Awesome-Road-Map

author:isLinXu
Github:https://github.com/isLinXu/awesome-road-map

1、基础知识(Fundamentals)

- 1.1 基础(Basic)
  - 1.1.1 矩阵和线性代数基础(Matrices & Linear Algebra Fundamentals)
  - 1.1.2 数据库基础(Database Basics)
  - 1.1.3 表格数据(Tabular Data)
  - 1.1.4 Pandas数据结构(Dataframes & Series)
  - 1.1.5 数据仓库技术(ETL,Extract,Transform,Load)
  - 1.1.6 业务分析与商业分析(Reporting vs BI vs Analytics)
  - 1.1.7 数据格式(Data Formats)
    - (1) JSON
    - (2) XML
    - (3) CSV
    - (4) TXT
  - 1.1.8 正则表达式(RegEx)
- 1.2 Python编程(Python Programming)
  - 1.2.1 Python基础(Basics)
    - (1) 表达式(Expressions)
    - (2) 变量(Variables)
    - (3) 数据结构(Data Structures)
    - (4) 函数(Functions)
    - (5) 安装包(via pip,conda or similar)
    - (6) 代码风格(CodeStyles,e.g.PEP8)
  - 1.2.2 重要库(Important Libraries)
    - (1) NumPy
    - (2) Pandas
  - 1.2.3 虚拟环境(Virtual Environments)
    - (1) miniconda
    - (2) anaconda
  - 1.2.4 Jupyter Notebook/Lab
- 1.3 数据源(Data Sources)
  - 1.3.1 数据挖掘(Data mining)
  - 1.3.2 网页抓取(Web Scraping)
  - 1.3.3 开源数据集(Pubilc Datasets)
  - 1.3.4 数据竞赛平台(e.g Kaggle)
- 1.4 数据处理与分析(Exploratory Data Analysis)
  - 1.4.1 主成分分析(PCA,Principal Component Analysis)
  - 1.4.2 降维及数值约简(Dimensionality & Numerosity Reduction)
  - 1.4.3 归一化(Normailzation)
  - 1.4.4 数据清理，处理缺失值(Data Scrubbing, Handling Miss Value)
  - 1.4.5 无偏估计(Unbiased Estimators)
  - 1.4.6 稀疏值分箱(Binning Sparse Value)
  - 1.4.7 特征提取(Feature Extraction)
  - 1.4.8 降噪(Denoising)
  - 1.4.9 采样(Sampling)

2、数据科学(Data Science)

- 2.1 数据统计(Statistics)
  - 2.1.1 概率论(Probability Theory)
    - (1) 随机性、随机变量和随机样本(Randomness,Random Variable and Random Sample)
    - (2) 概率分布(Probability Distribution)
    - (3) 条件概率和贝叶斯定理(Conditional Probability and Bayes's Theorem)
    - (4) 统计独立性(Statistical Independence)
    - (5) 独立同分布(IID,Independent and identically distributed)
    - (6) cdf,pdf,pmf
      - 1.累积分布函数(cdf,Cumulative Distribution Function)
      - 2.概率密度函数(pdf,Probability Density Function)
      - 3.概率质量函数(pmf,Probability Mass Function)
  - 2.1.2 连续分布(Continuous Distributions)
    - (1) 正态分布及高斯分布(Normal/Gaussian)
    - (2) 一致连续性(Uniform Continuous)
    - (3) Beta 分布(Beta Distributions)
    - 4.狄利克雷分布(Dirichlet Distribution)
    - (5)指数型分布(Exponential Distribution)
    - (6) 卡方(chi-square)
  - 2.1.3 离散分布(Discrete Distributions)
    - (1) 均匀离散分布(uniform discrete)
    - (2) 二项式分布(Binomial Distribution)
    - (3) 多项式分布(Multinomial Distribution)
    - (4) 超几何分布(Hypergeometric Distribution)
    - (5) 泊松分布(Poisson Distribution)
    - (6) 几何分布(Geometric Distribution)
  - 2.1.4 汇总统计(Summary Statistics)
    - (1) 期望和均值(Expectation and mean)
    - (2) 方差和标准差(Variance and Standard)
    - (3) 协方差与相关性(Covariance and Correlation)
    - (4) 中位数与四分位数(Median,Quartile)
    - (5) 四分位数范围(Interquartile Range)
    - (6) 百分位数/分位数(Percentile/Quantile)
    - (7) 众数(Mode (statistics))
  - 2.1.5 重要规则(Important Laws)
    - (1) 大数法则(Law of Large Numbers)
    - (2) 中心极限定理(Central Limit Theorem)
  - 2.1.6 估计(Estimation)
    - (1) 极大似然估计(Maximum Likelihood Estimation)
    - (2) 核密度估计(Kernel Density Estimation)
  - 2.1.7 假设检验(Hypothesis Testing)
    - (1) P值 (p-value)
    - (2) 卡方检验(chi-square-test)
    - (3) F检验(F-test)
    - (4) t 检验(t-test)
  - 2.1.8 置信区间(Confidence Interval)
  - 2.1.9 蒙特卡罗法(Monte Carlo Method)
- 2.2 可视化(Visualization)
  - 2.2.1 图表建议(Chart Suggestions)
  - 2.2.2 Python库
    - (1) Matplotlib
    - (2) Plotnine(like ggplot in R)
    - (3) Bokeh
    - (4) Seaborn
    - (5) ipyvolume(3D data)
  - 2.2.3 Web技术
    - (1) Vega-Lite
    - (2) D3.js
  - 2.2.4 仪表盘(Dashboards)
    - Dash
  - 2.2.5 BI
    - (1) Tableau
    - (2) PowerBI

