深度学习与中文自然语言处理

Deep Learning for Natural Language Processing

April-15-2018

- 1. 贝叶斯分类器；
    - 贝叶斯概率模型；
    - 贝叶斯分类器的实现；
    - 机器学习分类器的注意事项—baseline;
- 2. 机器学习
    - 什么是机器学习与如何衡量；
    - 机器学习的几种评测指标；
    - precision, recall, accuracy, AUC
    - overfitting
    - underfitting

- 3. 决策树与随机森林；
    - 利用决策树来解决泰坦尼克号乘生存问题；
    - 决策树与最大熵；
    - 决策树的其他应用； 特征的重要性排序；
    - 随机森林；

Question 1

- Q1: 桌子上有10张牌， 每张标记1, .., 10, 取出一张牌， 大于5， 问， 此张牌是8个概率是多少?

条件概率

贝叶斯公式-1-使用的背景

贝叶斯公式-2-原理

Q2: prior, likehood, evidence

- Q: 检测的正确率有99%, 1%的健康人会被检测出来为“假阳性”， 总体人群中， 千分之2的人有此种病. 某人检测出来为“阳性”. 此人生病的概率是?
  - A: 大约95%
  - B: 大约75%
  - C: 大约55%
  - D: 大约15%

贝叶斯公式-3-贝叶斯分类器

贝叶斯公式-3-朴素贝叶斯分类器

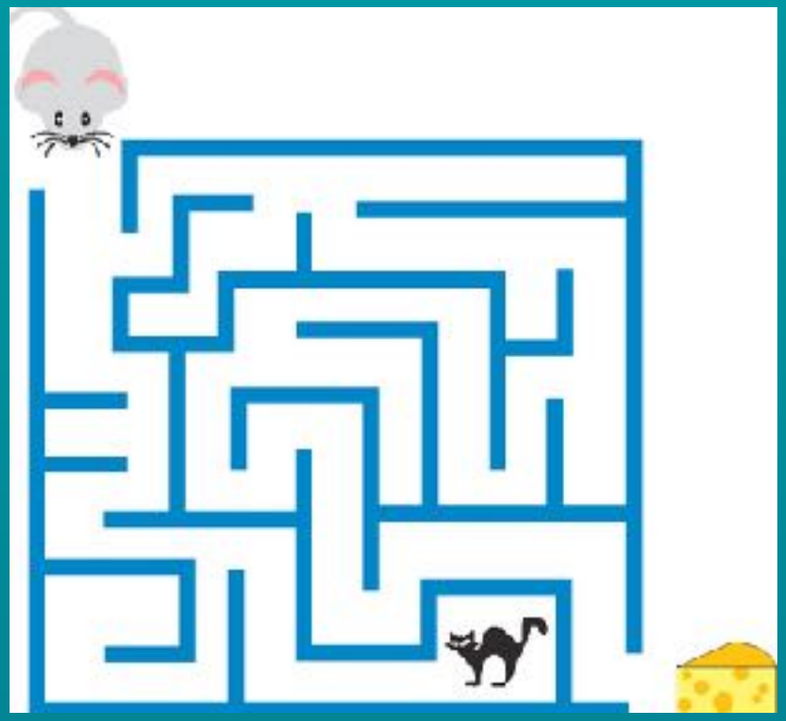
高斯贝叶斯分类器

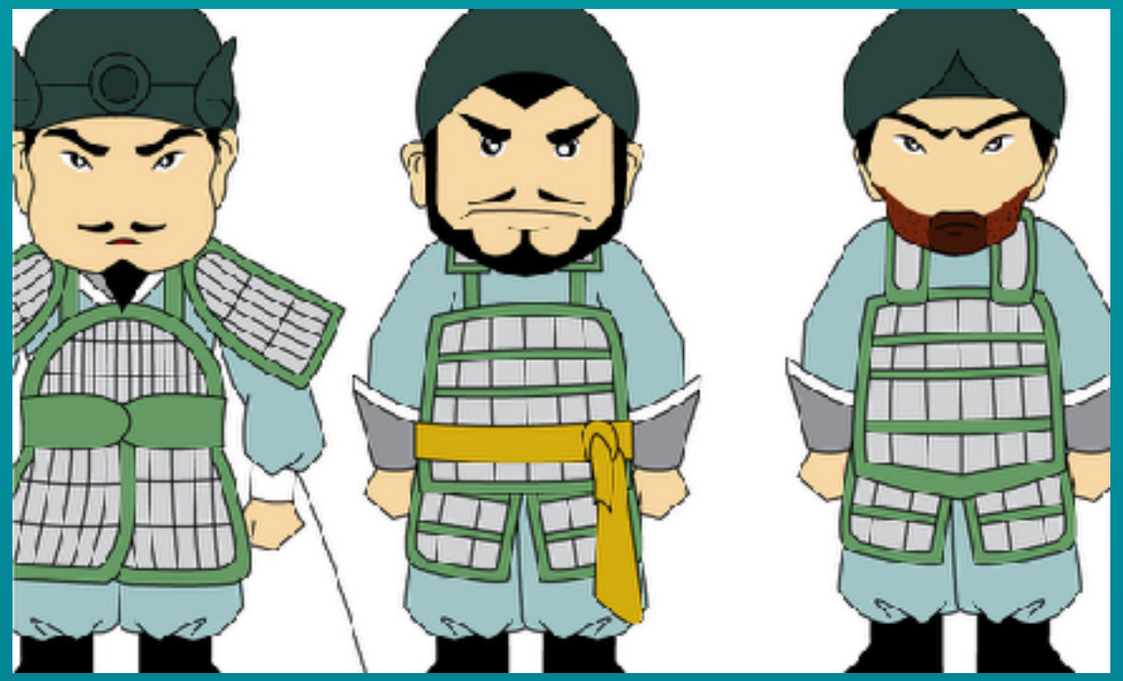
talk: 贝叶斯分类器的使用

- 1. 假设需要判断一笔交易是否为非法交易，如果需要用贝叶斯解决此问题， 需要考虑哪些东西?

