分类算法-NB

OutLine

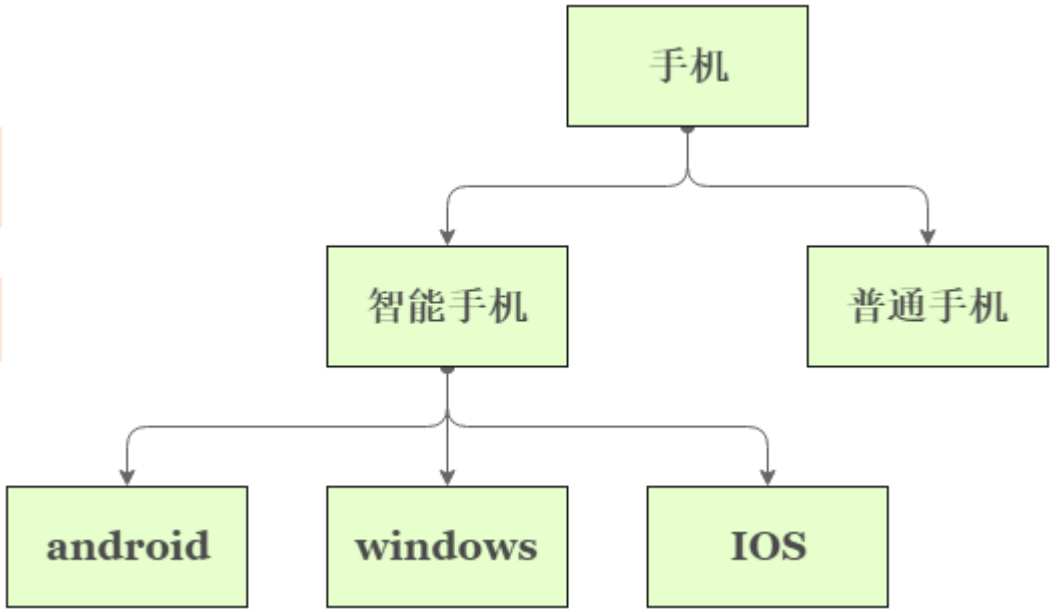
- 分类任务
- 朴素贝叶斯
- 【实践】基于MLlib的NB模型

分类技术概述

- 最常见的机器学习任务
- 定义：给定一个对象X，将其划分到预定义好的某一个类别Yi中
  - 输入：X
  - 输出：Y（取值于有限集合{y1,y2,.....,yn}）
- 应用:
  - 人群，新闻分类，query分类，商品分类，网页分类，垃圾邮件过滤，网页排序

不同类型的分类

- 类别数量
  - 二值分类
    - Y的取值只有两种，如：email是否垃圾邮件
  - 多值分类
    - Y的取值大于两个，如：网页分类{政治，经济，体育，……}
- 类别关系
  - 水平关系
    - 类别之间无包含关系
  - 层级关系
    - 类别形成等级体系