3、机器学习(Machine Learning)

- 3.1 综述概览(General)
  - (1) 概念、输入和属性(Concepts,Inputs & Attributes)
    - 1.分类变量(Categorical Variables)
    - 2.有序变量(Ordinal Variable)
    - 3.数值变量(Numerical Variable)
  - (2) 代价函数和梯度下降(Cost functions and Gradient Descent)
  - (3) 过拟合/欠拟合(Overfitting/underfitting)
  - (4) 训练、验证和测试数据(Training,Validation and test data)
  - (5) 精度和召回(Precision vs Recall)
  - (6) 偏差和方差(Bias & Variance)
  - (7) 优化提升(Model Lift)
- 3.2 方法(Method)
  - (1) 监督学习(Supervised Learning)
    - 1.回归(Regression)
      - a.线性回归(Linear Regression)
      - b.泊松回归(Poisson Regression)
    - 2.分类(Classification)
      - a.正确分类率(Classification Rate)
      - b.决策树(Decision Trees)
      - c.逻辑回归(Logistic Regression)
      - d.朴素贝叶斯分类器(Naive Bayes Classifiers)
      - e.K-最近邻(K-Nearest Neighbour)
      - f.支持向量机(Support Vector Machine)
  - (2) 无监督学习(Unsupervised Learning)
    - 1.关联规则学习(Association Rule Learning)
      - a.先验算法(Apriori Algorithm)
      - b.ECLAT算法(ECLAT Algorithm)
      - c.频繁模式树(FP-Tree,Frequent-Pattern tree)
    - 2.降维(Dimensionality Reduction)
      - a.主成分分析(PCA,Principal component analysis)
      - b.随机投影(Random Projection)
      - c.非负矩阵分解(NMF,Non-negative matrix factorization)
      - d.t-分布式随机邻域嵌入(T-SNE,t-Distributed Stochastic Neighbor Embedding)
      - e.一致流形近似和投影(UMAP,Uniform Manifold Approximation and Projection)
  - (3) 集成学习(Ensemble Learning)
    - 1.提升方法(Boosting)
    - 2.装袋算法(Bagging,Bootstrap aggregating)
    - 3.堆叠(Stacking)
  - (4) 强化学习(Reinforcement Learning)
    - Q-Learning
- 3.3 用例(Use Cases)
  - (1) 情绪分析(Sentiment Analysis)
  - (2) 协同过滤(Collaborative Filtering)
  - (3)标注(tagging)
  - (4)预测(Prediction)
- 3.4 工具(Tools)
  - (1)重要库(mportant Libraries)
    - 1.Scikit-Learn
    - 2.Spacy(NLP)

4、深度学习(Deep Learning)

