理解。

l **分类算法**：C4.5，朴素贝叶斯（Naive Bayes），SVM，KNN，Adaboost，CART

l **聚类算法**：K-Means，EM

l **关联分析**：Apriori

l **连接分析**：PageRank

1. C4.5

C4.5 算法是得票最高的算法，可以说是十大算法之首。C4.5 是决策树的算法，它创造性地在决策树构造过程中就进行了剪枝，并且可以处理连续的属性，也能对不完整的数据进行处理。它可以说是决策树分类中，具有里程碑式意义的算法。

2. 朴素贝叶斯（Naive Bayes）

朴素贝叶斯模型是基于概率论的原理，它的思想是这样的：对于给出的未知物体想要进行分类，就需要求解在这个未知物体出现的条件下各个类别出现的概率，哪个最大，就认为这个未知物体属于哪个分类。

3. SVM

SVM 的中文叫支持向量机，英文是 Support Vector Machine，简称 SVM。SVM 在训练中建立了一个超平面的分类模型。如果你对超平面不理解，没有关系，我在后面的算法篇会给你进行介绍。

4. KNN

KNN 也叫 K 最近邻算法，英文是 K-Nearest Neighbor。所谓 K 近邻，就是每个样本都可以用它最接近的 K 个邻居来代表。如果一个样本，它的 K 个最接近的邻居都属于分类 A，那么这个样本也属于分类 A。

5. AdaBoost

Adaboost 在训练中建立了一个联合的分类模型。boost 在英文中代表提升的意思，所以 Adaboost 是个构建分类器的提升算法。它可以让我们多个弱的分类器组成一个强的分类器，所以 Adaboost 也是一个常用的分类算法。

6. CART

CART 代表分类和回归树，英文是 Classification and Regression Trees。像英文一样，它构建了两棵树：一棵是分类树，另一个是回归树。和 C4.5 一样，它是一个决策树学习方法。

7. Apriori

Apriori 是一种挖掘关联规则（association rules）的算法，它通过挖掘频繁项集（frequent item sets）来揭示物品之间的关联关系，被广泛应用到商业挖掘和网络安全等领域中。频繁项集是指经常出现在一起的物品的集合，关联规则暗示着两种物品之间可能存在很强的关系。

8. K-Means

K-Means 算法是一个聚类算法。你可以这么理解，最终我想把物体划分成 K 类。假设每个类别里面，都有个“中心点”，即意见领袖，它是这个类别的核心。现在我有一个新点要归类，这时候就只要计算这个新点与 K 个中心点的距离，距离哪个中心点近，就变成了哪个类别。

9. EM

EM 算法也叫最大期望算法，是求参数的最大似然估计的一种方法。原理是这样的：假设我们想要评估参数 A 和参数 B，在开始状态下二者都是未知的，并且知道了 A 的信息就可以

得到 B 的信息，反过来知道了 B 也就得到了 A。可以考虑首先赋予 A 某个初值，以此得到 B 的估值，然后从 B 的估值出发，重新估计 A 的取值，这个过程一直持续到收敛为止。

EM 算法经常用于聚类和机器学习领域中。

10. PageRank

PageRank 起源于论文影响力的计算方式，如果一篇文论被引入的次数越多，就代表这篇论文的影响力越强。同样 PageRank 被 Google 创造性地应用到了网页权重的计算中：当一个页面链出的页面越多，说明这个页面的“参考文献”越多，当这个页面被链入的频率越高，说明这个页面被引用的次数越高。基于这个原理，我们可以得到网站的权重划分。

算法可以说是数据挖掘的灵魂，也是最精华的部分。这 10 个经典算法在整个数据挖掘领域中的得票最高的，后面的一些其他算法也基本上都是在这个基础上进行改进和创新。今天你先对十大算法有一个初步的了解，你只需要做到心中有数就可以了，具体内容不理解没有关系，后面我会详细给你进行讲解。

数据挖掘的数学原理

我说了这么多数据挖掘中的经典算法，但是如果你不了解概率论和数理统计，还是很难掌握算法的本质；如果你不懂线性代数，就很难理解矩阵和向量运作在数据挖掘中的价值；如果你没有最优化方法的概念，就对迭代收敛理解不深。所以说，想要更深刻地理解数据挖掘的方法，就非常有必要了解它后背的数学原理。

1. 概率论与数理统计

概率论在我们上大学的时候，基本上都学过，不过大学里老师教的内容，偏概率的多一些，统计部分讲得比较少。在数据挖掘里使用到概率论的地方就比较多了。比如条件概率、独立性的概念，以及随机变量、多维随机变量的概念。

