法律声明

- 本课件包括：演示文稿，示例，代码，题库，视频和声音等，小象学院拥有完全知识产权的权利；只限于善意学习者在本课程使用，不得在课程范围外向任何第三方散播。任何其他人或机构不得盗版、复制、仿造其中的创意，我们将保留一切通过法律手段追究违反者的权利。

关注 小象学院

XGBoost

定义

- eXtreme Gradient Boosting： A scalable machine learning system for tree boosting
- 论文链接 http://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf
- 项目开源实现 https://github.com/dmlc/xgboost
- 不仅仅是个算法，是个具有并行化数据处理能力、可处理大规模数据的系统

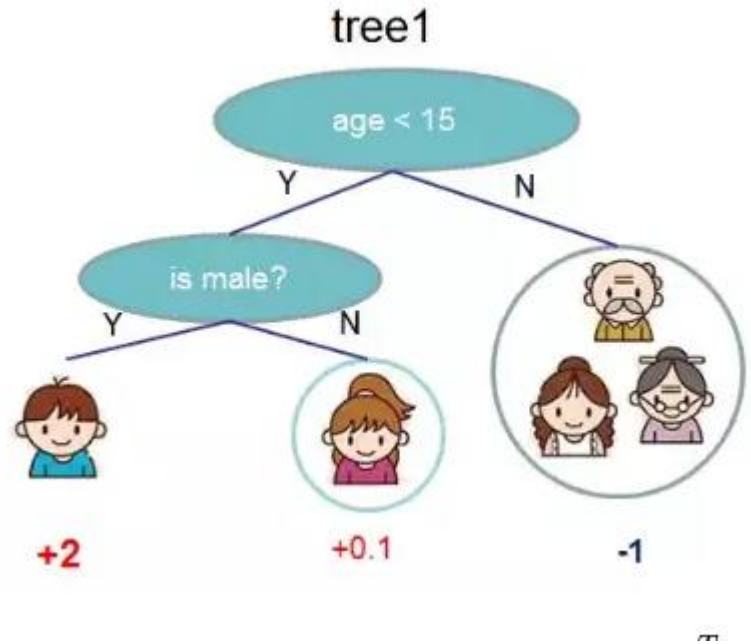
XGBoost与GBDT

- 基本思想与GBDT相同
  - Boosting tree
  - 下一棵树学习上一课数的残差
- 不同点
  - 算法
    - 误差函数中引入正则项（更高的泛化能力）

$$L(\phi) = \sum_i l(\hat{y}_i - y_i) + \sum_k \Omega(f_k)$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

XGBoost与GBDT（Cont.）



$$\Omega = \gamma 3 + \frac{1}{2}\lambda \sum_{j=1}^{T} (4 + 0.01 + 1)$$

XGBoost与GBDT（Cont.）

- 不同点
  - 算法
    - 分裂节点，使节点分裂之后树所得分数提高最多

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma$$

XGBoost与GBDT（Cont.）

- 不同点
  - 系统（计算性能更好，支持分布式）
    - 并行化
      - 将各列分块存储，在进行各列的分裂点计算查找时，可以并行进行
    - 分裂点选择
      - 支持枚举，也支持分位点算法
    - 运用计算机工程领域的缓存技术，增加数据的连续访问命中率
    - 可以在Hadoop、Spark、MPI等并行计算框架上执行

使用

- 安装

  conda install -c anaconda py-xgboost

- API

https://xgboost.readthedocs.io/en/latest/python/index.html

  - Scikit-Learn API
  - Learning API

联系我们

小象学院：互联网新技术在线教育领航者

– 微信公众号：**小象学院**