



小组项目题目列表 (Topic List)

Topic 01 (机器学习) 逻辑回归

Description 逻辑回归 (Logistic Regression) 是一个简单高效的有监督分类算法, 广泛应用于机器学习、生物医学研究、经济预测等领域。本项目希望同学们介绍逻辑回归的基本原理和优缺点, 对比线性回归, 并通过生活中的实际例子来了解逻辑回归的应用。

Topic 02 (机器学习) 朴素贝叶斯分类器

Description 朴素贝叶斯分类器 (Naive Bayes Classifier) 是以贝叶斯定理为基础的简单分类器, 在数据挖掘、模式识别、及生命科学研究中应用相对广泛。本项目希望同学们介绍朴素贝叶斯分类器的基本原理, 了解其优势和局限性, 通过具体示例演示其在实践中的应用。

Topic 03 (机器学习) 反向传播算法

Description 在课堂上我们学习了神经网络的基本架构, 了解到在神经网络的训练过程中, 权重的更新是通过目标函数的误差反向传播 (Error Backpropagation) 来进行的。本项目希望同学们介绍全连接神经网络误差反传的具体过程, 了解反向传播的目的, 以及为什么要从后向前进行误差反传, 可能会存在什么问题。

Topic 04 (机器学习) 非线性激活函数

Description 课程上我们介绍非线性激活函数 (Activation Function) 为神经网络引入了非线性性, 提高了神经网络拟合函数的复杂度。本项目希望同学们介绍不同类型的非线性激活函数, 例如, Sigmoid、Tanh、ReLU、LReLU、Softmax 等, 深入了解它们的工作原理和优缺点, 讨论在实际应用中如何选择非线性激活函数。

Topic 05 (计算机视觉) VGG

Description 在课程上我们介绍了神经网络的基本构造, 但在实际使用中, 如何选择卷积核的大小和卷积层的层数? 本项目希望同学们学习和介绍一个经典的神经网络模型 VGG, 通过阅读以下论文, 了解 VGG 网络的构架和设计原理, 讨论其在实际应用中的使用效果。

Reference: Simonyan, Karen, and Andrew Zisserman, Very Deep Convolutional Networks for Large-Scale Image Recognition, arXiv preprint arXiv:1409.1556, 2014.

Topic 06 (计算机视觉) ResNet

Description 随着神经网络层数的增加, 神经网络的训练难度增加, 会出现梯度消失或爆炸的问题, 特别是梯度消失问题, 通常不容易检测和避免。残差网络 ResNet 的提出使得超深神经网络的构建和优化变成可能。本项目希望同学们学习和介绍另一个经典的神经网络模型 ResNet, 通过阅读以下论文, 了解 ResNet 网络的构架和设计原理, 讨论为什么 ResNet 可以构建超过一百层的神经网络, 以及其在实际应用中的表现。

Reference: He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, Deep Residual Learning for Image Recognition, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778. 2016.

Topic 07 (计算机视觉) YOLO

Description 在课堂上我们提及目标检测是神经网络在计算机视觉中的典型应用。本项目希望同学们学习和介绍一个经典的实时目标检测网络 YOLO (You Only Look Once), 通过阅读以下论文, 了解 YOLO 网络的构架和设计原理, 讨论为什么 YOLO 可以达到实时检测的目的。

Reference: Redmon, Joseph, Santosh Divvala, Ross Girshick, Ali Farhadi, You Only Look Once: Unified, Real-Time Object Detection, In IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779-788, 2016.



Topic 08 (计算机视觉) OpenPose

Description OpenPose 是首个基于深度学习的实时多人二维姿态估计的算法，可实现人体动作、面部表情、手指运动等姿态估计。本项目希望同学们通过阅读以下论文和搜集相关材料，介绍 OpenPose 的基本原理和应用实例，讨论其优缺点。

Reference: Cao, Zhe, Gines Hidalgo, Tomas Simon, Shih-En Wei, Yaser Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields, IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 43, no. 1, pp. 172–186, Jan. 2021.

