安德鲁·摩尔

家
传记
教程
论文
有用
常见问题
接触
推特

下面的链接指向一组关于统计数据挖掘的许多方面的教程，包括概率的基础、统计数据分析的基础以及大多数经典的机器学习和数据挖掘算法。

下面是.ppt(PowerPoint)形式的所有教程。请随意使用和调整他们任何你喜欢的方式，但请包括一个指针回到这个网页。谢谢

这些算法包括决策树、神经网络、贝叶斯分类器、支持向量机和基于案例的(也称非参数)学习。它们包括多元多项式回归、MARS、局部加权回归、GMDH和神经网络等回归算法。它们还包括其他数据挖掘操作，如聚类(混合模型、k均值和分层)、贝叶斯网络和强化学习。

我希望它们是有用的(如果它们是有用的，或者如果你有建议或错误更正，请告诉我)。单击此处可获得简短的主题列表。

- 决策树。决策树是目前数据挖掘和机器学习中最常用的分类算法之一。本教程可用作数据挖掘的风格和术语的自成一体的介绍，而不需要审查许多统计或概率先决条件。如果你是新的数据挖掘，你会喜欢它，但你的眉毛会扬起，这一切是多么简单！在定义了分类工作之后，我们解释了如何使用信息增益(下一个Andrew教程)来查找预测输入属性。我们展示了如何递归地应用这个过程来构建一个决策树来预测未来的事件。然后，我们仔细研究一个非常基本的问题，它是所有统计学和机器学习理论的基础：如何在一个非常适合数据的复杂模型和一个“Occam‘s剃刀”模型之间进行选择，该模型简洁但不擅长拟合数据(这个主题将在安德鲁以后的讲座中重新讨论，包括“交叉验证”和“VC-维度”)。我们还讨论了非常广泛的世界的改进和调整的基本决策树的想法。

- 信息增益本教程详细介绍了信息论的思想，最终导致信息获取.是目前数据挖掘中使用的最流行的关联度量之一。我们沿途考察了熵和条件熵的思想。在连续概率密度函数的情况下，请看关于高斯的讲座来讨论熵。

- 数据矿工的可能性。本教程回顾从地面开始的概率。可以说，在尝试数据挖掘、机器学习或应用统计的高级算法之前，完全满足于概率是一项有用的投资。除了为在剩下的教程中一遍又一遍地使用技术设置舞台之外，本教程还介绍了密度估计作为一种重要操作的概念，然后介绍了贝叶斯分类器，例如过度拟合的联合密度Bayes分类器和抗过拟合的朴素Bayes分类器。

- 概率密度函数回顾你以前可能遇到的世界：实值随机变量，概率密度函数，以及如何处理多元(即高维)概率密度。在这里，您可以查看诸如期望值、协方差矩阵、独立性、边际分布和条件分布之类的内容。一旦你对这些东西感到满意，你就不会成为一个数据挖掘者，但是你将拥有很快成为一个工具的工具。

- 高斯人。高斯，无论是友好的单变量类型，还是slightly-reticent-but-nice-when-you-get-to-know-them多元类，在统计数据挖掘的许多方面都是非常有用的，包括许多数据挖掘模型，在这些模型中，底层的数据假设是高度非高斯的。你需要和多元高斯人做朋友。

- 最大似然估计MLE是学习数据挖掘模型参数的有力工具。这是一种试图做两件事的方法。首先，当您想从数据中学习某种模型时，这是一种相当有原则的方法，可以计算出您应该做什么计算。其次，它通常是相当容易计算的。无论如何，重要的是，为了理解多项式回归，神经网络，混合模型，隐马尔可夫模型和其他很多东西，如果你对MLE很满意的话，它真的会有帮助。

- 高斯贝叶斯分类器一旦你与高斯人成为朋友，就很容易将它们作为贝叶斯分类器的子组件使用。本教程向您展示了如何。

- 交叉验证。交叉验证是几种方法之一，用于评估您从一些培训数据中学到的模型在未来尚未见过的数据上的表现。我们将回顾测试集验证，留出交叉验证(LOOCV)和k折叠交叉验证，我们将讨论这些技术可以使用的各种地方。我们还将讨论简历应该呈现的可怕现象。最后，我们的头发会站在一边，因为我们意识到，即使在使用简历时，你仍然可能过分随意地不健康。

