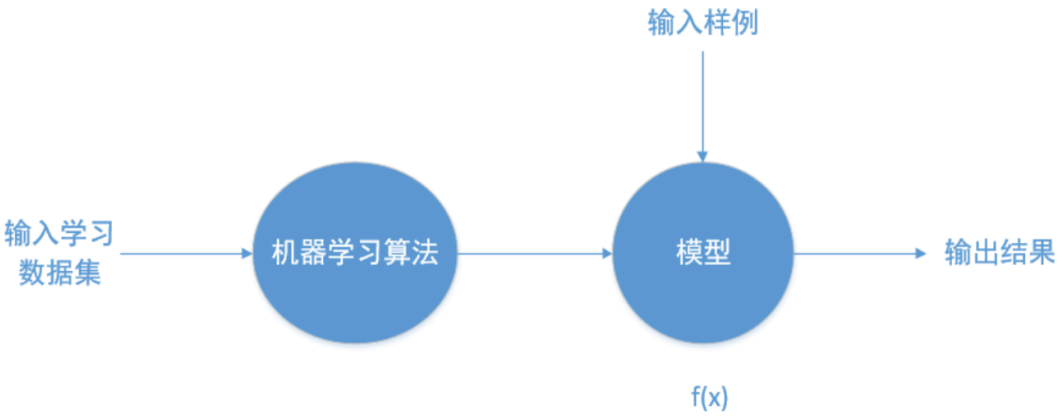
机器学习相关资料

2019 年 11 月 7 日 星期四

- **什么是机器学习？**

  机器学习的核心思想是创造一种普适的算法，类似于黑匣子，它能从数据中挖掘出有规律的东西，而不需要针对某个问题去写代码。你需要做的只是把数据“投喂”给这个算法，然后它会在数据上建立自己的逻辑。最基本的机器学习算法是解决分类和回归两大类问题。



机器学习

参考阅读：

机器学习系列（一）——机器学习简介

https://www.jianshu.com/p/b9583c7cb6c3

什么是机器学习？有哪些算法和分类？又有哪些应用？看完你就懂了

https://mp.weixin.qq.com/s/UfYI-HY2JEzyzJO3alruyg

机器学习

- **机器学习/数据分析/数据科学家人才技能**

  从某种程度来说，数据分析/数据挖掘/算法工程师/数据科学家的职责和内容，基本都可以概括为利用数据（大数据）来进行价值应用，包括数据探索、业务决策、算法建模、智能应用等等过程。

在这个行业，离不开两个核心的知识：数学和统计。

数学基础：

最新版《机器学习数学基础》发布，417 页 PDF 免费下载

https://mp.weixin.qq.com/s/R0jHIWHkE8YXDmhP6-TqkA

统计基础：

一文读懂统计学与机器学习的本质区别（附案例）

https://mp.weixin.qq.com/s/Mq6XEU8SheiOoEdIhNHaOg

值得收藏！数据分析最常用的 18 个概念，终于有人讲明白了

https://mp.weixin.qq.com/s/ZKCM6D4bHjgFZQax0p9P3A

数据分析的内容：大多数情况下，数据分析的过程必须包括数据探索的过程。数据探索可以有两个层面的理解：一是仅利用一些工具，对数据的特征进行查看；二是根据数据特征，感知数据价值，以决定是否需要对别的字段进行探索，或者决定如何加工这些字段以发挥数据分析的价值。字段的选取既需要技术手段的支撑，也需要数据分析者的经验和对解决问题的深入理解。

AI 时代的稀缺人才：解读数据科学家成长的 4 个阶段

https://mp.weixin.qq.com/s/QunD-bCmqyWy_DvJpLEqrA

- **机器学习的分类：**

机器学习粗的分类是有监督学习、无监督学习、增强学习，有监督学习需要标识数据，无监督学习不需要标识数据，增强学习介于两者之间（有部分标识数据）。

- **监督学习**：给机器的训练数据拥有标记或标签的学习方式是监督学习。监督学习主要处理分类和回归问题，主要的监督学习算法有：k 近邻 线性回归和多项式回归 逻辑回归 SVM 支持向量机 决策树和随机森林

- **无监督学习**：给机器的训练数据没有任何标记或标签答案。它经常对这些数据做聚类分析型分类和异常值检测。另外非监督学习可用于对数据进行降维，降维包括特征提取和特征压缩，经典的 PCA 算法就是非监督学习算法用于实现特征压缩，降维把高纬特征向量变为低纬，方便计算和可视化。
- **增强学习**：也叫强化学习，它根据周围环境的情况采取行动，根据每次行动的结果和反馈，学习和调整行动方式，它必须学习什么是最好的策略从而随着时间推移能获得最大回报。如 AlphaGo 内部的算法。现在无人驾驶，机器人等都是这种方式进行学习。监督学习和半监督学习依然是增强学习的基础。

简单来说，有监督和无监督的主要差别在于能够对预测结果用数据进行校验和判断，比如预测用户是否点击某商品，是可以用点击结果来对预测的结果进行打标的（预测对了就是对了，错了就是错了），但是假设我要对用户进行分群，如把用户进行下沉用户和非下沉用户的分类，我没有数据对预测进过进行打标判断的（预测结果我也不知道对不对，或者没有绝对标准说对不对），那就是无监督学习，因为无监督学习不能对算法结果进行评判，模型好坏就很难得知，所以大多数算法工程师都会绞尽脑汁把无监督问题转化为有监督问题，或者采用增强学习的方案，对部分数据进行打标，然后持续迭代，来解决不知道如何制定‘KPI’的问题。--by liulinghan