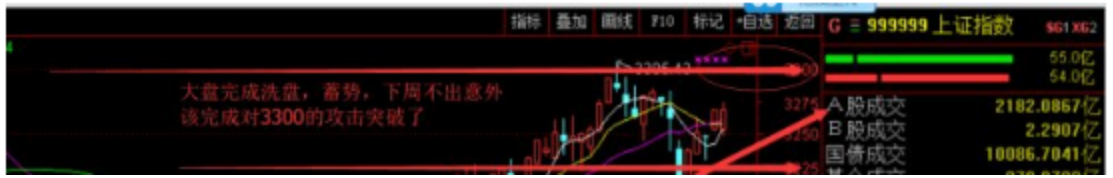
新 闻 分 类

下周将突破3300迎来一片红！

2017-08-18 17:08:32　评论(0)　财经　大盘　分析　个股　新闻

1

最近几日龙哥外出有事，事情比较急来不及请假了，很抱歉！下周一直播恢复正常、接下来来简单的分析下！周五大盘被外围股市影响，开盘明显低开低走，一度跌破3250点，迅速回补缺口后开始弱势低位震荡，早盘低位多次强调大盘正常借势洗盘，下午红盘概率大，果然大盘下午逐波拉红，个股来说跌多涨少，下午能红盘几乎都是权重股的功劳，期指最近和市场步调不一致，大部分昨天走强的个股周五极度低迷，说明市场持续性炒作氛围不够，央企混改题材股成为今天相对最强的热点.混改题材股涵盖了很多中字头个股，需要资金量较大，现在市场这点量能想撬动有很大难度，所以从盘口就看到了此消彼涨的态势，早盘刚开始雄安较强，但混改确认强势后雄安就明显出现冲高回落态势，所以从分时走势来看，雄安大部分个股日内走的非常难看，而混改股正好反过来，下午越走越强势，我们最好跟着资金走，建议适当关注混改股。



新闻分类

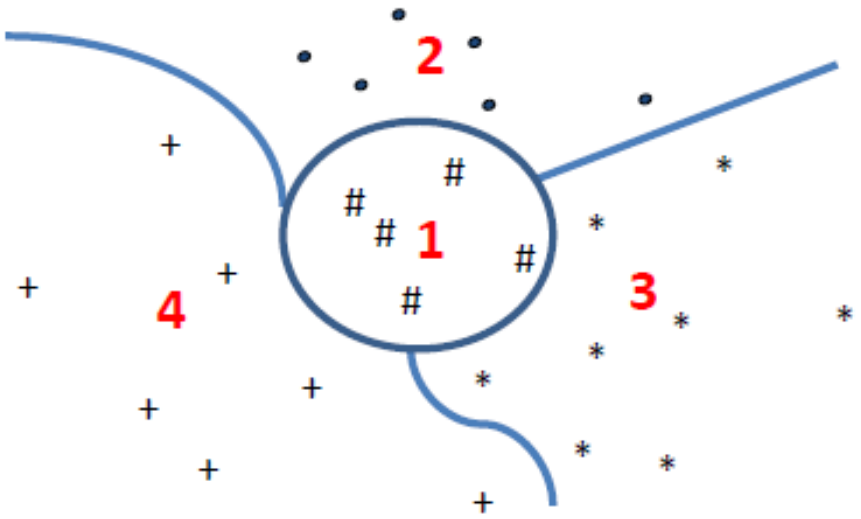
- 任务
  - 为任一新闻，例如{股市，反弹，有力，基金，建仓，加速……}
  - 指定其类别=>{军事，科技，财经，生活……}
- 基于规则的方式
  - 列举每个类别的常用词
    - 军事：导弹，军舰，军费……
    - 科技：云计算，siri，移动互联网……
  - 问题
    - 如何保证列举全?
    - 冲突如何处理？苹果：科技？生活?
    - 不同的词有不同的重要度，如何决定?
    - 如果类别很多怎么办?

分类任务解决流程

- 新闻分类
- 特征表示：X={昨日，是，国内，投资，市场.....}
- 特征选择：X={国内，投资，市场.....}
- 模型选择：朴素贝叶斯分类器
- 训练数据准备
- 模型训练
- 预测（分类）
- 评测

分 类 技 术

- 概率分类器
  - NB
  - 计算待分类对象属于每个类别的概率，选择概率最大的类别作为最终输出
- 空间分割
  - SVM
- 其他
  - KNN



OutLine

- 分类任务
- 朴素贝叶斯
- 【实践】基于MLlib的NB模型

朴素贝叶斯分类

- 朴素贝叶斯（NaiveBeyesian Classification，NB）分类器
  - 概率模型
  - 基于贝叶斯原理

$$p(y_i|X) = \frac{p(X|y_i)P(y_i)}{P(X)} = \frac{P(y_i)\prod_j P(x_j|y_i)}{P(X)}$$