- 神经网络我们从线性回归开始.神经网络的祖先。我们研究线性回归如何使用简单的矩阵运算从数据中学习。我们高兴地咯咯地笑着，因为我们看到了为什么一个初始假设不可避免地导致了试图最小化平方误差之和的决定。然后，我们探索了另一种计算线性参数的方法-梯度下降。然后我们利用梯度下降来允许分类器，除了回归器，最后允许高度非线性的模型-全神经网络。

- 基于实例的学习(又称基于案例的学习或基于内存的学习或非参数学习)。一个多世纪以来，这种数据挖掘形式仍然得到统计学家和机器学习者的广泛使用。探讨了最近邻学习、k最近邻、核方法和局部加权多项式回归.本教程中算法的软件和数据可从 http：/www.cs.cmu.edu/~AWM/vizier。此幻灯片集中的示例图形是用相同的软件和数据创建的。

- 八种回归算法你得等着找出安德鲁的排序，但根据你到目前为止所涵盖的所有基础，我们很快就能通过：回归树、级联相关、组方法数据处理(GMDH)、多元自适应回归样条(MARS)、多元线性插值、径向基函数、稳健回归、级联相关+投影寻踪。

- 预测实际价值产出：回归简介。这个讲座完全是从神经网络讲座开始的材料和"最喜欢的回归算法"讲座中的一个主题子集组成的。我们讨论线性回归，然后讨论这些主题：变噪声、非线性回归(非常简单)、多项式回归、径向基函数、稳健回归、回归树、多线性插值和MARS。

- 贝叶斯网络本教程首先回顾概率的基本原理(但要正确地做到这一点，请参阅前面关于概率数据挖掘的Andrew讲座)。然后讨论了联合分布在不确定知识表示和推理中的应用。在讨论了联合分发作为一种通用工具的明显缺陷(维度的诅咒)之后，我们访问了涉及独立性和条件独立性的聪明技巧，这些技巧使我们能够更简洁地表达我们的不确定知识。当我们意识到我们已经掌握了我们所需要的大部分知识来理解和欣赏贝叶斯网络时，我们感到非常高兴。本教程的其余部分介绍了如何使用贝叶斯网络进行推理的重要问题(也请参阅下一节安德鲁讲座)。

- 贝叶斯网络中的推理(斯科特戴维斯和安德鲁摩尔)。这些幻灯片大部分是由斯科特·戴维斯。一旦你掌握了一个贝叶斯网络，还有一个问题就是你如何使用它进行推理。推理是用已知值给出一些属性子集的操作，我们必须使用Bayes网来估计一个或多个剩余属性的概率分布。一个典型的推论是"我发烧101，我是37岁的男性，我的舌头感觉有点奇怪，但我没有头痛。"我得了黑死病的可能性有多大？"

- 学习贝叶斯网络这篇简短的教程概述了从数据中学习贝叶斯网络的问题，以及使用的方法。这是包括安德鲁和他的学生在内的许多研究小组积极研究的领域。奥顿实验室网站了解更多细节)。

- 简单介绍朴素的贝叶斯分类器。我建议使用数据挖掘概率为了更深入地介绍密度估计和Bayes分类器的一般用途，以朴素Bayes分类器为特例。但如果你只是想要执行摘要的底线学习和使用朴素贝叶斯分类器的分类属性，那么这些是你的幻灯片。

- **Bayes**网简介。这是一个非常短的5分钟的"执行概述"的直觉和洞察力背后的贝叶斯网络。阅读全文贝叶斯网络教程想了解更多信息。

- 高斯混合模型高斯混合模型(GMMS)是统计上最成熟的聚类方法之一(尽管它们也被广泛用于密度估计)。在本教程中，我们介绍了聚类的概念，并了解了聚类的一种形式.在这种聚类中，我们假设单个数据点是通过首先从一组多元高斯人中选择一个，然后从它们中取样而产生的.可以是一种定义良好的计算操作。然后，我们将看到如何从数据中学习这样的东西，并且我们发现以前的Andrew教程中没有使用的优化方法在这里会有很大的帮助。这种优化方法称为期望最大化(EM)。我们将花一些时间给出一些关于EM的高级解释和演示，这对于除高斯混合模型之外的许多其他算法都是很有价值的(我们将在后面关于隐马尔可夫模型的Andrew教程中再次遇到EM)。可以找到正文中提到的疯狂的"n"代数(手写的)。这里.