很多算法的本质都与概率论相关，所以说概率论与数理统计是数据挖掘的重要数学基础。

2. 线性代数

向量和矩阵是线性代数中的重要知识点，它被广泛应用到数据挖掘中，比如我们经常会把对象抽象为矩阵的表示，一幅图像就可以抽象出来是一个矩阵，我们也经常计算特征值和特征

向量，用特征向量来近似代表物体的特征。这个是大数据降维的基本思路。

基于矩阵的各种运算，以及基于矩阵的理论成熟，可以帮我们解决很多实际问题，比如 PCA 方法、SVD 方法，以及 MF、NMF 方法等在数据挖掘中都有广泛的应用。

3. 图论

社交网络的兴起，让图论的应用也越来越广。人与人的关系，可以用图论上的两个节点来进行连接，节点的度可以理解为一个人的朋友数。我们都听说过人脉的六度理论，在 Facebook 上被证明平均一个人与另一个人的连接，只需要 3.57 个人。当然图论对于网络结构的分析非常有效，同时图论也在关系挖掘和图像分割中有重要的作用。

4. 最优化方法

最优化方法相当于机器学习中自我学习的过程，当机器知道了目标，训练后与结果存在偏差就需要迭代调整，那么最优化就是这个调整的过程。一般来说，这个学习和迭代的过程是漫长、随机的。最优化方法的提出就是用更短的时间得到收敛，取得更好的效果。

总结

今天我列了下学习数据挖掘你要掌握的知识清单，只有你对数据挖掘的流程、算法、原理有更深的理解，你才能在实际工作中更好地运用，我将在后面的章节中对它们进行一一介绍。

最后给你留道思考题吧。

今天我给你讲了如何学习数据挖掘，你从中有什么样的体会呢？如果某电商网站想挖掘商品之间的关联关系，从而提升销售额，你觉得可以采用上面的哪个算法？为什么？

欢迎在留言区和我讨论，也欢迎点击“请朋友读”，把这篇文章分享给你的朋友或者同事，一起来交流，一起来进步。

上一篇　01 | 数据分析全景图及修炼指南

下一篇　03 | Python基础语法：开始你的Python之旅

精选留言 (159)

写留言

Alex王伟健 置顶
2018-12-19

29

https://mubu.com/doc/y6YuGg_UA0
有些挺耳熟，不过都还给老师了。工作中或者生活中多用应该就忘得少了

编辑回复: 赞

三年二班邱...
2018-12-19

56

老师你好，数学原理里面的内容需要到什么程度，才可以呢？数学这一模块是我很担心的，因为数学实在不怎么样。不知道有什么书籍可以提升这个方面的知识呢。以后常用的也就是这十大算法吗？

展开

作者回复: 如果很多人都有这个情况的话，我想抽个时间，给你整理一篇“白话数学基础：数学基础不好的人，如何理解数据挖掘算法”

Cathy
2018-12-19

17

体会：
①学渣与学霸最大的区别不是智商，而是学习方法和学习态度。作为一名计算机出身的工科女，曾经差点溺死在各类算法的海洋里，目前初入社会做产品，又差点迷失在数据的大山。个人还需要调整自己的学习方法和学习态度。
②当前个人接触的仅仅是数据收集、数据处理、数据分析、数据展现，看到老师的数...

展开

作者回复: 总结的不错👍

JingZ
2018-12-19

12

(1)数据挖掘学习方法体会：有了知识清单，相当于有了一个系统思维在那，对快速识别问题的确很有帮助~很好的方法方便实践，就像巴菲特和芒格的投资是使用的公司尽调清单一样，MECE的解决问题
(2)基于电商商品的关联进行推荐从而提高销售的话，个人认为是Apriori算法，其为了提取频繁项集和一定置信度的关联规则，即用户购买了X产品有多大概率去买Y，根据置信度...

展开

作者回复: 总结的很好，大家可以看下。尤其是用到了MECE原则👍

vincent

11

2018-12-19

数学太差，毕业很久了怎么学习呢?

展开

作者回复: 不用担心，很多人都有这个问题。我觉得你可以尝试：
1、培养兴趣：兴趣是最好的老师，我们大自然的很多科学都是和数学相关，比如为什么雪花是六边形?
2、刻意训练：你不需要通过做项目来做完整的数学训练，比如你和朋友去吃饭的时候，你可以脑算下一共花了多少钱？很多时候，心算是数学的一个能力
3、价值暗示：数学可以帮你很多，尤其是在算法效率、代码质量上。很明显，数学好的人，写出来的算法效率也更高。
所以一个代码完成后，你可以问自己个问题：还有没有更好的方法?