- P(X)：待分类对象自身的概率，可忽略
- P(yi)：每个类别的先验概率，如P(军事)
- P(X|yi)：每个类别产生该对象的概率
- P(xi|yi)：每个类别产生该特征的概率，如P(苹果|科技)

模型训练、参数估计

- 策略：最大似然估计（maximum likehood estimation，MLE）
  - P(Yi)
    - Count(yi)：类别为yi的对象在训练数据中出现的次数
  - 例如：
    - 总共训练数据1000篇，其中军事类300篇，科技类240篇，生活类140篇，……
    - P(军事)=0.3， P(科技)=0.24， P(生活)=0.14，……

模型训练、参数估计

- 最大似然估计（maximum likehood estimation，MLE）

$$p(x_{j|}|y_i) = \frac{Count(x_j, y_i)}{Count(y_i)}$$

  - P(xj|yi)
    - Count(xj, yi)：特征xj和类别yi在训练数据中同时出现的次数
  - 例如：
    - 总共训练数据1000篇，其中军事类300篇，科技类240篇，生活类140篇，……
    - 军事类新闻中，谷歌出现15篇，投资出现9篇，上涨出现36篇
    - P(谷歌|军事)=0.05， P(投资|军事)=0.03， P(上涨|军事)=0.12，……

模 型 示 例

$$p(y_i|X) = \frac{P(X|y_i)P(y_i)}{P(X)} = \frac{P(y_i)\prod_j P(x_j|y_i)}{P(X)}$$

- P(yi)
  - p(军事)=0.3， p(科技)=0.24， p(生活)=0.14，……
- P(xj|yi)
  - P(谷歌|军事)=0.05， P(投资|军事)=0.03， P(上涨|军事)=0.12，……
  - P(谷歌|科技)=0.15， P(投资|科技)=0.10， P(上涨|科技)=0.04，……
  - P(谷歌|生活)=0.08， P(投资|生活)=0.13， P(上涨|生活)=0.18，……
  - ……

预 测

- 分类原则

$$p(y_i|X) = \frac{P(X|y_i)P(y_i)}{P(X)} = \frac{P(y_i)\prod_j P(x_j|y_i)}{P(X)}$$

  - 给定X，计算所有的p(yi|X)，选择概率值最大的yi作为输出
    - X={国内，投资，市场，……}
    - P(军事|X)=P(国内|军事)* P(投资|军事)* P(市场|军事)……P(军事)
    - 同样计算P(科技|X) P(生活|X)
  - 二值和多值分类同样的做法

评 测

- 测试数据
  - (微软更新必应搜索，科技)
  - (名企精装修直降30万，房产)
  - (国际版块利空突袭 周一大盘堪忧，财经)
  - ……
- 混淆表

混淆表(confusion table)		分类器预测的类别	
		y1	y2
实际的类别	y1	C11	C12
	y2	C21	C22

评 测 指 标

混淆表(confusion table)		分类器预测的类别	
		y1	y2
实际的类别	y1	C11	C12
	y2	C21	C22

- 准确度Accuracy：(C11+C22)/(C11+C12+C21+C22)
- 精确率Precision（y1）：C11/(C11+C21)
- 召回率Recall（y1）：C11/(C11+C12)

评 测 指 标

混淆表(confusion table)		分类器预测的类别	
		军事	科技
实际的类别	军事(60)	50	10
	科技(40)	5	35

- 准确度Accuracy：(50+35)/(35+5+10+50)=85%
- 精确率Precision（y1）：50/(50+5)=90.9%
- 召回率Recall（y1）：50/(50+10)=83.3%

朴素贝叶斯分类特点

- 优点：
  - 简单有效
  - 结果是概率，对二值和多值同样适用
- 缺点：
  - 独立性假设有时不合理

OutLine

- 分类任务
- 朴素贝叶斯
- 【实践】基于MLlib的NB模型

Q & A

@八斗学院