- **K-**均值和层次聚类。K-均值是最著名的聚类算法.在本教程中，我们回顾了集群所试图实现的目标，并详细说明了k均值方法聪明地优化一些非常有意义的东西的原因。哦，是的，我们会告诉你(并告诉你)k-的算法实际上是做什么的。您还将了解另一个著名的集群类：分层方法(生命科学中最受欢迎的方法)。"层次聚集聚类"和"单链接聚类"等短语将被广泛使用。

- 隐马尔可夫模型在本教程中，我们将首先回顾马尔可夫模型(也称马尔可夫链)，然后.我们将隐藏它们！这模拟了一个非常普遍的现象.有一些潜在的动态系统是根据简单和不确定的动力学而运行的，但我们看不到它。我们所能看到的只是一些来自底层系统的噪音信号。从这些嘈杂的观测中，我们想要做的事情是预测最有可能的潜在系统状态，或者状态的时间历史，或者下一次观测的可能性。它在故障诊断、机器人定位、计算生物学、语音理解等领域有着广泛的应用。在本教程中，我们将描述如何愉快地处理与HMM有关的大多数无害的数学，以及如何使用一种名为动态规划(DynamicProgramming，DP)的方法来高效地完成您可能想做的大部分HMM计算。这些操作包括状态估计，估计潜在状态的最可能路径，并作为一个宏大的(和EM填充的)终结，从数据中学习hmms。

- **VC**维数本教程涉及一个著名的机器学习理论.如果一方面有学习算法，另一方面有数据集，那么在多大程度上可以判断学习算法是否存在过度拟合或不适当的危险？如果你想对这个迷人的问题进行一些正式的分析，那么这是你的教程。除了对过度拟合现象有很好的理解外，您还得到了一种估计算法对未来数据的性能的方法，该方法完全基于其训练集误差，以及学习算法的一个属性(VC维)。因此，VC维为选择分类器提供了一种替代交叉验证的方法，称为结构风险最小化(Structure Risk Minimation，SRM)。我们会讨论的。我们还将非常简单地比较CV和SRM与其他两种模式选择方法：AIC和BIC。

- 支持向量机我们回顾了分类器的边缘的概念，以及为什么这可能是衡量分类器的可取性的一个很好的标准。然后，我们考虑了寻找最大边缘线性分类器的计算问题。在这一点上，我们尴尬地看着自己的脚趾，注意到我们只做了适用于无噪音数据的工作。但我们振作起来，展示了如何创建一个抗噪声分类器，然后是一个非线性分类器。然后我们在显微镜下观察SVMS著名的两种东西-将数据投射到万亿维的计算能力和统计能力，在乍一看是一个典型的过度拟合陷阱。

- **PAC**学习PAC的意思是"可能大致正确"，并考虑到一种很好的形式，用于决定您需要收集多少数据，以便给定分类器能够实现对未来测试数据的给定部分进行正确预测的给定概率。由此得出的估计有些保守，但仍然代表着一种有趣的途径，计算机科学试图通过这一途径深入研究统计部门通常会发现的那种分析问题。

- 马尔可夫决策过程如果你的行动的结果是不确定的，你如何有效地计划？有一些非常好的消息，也有一些重大的计算困难。我们首先讨论马尔可夫系统(没有行动)和马氏系统的概念。然后，我

们激励和解释无限地平线的想法，折扣未来的回报。然后我们考虑两种相互竞争的方法来处理以下计算问题：给定一个带有奖励的马尔可夫系统，计算预期的长期折扣报酬。这两种方法都是直线代数法和动态规划法。然后我们跳到马尔可夫决策过程，发现我们已经完成了82%的工作，不仅计算了每个MDP状态的长期回报，而且还计算了在每个状态下采取的最优行动。
    - 除了这些幻灯片外，关于强化学习的调查，请参见本文或萨顿和巴托的书.
    - 马尔可夫决策过程的可视化仿真及RoHIT Kelkar和Vivek Mehta的强化学习算法。

- 强化学习在进行强化学习之前，您需要对马尔可夫决策过程(之前的Andrew教程)感到高兴。它涉及一个迷人的问题，你是否可以训练一个控制器，使其在一个可能需要吸取一些短期惩罚才能获得长期回报的世界中发挥最佳的作用。我们将讨论确定性-等价RL，时间差异(TD)学习，最后Q-学习。维度的诅咒将是不断地从我们的肩上学习，垂涎三尺，咯咯地笑。
    - 除了这些幻灯片外，关于强化学习的调查，请参见本文或萨顿和巴托的书.
    - 马尔可夫决策过程的可视化仿真及RoHIT Kelkar和Vivek Mehta的强化学习算法。