Key.
2018-12-19

10

理解了数据比选择算法建立模型更重要。我觉得电商网站可以采用Apriori算法，因为通过挖掘频繁项集，可以探索到物品之间的联系，从而为商家提供销售思路!

展开

作者回复: 是的，Aprior是个挖掘商品关联关系的常用算法

五岳寻仙
2018-12-19

9

总结与思考：

1. 商业理解：如老师之前所讲，数据挖掘是工具，要么帮我批处理，要么拓展我们思考的规模。也就是说问题本身是人能够处理得了的，只是受限与时间太长或者规模太大，需要借助计算机。人工智能是人思考的放大，如果一个问题人都想不通，指望借助人工智能...

展开

HxScript
2018-12-19

7

文中的引子我深有体会：
我本科就是学的石油工程。油藏的勘探、储量预测、钻井、采油的确对应了数据挖掘的发

现业务中的key points、收集业务中的相关数据并建模、再将模型反代入业务进行模型持续的评估、输出可视化的数据分析结论以及报告。

...

展开

Robin
2018-12-20

5

apriori

展开

sarach
2018-12-19

4

一直对数据挖掘感兴趣，但没有找到合适的学习方法，通过这节课 系统的对数据挖掘算法整体有了个认识；希望之后的每一天都可以进步· ~

课后思考题：...

展开

花生
2019-02-11

3

觉得最难的不是算法，而是数据到算法选择过程中的衔接工作，比如特征工程。还有就是得到分析结果并不难，解释结果怎么来的，合理性分析很难。

双木公子
2019-01-20

3

发现我天然具有学数据挖掘的条件，基础数学理论知识掌握的比较牢固，算法中的图论知识也比较感兴趣。

Chen
2018-12-26

3

决策树这块，C4.5和CART主要不同在哪呢？一般什么时候用C4.5，什么时候用CART呢？CART即是分类树，又是回归树，是即可以解决分类问题，又可以解决回归问题吗？怎么用呢？

十二先森
2018-12-19

3

我大学不是计算机专业，学习这个概率和统计学从哪方面下手

展开

作者回复: 感谢关注，其实高中的时候，我们也会接触简单的概率论知识。这里你可以带着问题，去思考。先知道每个概念代表的意义即可，如果不能推导公式，没有关系。不影响你对“条件概率” “联合概率”的理解，也不会影响你使用这些工具，因为在python中都有相应的类库
在使用的基础上，如果你想进一步探索概率论的原理，可以自己推导下这些公式，也可以多做一些相关练习，来加强自己的理解

追梦小乐
2019-01-01

2

咦， 怎么没有隐马尔科夫HMM?

展开

作者回复: 这里只介绍十大经典算法，有一些算法没有放进去，深度学习，HMM这些确实用的也比较多

denzel.m...
2018-12-23

2

https://mubu.com/doc/fplKTT3GIn
楼上推荐的幕布真是个好应用，总结特别方便，理解也更加深刻了，还可以导出思维导图。

展开

作者回复: 幕布确实很好用，我也是看到留言中不少人在用👍

Louie Zha...
2018-12-21

2

可以使用Apriori算法得到各样品之间关联的程度大小，关联性越大，那么可将该对应商品

捆绑销售，可达到提升销售额的目的。还望老师批评指正，谢谢！

凛冬里的匍...

2018-12-19

2

C4.5算法中的剪枝是什么意思？机器学习中的梯度下降法是不是也是以最优化方法为数学基础的？

Jane

2018-12-19

2

随机森林，xgboost这种在经典算法基础上衍生出来的算法老师能不能在讲基础算法的时候拓展介绍一下啊。
应该可以通过Apriori将相关商品关联起来，比如亚马逊“购买过此类商品的人通常也会购买”这种商品推荐。

> 作者回复: 很好的建议！其实不光是随机森林，xgboost，还有逻辑回归都是很常用的算法，有时间一起介绍下

hillw4h

2018-12-19

2

最难的部分应该是数据获取吧？

展开

> 作者回复: 不同阶段，各有各的难点。数据获取是前提，在数据获取中，更主要的是各个网站的反爬虫机制。尤其针对手机，你还需要进行模拟。切换不通的IP，有些网站需要多个账号登录，你就需要准备多个账号轮流切换。
> 另外在数据挖掘过程中，模型算法的选择也很重要。有时候选择比努力更重要，不是说你参数优化调整不好，而是在最开始的时候，就可以选择一种适合的算法模型，这也是为什么，很多人针对一个项目，会使用多种算法，看下哪个算法的效果好，再确定采用哪个模型