- **机器学习十大经典算法：**

有监督学习算法：决策树、朴素贝叶斯分类器、最小二乘法、逻辑回归、支持向量机、集成学习、

无监督学习算法：聚类算法、主成分分析 PCA、SVD 矩阵分解、独立成分分析 ICA

参考阅读：

关于机器学习的知识点，全在这篇文章里了

https://mp.weixin.qq.com/s/tPdHzinzcDhJDRvIx8dSNw

图解十大经典机器学习算法入门

https://blog.csdn.net/jrunw/article/details/79205322

机器学习十大算法都是何方神圣？看完你就懂了：

http://tech.sina.com.cn/it/2016-12-24/doc-ifxyxury8364458.shtml

梯度下降的原理小结：

https://www.cnblogs.com/pinard/p/5970503.html

GBDT 原理：

https://www.cnblogs.com/pinard/p/6140514.html

- **机器学习在电商中的应用：**

通常机器学习在电商领域有三大应用，推荐、搜索、广告。在电商推荐中，点击率预估算法主要使用的机器学习是逻辑回归（LR）和决策树（常用 GBDT，XGboost，LightGBM 算法）。当然目前大多数 CTR 预估模型都使用深度学习了。

参考阅读：

机器学习在电商应用中的三个境界：爆款模型、转化率模型及个性化模型

https://www.jianshu.com/p/6bbfe3ff18ed

- **机器学习系统的核心要素：**

模型算法、日志流、训练系统、特征系统、评估系统

特征系统：主要解决线上和线下特征一致性的问题

点击率预估：

https://blog.csdn.net/wuxiaosi808/article/details/77985656

其他扩展内容：

京东：包勇军_京东电商广告和推荐系统的机器学习系统实践

演讲内容：http://bigdata.it168.com/a2016/0913/2915/000002915998.shtml

ppt 内容：https://wenku.baidu.com/view/7081273d80eb6294dc886c84.html

备注：广告侧 16 年的线上推荐和离线训练系统，深度学习尝试就已经比我们现在还要领先很多了。虽然他们当时线上算法主要还是 LR，但是后面应该很快就 GBDT 和深度学习部署了。而且 3 年前已经尝

试了 10 亿特征，要知道我们现在目前最多才使用 1000 个特征（有用的特征只有 500 个）。-by liulinghan

KAGGLE 点击率预估案例：

利用 Python 使用 LR 模型训练一个简单 CTR 预估模型（含代码教程）：

https://blog.csdn.net/zhouwenyuan1015/article/details/72474026

利用 SPARK 使用 LR 模型训练一个简单 CTR 预估模型（含代码教程）：

http://www.sohu.com/a/121498635_500653

特征工程相关：

特征预处理：

https://www.cnblogs.com/pinard/p/9093890.html

特征选择：

https://www.cnblogs.com/pinard/p/9032759.html

特征表达：

https://www.cnblogs.com/pinard/p/9061549.html

- **什么是深度学习？**

对许多机器学习问题来说，特征提取不是一件简单的事情。在一些复杂问题上，要通过人工的方式设计有效的特征集合需要很多的时间和精力，有时甚至需要整个领域数十年的研究投入。例如，假设想从很多照片中识别汽车。现在已知的是汽车有轮子，所以希望在图片中抽取“图片中是否出现了轮子”这个特征。但实际上，要从图片的像素中描述一个轮子的模式是非常难的。虽然车轮的形状很简单，但在实际图片中，车轮上可能会有来自车身的阴影、金属车轴的反光，周围物品也可能会部分遮挡车轮。实际图片中各种不确定的因素让我们很难直接抽取这样的特征。

深度学习解决的核心问题之一就是自动地将简单的特征组合成更加复杂的特征，并使用这些组合特征解决问题。深度学习是机器学习的一个分支，它除了可以学习特征和任务之间的关联以外，还能自动从简单特征中提取更加复杂的特征。图 1 中展示了深度学习和传统机器学习在流程上的差异。如图 1 所

示，深度学习算法可以从数据中学习更加复杂的特征表达，使得最后一步权重学习变得更加简单且有效。在图 2 中，展示了通过深度学习解决图像分类问题的具体样例。深度学习可以一层一层地将简单特征逐步转化成更加复杂的特征，从而使得不同类别的图像更加可分。比如图 2 中展示了深度学习算法可以从图像的像素特征中逐渐组合出线条、边、角、简单形状、复杂形状等更加有效的复杂特征

「回顾」TensorFlow 技术发展与落地实践

https://mp.weixin.qq.com/s/_p77pDl2iAN2t7RAeesPFg

深入浅出 Tensorflow（一）：深度学习及 TensorFlow 简介

https://www.jianshu.com/p/5905da318b48

手把手教你训练一个神经网络，打爆 21 点！

https://mp.weixin.qq.com/s/6AXqJL1jVw_XpXq2AL2kow

深度学习与神经网络

https://www.jianshu.com/p/cc15bc0a75f1

什么是 embedding？

http://blog.csdn.net/weixin_44493841/article/details/95341407

Embedding 从入门到专家必读的十篇论文

https://cloud.tencent.com/developer/article/1457042

理解 BERT:一个突破性 NLP 框架的综合指南 https://www.jianshu.com/p/271e34e0daa7