- 4.1 论文(Papers)
  - (1) Deep Learning Papers Reading Roadmap
  - (2) Paper with code
  - (3) Paper with code-state of the art
- 4.2神经网络(Neural Networks)
  - (1) 理解神经网络(Understanding Neural Networks)
  - (2) 损失函数(Loss Functions)
  - (3) 激活函数(Activation Functions)
  - (4) 权重初始化(Weight Initialization)
  - (5) 梯度消失/爆炸问题(Vanishing/Exploding Gradient Problem)
- 4.3 结构(Architectures)
  - (1) 前馈神经网络(Feedforward Neural Network)
  - (2) 自编码器(Autoencoder)
  - (3) 卷积神经网络(CNN,Convolutional Neural Network)
  - (4) 循环神经网络(Recurrent Neural Network)
  - (5) Transformer
  - (6) 孪生神经网络(Siamese Network)
  - (7) 对抗生成网络(GAN,Generative Adversarial Network)
  - (8) 进化发展结构(Evolving Architectures /NEAT)
  - (9) 残差连接(Residual Connections)
- 4.4 训练(Training)
  - (1) 优化器(Optimizers)
  - (2) 学习率调度器(Learning Rate Schedule)
  - (3) 批量归一化(Batch Normalization)
  - (4) 批量大小影响(Batch Size Effects)
  - (5) 正则化(Regularization)
  - (6) 多任务学习(Multitask Learning)
  - (7) 迁移学习(Transfer Learning)
  - (8) 课程学习(Curriculum Learning)
- 4.5 工具框架(Tools&Frame)
  - a.Awesome Deep Learning
  - b.Huggingface Transformers
  - (1) Important Libraries
  - (2) TensorFlow
  - (3) Pytorch
  - (4) TensorBoard
  - (5) MLFlow
- 4.6 高级模型优化(Model optimization advanced)
  - (1) 蒸馏(Distillation)
  - (2) 量化(Quantization)
  - (3) 神经架构搜索(NAS,Neural Architecture Search)

5、数据工程(Data Engineering)

- 5.1 数据格式总结(Summary of Data Formats)
- 5.2 数据发现(Data Discovery)
- 5.3 数据来源&采集(Data Source & Acquisition)
- 5.4 数据集成(Data Integration)
- 5.5 数据融合(Data Fusion)
- 5.6 数据转换与数据增强(Transformation & Enrichment)
- 5.7 数据调查(Data Survey)
- 5.8 数据整理(OpenRefine)
- 5.9 数据量化(How much Data)
- 5.10 实施有效的提取、转换、加载流程(using ETL，Extract, Transform and Load)
- 5.11 数据湖与数据仓库(Data Lake vs Data Warehouse)
- 5.12 将您的 Python 应用程序 Docker 化(Dockerize your Python Application)

6、大数据(Blg data)

- 6.1 大数据架构(Blg Data Architectures)
  - (1) 架构模式和最佳实践(Architectural Patterns & Best Practices)
- 6.2 原则(Principles)
  - (1) 水平竖直缩放(Horizontal vs Vertical Scaling)
  - (2) Apache Hadoop- MapReduce
  - (3) 数据复制(Data Replication)
  - (4) 名称节点与数据节点(Hadoop Name & Data Noddes)
  - (5) 工作和任务跟踪器(Job & Task Tracker)
- 6.3 工具(Tools)
  - (1) Check the Awesome Big Data List
  - (2) Hadoop(Large Data)
  - (3) Spark(in memory)
  - (4) RAPIDS(On GPU)
  - (5) Flume,Scribe：For Unstruct Data
  - (6) Data Warehouse with Hive
  - (7) Elastic(ELK)Stack
  - (8) Avro
  - (9) Flink
  - (10) Numba
  - (11) Onnx
  - (12) OpenVINO
  - (13) MLFlow
  - (14) Kafka & KSQL
  - (15) Databases
    - 1.Cassandra
    - 2.MongoDB,Neo4j
    - 3.AWS SageMaker
    - 4.Google ML Engine
    - 5.Microsoft Azure Machine Learning Studio
  - (16) Scalability
    - 1.ZooKeeper
    - 2.Kubernetes
  - (17) Cloud Services
  - (18) Awesome Production ML