Topic 09 (数据库) Oracle

Description Oracle 数据库是甲骨文公司的一款关系数据库管理系统，也是世界上使用最为广泛的通用数据库系统。本项目希望同学们介绍 Oracle 数据库的系统构架、数据特点，分析为什么 Oracle 公司能在世界数据库市场长期占据首位，并进一步了解 Oracle 数据库的特点和应用范围。

Topic 10 (数据库) OceanBase

Description OceanBase 是阿里巴巴集团研发的分布式数据库软件，用于淘宝网和诸多阿里集团的云服务、部分政府结构和银行等。本项目希望同学们介绍 OceanBase 数据库的系统构架、数据特点，并与国际著名数据库公司做比较，分析其优势和不足。

Topic 11 (数据挖掘) Apriori

Description 先验算法 (Apriori Algorithm) 是数据挖掘中关联规则学习的经典算法之一，课堂上我们进行了简单的介绍。本项目希望同学们深入了解 Apriori 算法的核心思想，介绍它是如何挖掘到 k 项频繁项集的，并通过实例来说明算法的执行过程。

Topic 12 (数据挖掘) SVM

Description 支持向量机 (Support Vector Machine, SVM) 是机器学习和数据挖掘中经典的有监督分类算法。本项目希望同学们介绍 SVM 的基本原理，介绍其是如何工作的，能否处理非线性可分的分类问题，以及 SVM 的实际应用例子。

Topic 13 (数据挖掘) K-Means 变体

Description 在课堂中，我们介绍了 K-Means 聚类算法，它是一种无监督聚类算法，因其简单和不错的聚类效果，有着广泛的应用。K-Means 算法有大量的变体，例如，K-Means++，K-Medoids，K-Medoids++，Mini Batch K-Means 等。本项目希望同学们介绍 K-Means 的两到三个变体，分析它们的异同和优缺点，以及对比其在什么场景下最为适用。

Topic 14 (数据挖掘) PCA

Description 主成分分析 (Principal Component Analysis, PCA) 是用于数据降维的经典算法，通过提取数据的主要特征分量，将高维数据转换到低维，广泛应用于数据压缩，消除冗余或数据噪声等。本项目希望同学们介绍 PCA 的主要思想以及工作步骤，以一个具体案例展示 PCA 的运算过程，并分析其应用和不足。

Topic 15 (数据挖掘) 决策树熵划分准则

Description 在课堂中，我们介绍了决策树在构建过程中，需要根据特征进行划分，例如采用 Gini 系数来划分，使得每个节点的纯度越来越高。在实际应用中，熵划分准则是一个常用的技术，它使用信息熵 (entropy) 来衡量每个节点的不确定性。本项目希望同学们介绍“熵”的概念，分析熵划分准则是怎样应用到决策树划分的，并对比其与基于 Gini 系数划分有什么优缺点。



Topic 16（数据挖掘）随机森林

Description 深度学习之前，随机森林（Random Forest）是常用的机器学习经典算法，是一种可用于分类（Classification）和回归（Regression）等任务的集成学习方法。本项目希望同学们了解随机森林算法的工作原理，通过实例阐明随机森林是如何应用于分类或回归问题的，以及其有何优缺点。

Topic 17（推荐系统）协同过滤

Description 协同过滤（Collaborative Filtering）是利用某兴趣相投或拥有共同经验的群体的喜好来推荐用户感兴趣的信息，个人通过合作的机制给予信息一定程度的回应（如评分）并记录下来以达到过滤的目的，进而帮助别人筛选信息。本项目希望同学们了解一种协同过滤算法，介绍其工作原理，并通过实例进行说明。

Topic 18（推荐系统）推荐链路

Description 推荐系统的推荐链路包括召回、粗排、精排、重排，现有的大多数应用软件正是通过这四个步骤从千万甚至更高量级的视频/广告/产品中选出最合适的一个或多个推荐至你的电脑/手机终端。本项目希望同学们学习推荐链路的四个阶段，介绍它们是如何工作的，以及每个阶段的作用和意义。（也可以介绍不同IT公司的不同推荐系统链路研发前沿，比如腾讯目前正在尝试将召回和粗排合并，以一套模型输出结果）

Reference: <https://www.bilibili.com/video/BV1hF411M7b5/>

Topic 19（推荐系统）PageRank

Description 网页排名（PageRank）是 Google 使用的对网页搜索结果进行排名的算法，决定了网页在搜索结果中的相对重要性。本项目希望同学们了解 PageRank 的核心思想和算法流程，讨论 PageRank 是如何应用到网页排序中的，并思考这一方法本质的科学原理。

——本文档由洪义撰写，高晓飒终校

胡永祎、凯丽布努尔·吐孙提出修订意见