基本原理

- 机器如何学习
  - 让机器学习规律与事物间联系　算法对数据进行运算，产生模型
  - 关于数据　计算机处理的是数值和运算，要对源数据转换（向量）
  - 一些概念　数据样本 特征 标签 标注
  - 有监督模型与无监督模型
    - 有监督：通过标注数据进行学习
    - 无监督：没有标签 只有样本
- 模型的获取和改进
  - 获取模型的过程
    - Step-1：数据准备
      - Step-1.1 数据预处理　收集数据、清洗数据、标注数据。
      - Step-1.2 构建数据的向量空间模型　将文本、图片、音频、视频等格式的数据转换为向量
      - Step-1.3 分集　将构建好向量空间模型的数据分为训练集、验证集和测试集　每个集合独立，验证集和测试集可以合二为一，但训练集总是占大头
    - Step-2：训练
      - Step
        - Step-2.1：编写训练程序
          - Step-2.1.1：选择模型类型。
          - Step-2.1.2：选择优化算法
          - Step-2.1.3：根据模型类型和算法编写程序。
        - Step-2.2：训练 -> 获得临时模型
        - Step-2.3：在训练集上进行临时模型，获得训练集预测结果
        - Step-2.4：在验证集上进行临时模型，获得验证集预测结果
        - Step-2.5：综合参照 Step-2.3 和 Step-2.4 的预测结果，改进模型
        - Step-2.6：Step-2.2 到 Step-2.5 反复迭代，直至获得让我们满意，或者已经无法继续优化的模型
      - 定义：将训练集输入给训练程序，进行运算。
      - tips：核心是算法，训练程序输出的结果，就是模型
    - Step-3：测试　将测试集数据输入给训练获得的模型，得到预测结果；再将预测结果与这些数据原本预期的结果进行比较。
  - 改进模型　Step 2.5——从三个方面来进行：数据、算法和模型。
    - 数据
      - 有限的数据
        - 大量的高质量训练数据，是提高模型质量的最有效手段
        - 对数据进行归一化（Normalization，又译作正规化、标准化）等操作。
        - 采用 Bootstrap 等采样方法处理有限的训练/测试数据，以达到更好的运算效果。
        - 根据业务进行特征选取：从业务角度区分输入数据包含的特征，并理解这些特征对结果的类型。
    - 调参（算法）
      - 调参过程
        - 训练过程的输出的参数并不需要我们来调。超参数是需要模型训练者自己来设置和调整的
        - 1.制定目标。
        - 2.制定策略
        - 3.执行
        - 4.验证
        - 5.调整策略 -> 3
        - 6.迭代3-5
      - 组合调参较难，深入理解原理后才有可能找到正确的调参方向
    - 模型类型选择
      - 对某个模型再怎么调参，都无法满足在某个特定指标上的要求时，考虑换模型
      - tips
        - DL 模型（CNN、DNN、RNN、LSTM 等等）对训练数据的量需求较大，训练时间普遍较长，可解释性较弱，一旦出现问题，很难 Debug
        - 先好好学统计学习模型。
- 机器学习三要素
  - 数据
    - 源数据（文字、图片、音频、视频）
    - 数据处理步骤
      - 1.特征工程
        - a.确定用哪些特征来表示数据
        - b.确定用什么方式表达这些特征
      - 2.VSM转换　构建一个向量空间模型（VSM）。VSM 负责将格式（文字、图片、音频、视频）转化为一个个向量。
    - 一些概念
      - 无标注数据　这些特征来做聚类，转化成的向量称为这份数据的特征向量
      - 数据标注　给训练样本打标签（用变量 y 表示，离散值：0、1）
  - 模型
    - 定义　模型是机器学习的结果，学习过程称为训练（Train）
    - 输出结果　模型视为函数，输入数据，得到结果，结果可能是一个数值（回归），也可能是一个标签（分类）
    - 训练过程
      - 情景：选定模型类型，不知道其中参数。
      - 结合训练数据计算出各个参数的具体取值的过程
  - 算法
    - 定义　训练过程依据某种做法进行运算，该做法就是算法。
    - 目标　有监督学习的目标：让训练数据的所有 x 经过 f(x) 计算后，获得的 y' 与它们原本对应的 y 的差别尽量小。
    - 一些概念
      - 损失函数（Loss Function）　L（y, y'）= L(y, f(x))，描述 y' 与 y 之间的差别。
      - 代价函数（Cost Function）　J（theta），描述整体的损失
        - theta 表示 f(x) 中所有待定的参数（theta 也可以是一个向量，每个维度表示一个具体的参数）
      - 目标函数　最小化J（theta）
    - 常用优化算法　梯度下降法（Gradient Descent），共轭梯度法（Conjugate Gradient），牛顿法和拟牛顿法、模拟退火法（Simulated Annealing）
- 模型的质量和评判指标
  - 概念
    - 精准率（Precision）
    - 召回率（Recall）
    - 综合二者的F1Score
  - 关于指标的tips
    - P 和 R 是矛盾的 很难保证双高
    - P、R、F1Score 在分类问题中都是对某一个类而言的，每一个类都有一套独立的值，衡量模型整体质量，要综合看所有指标。
    - 指标还和测试数据有关
      - 同样的模型，换一套测试数据后，很可能 P、R、F1Score 会有变化
      - 考虑是否存在 bias 或者 overfitting 的情况
    - 任意的评价指标，都同时指向一个模型和一个数据集，两者缺一不可。
    - 欠拟合（Underfitting）和过拟合（Overfitting）
- 最常用的优化算法——梯度下降法
  - 前提：确认目标函数是凸函数
  - 梯度下降法
    - 0.首先需要确定步进长度α
    - 1.任意找一点x0
    - 2.计算f(x0)及f'(x0)
    - 3.x=x0-f'(x0)*α
    - 4.将上述所得的x作为新的x0
    - 5.重复2-4（即迭代过程）
  - 超参数　步长
  - 推导　https://blog.csdn.net/pengchengliu/article/details/80932232

试用版 XMind:ZEN 试用版