- hint: 收集哪些数据? 考虑哪些指标? 如何衡量好坏?

机器学习





- 为什么出现了机器学习？ 解决哪些问题?
- 监督学习， 非监督学习， 强化学习 ..

什么是feature?

- 1. 设计feature
- 2. 深度学习的视角

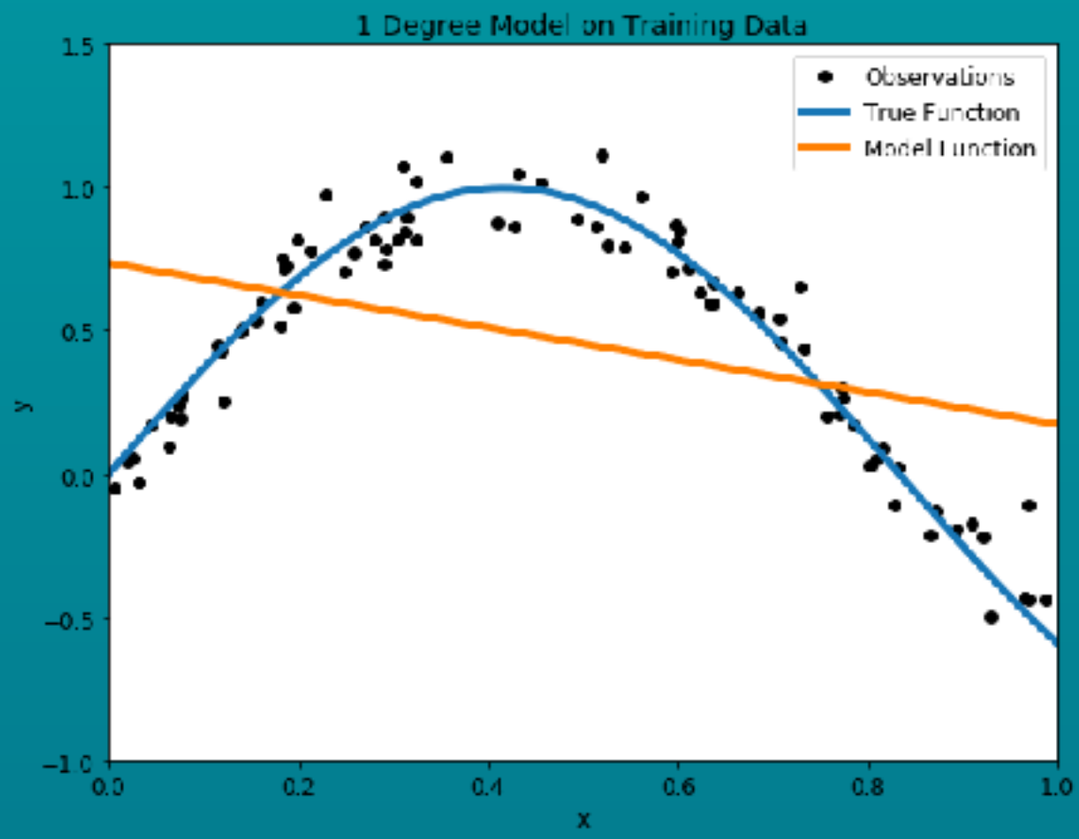
传统视角与深度学习的视角

如何衡量机器学习的好坏

- 训练时候的表现
  - loss函数
  - accuracy
  - precision
  - recall
  - AOC/AUC, f1_score, f2_score
- 泛化能力

Overfitting & Underfitting

- 以函数为例:
-



小作业

- 课后总结Overfitting与Underfitting产生的原因并总结成文档， 发送到钉钉群文件中；

- deadline: 2018-4-21

决策树

- 从泰坦尼克号沉船开始说起

Q: 如何决策?

熵和混乱程度

小作业

- 总结贝叶斯分类器和决策树分类的优缺点， 并总结成文字版发送到钉钉群文件中；
- deadline: 2018-4-21

Assignment-01

- Python网络爬虫
- Requests
- BeautifulSoup
- 广度优先和深度优先
- Task: 爬取豆瓣的电影评论
  - <豆瓣电影id， 电影名，评论， 5颗星>
  - 源代码参考： https://github.com/fortyMiles/get_douban_comments

Project-01

- 使用贝叶斯或决策树 在 scikit-learning 中建立模型， 预测其文章是否为新华社所发， 合理选取 feature， 避免过拟合。

- Deadline: 2018-4-28日

总结

- 1. 条件概率与贝叶斯分类器；
- 2. 机器学习的基本概念；
- 3. 决策树；
- 4. BSF, DFS的python实现；
- 5. 网络爬虫的基本实现；
- **6. 两个小作业， 一个大作业， 一个Project**