- 生物监测：一个例子。我们回顾了其他生物监测幻灯片中描述的方法，这些方法适用于2000年WalkertonCryptosporidium爆发的住院数据。这是作为ECADS项目的一部分执行的工作。

- 初等概率和朴素贝叶斯分类器。这张幻灯片重复了安德鲁的教程系列的主要概率幻灯片的大部分材料，但这套幻灯片集中于疾病监测的例子，包括一个非常详细的描述给非专家如何在实践中使用Bayes规则，关于Bayes分类器，以及如何从数据中学习朴素的Bayes分类器。

- 空间监视本教程讨论扫描统计，这是一种著名的流行病学方法，用于发现疾病病例的过度密度。

- 时间序列方法本教程回顾了一些基本的单变量时间序列方法，重点是在观察序列开始异常时使用时间序列来发出警报。

- 游戏树搜索算法，包括**Alpha-Beta**搜索。电脑游戏算法简介。我们描述了关于完全信息的二人零和离散有限确定性对策的假设.我们还练习一口气说名词短语.在恢复团队完成了他们的工作之后，我们讨论了用极大极小和α-β搜索来解决这样的游戏。我们还讨论了动态规划方法，这是最常用的结束游戏.我们还讨论了启发式评价函数在游戏中的理论和实践。

- 零和博弈论。想知道如何和为什么在扑克中虚张声势？如何将游戏编译成矩阵形式？并对游戏中隐藏信息的基本知识进行了一般性的讨论？这是给你的幻灯片。它可能会帮助你从阅读开始游戏树搜索的幻灯片。

- 非零和博弈论。拍卖和电子谈判是一个引人入胜的话题。这些幻灯片介绍了大多数假设、形式主义和非零和博弈论背后的数学。它可能会帮助你从阅读开始游戏树搜索的幻灯片和具有隐藏信息的零和博弈论可从同一套教程中获得。在本教程中，我们将讨论多人非零和博弈的定义，策略的支配，纳什均衡。我们处理离散博弈，以及策略包含实数的游戏，比如你在两人双拍卖谈判中的出价。我们讨论了囚徒困境、公地悲剧、双重拍卖和多人拍卖，如第一次价格密封拍卖和第二次价格拍卖。双拍卖分析的数学可以在http://www.cs.cmu.edu/~awm/double_auction_math.pdf .

- 基于时间序列的异常检测算法简介。本简单教程概述了一些检测生物监测时间序列异常的方法。幻灯片不完整：演示文稿的口头评论尚未作为解释性文本框。请告诉我(AWM@cs.cmu.edu)，如果您对这些幻灯片和/或访问实现和绘制各种单变量方法的软件感兴趣。如果我收到足够的请求，我将尝试使上述两个可用。

- **AI**类介绍。对不同类型人工智能研究动机的快速非正式讨论

- 搜索算法什么是搜索算法？它是做什么工作的，它可以在哪里应用？我们介绍了宽度优先搜索和深度优先搜索的各种风格，然后讨论了各种备选方案和改进，包括迭代、深化和双向搜索。然后我们皱着眉头看了看一个叫做“最佳第一搜索”的想法。这将是我们第一次看到能够利用启发式函数的搜索算法。

- 一颗星的启发式搜索。经典的求最短路径的算法，给出了一种可接受的启发式算法。我们将讨论可否受理的概念(摘要：容许=乐观)。我们展示了如何证明A*的性质。我们还将简要讨论IDA*(迭代深化A*)。

- 约束满足算法，在计算机视觉和调度中的应用。本教程从人工智能文献中讲授关于约束满意度的概念。伴随的动画在http://www.cs.cmu.edu/~awm/animations/constraint。这是一种不知情搜索的特殊情况，在这种情况下，我们希望为满足一组约束的一组变量找到解决方案配置。例题包括图着色、8皇后、幻方、解释线条图的Waltz算法、多种调度，最重要的是扫雷车的演绎阶段。我们将研究的算法包括回溯搜索、前向检查搜索和约束传播搜索。我们还将研究通用启发式算法，以获得更多的搜索速度。

- 机器人运动规划我们回顾了一些聪明的路径规划算法，一旦我们到达实值连续空间，而不是安全和温暖的离散空间，到目前为止，我们一直隐藏。我们看配置空间，可见性图，单元分解，基于Voronoi的规划和势场方法.不幸的是，PDF版本中缺少了一些数字。

- 爬山，模拟退火和遗传算法。一些非常有用的算法，只在紧急情况下使用。