数据分析与可视化

Fred

数据分析与数据挖掘、机器学习的关系

数据分析

维基百科

数据分析是指用适当的统计分析方法对收集来的大量数据进行分析，提取有用信息和形成结论而对数据加以详细研究和概括总结的过程。这一过程也是质量管理体系的支持过程。在实用中，数据分析可帮助人们作出判断，以便采取适当行动。

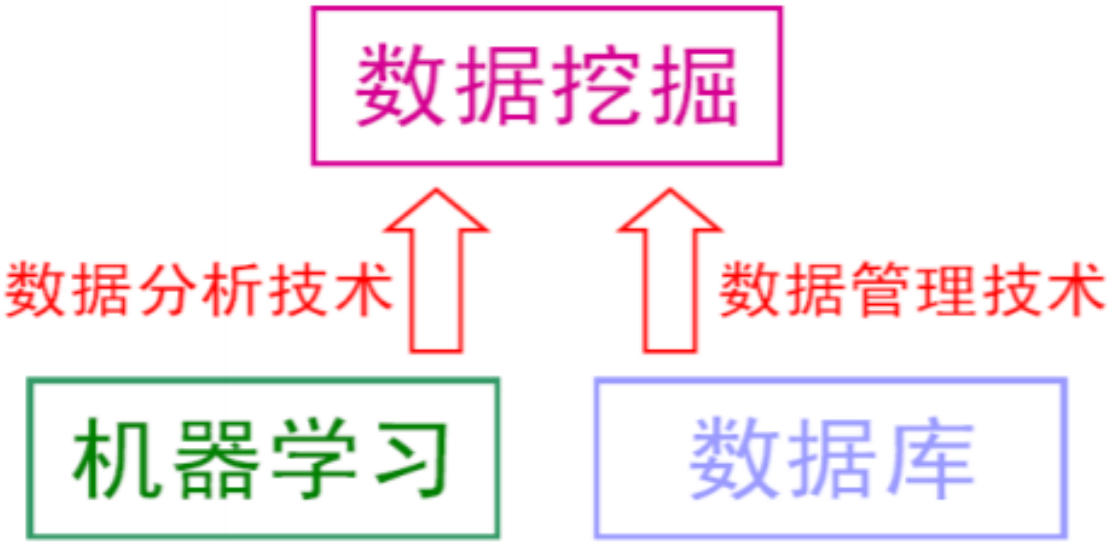
数据分析的数学基础在 20 世纪早期就已确立，但直到计算机的出现才使得实际操作成为可能，并使得数据分析得以推广。数据分析是数学与计算机科学相结合的产物。

数据分析与数据挖掘的关系

数据分析只是在已定的假设，先验约束上处理原有计算方法，统计方法，将数据分析转化为信息，而这些信息需要进一步的获得认知，转化为有效的预测和决策，这时就需要数据挖掘，也就是我们数据分析师系统成长之路的“更上一楼”。

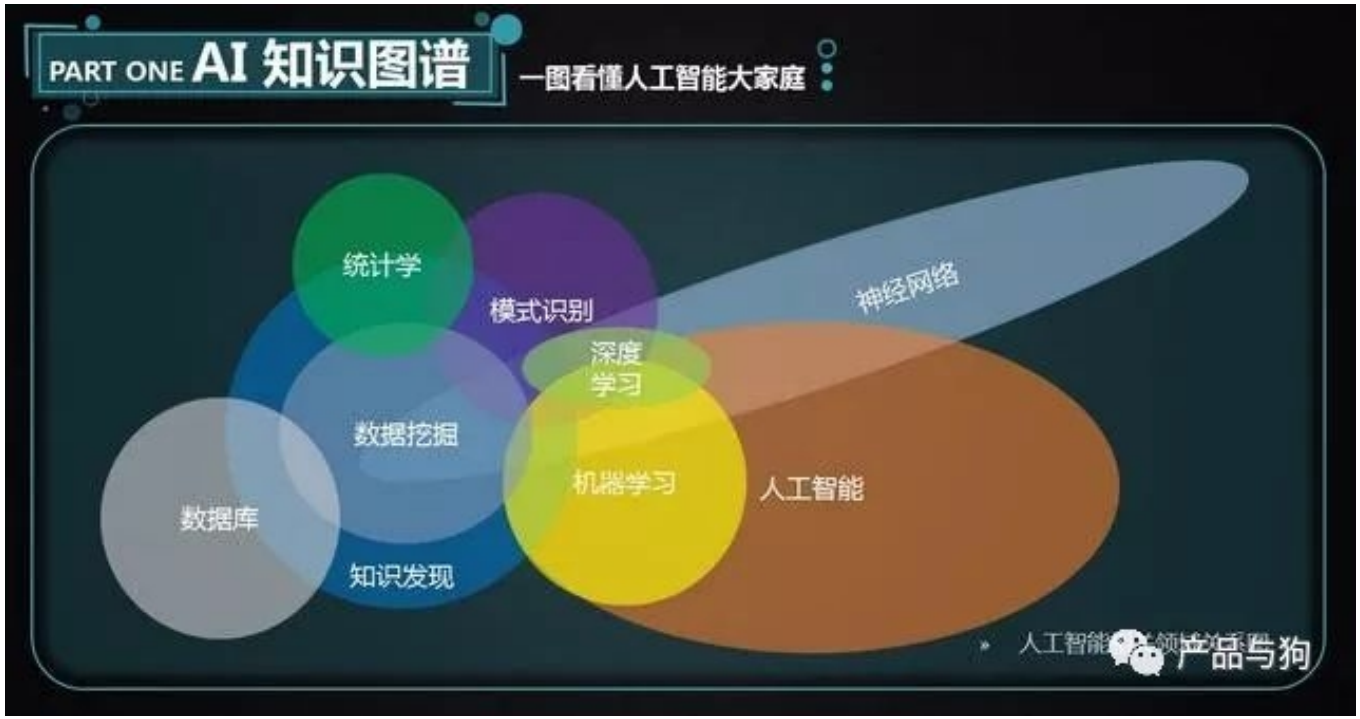
数据挖掘与数据分析两者紧密相连，具有循环递归的关系，数据分析结果需要进一步进行数据挖掘才能指导决策，而数据挖掘进行价值评估的过程也需要调整先验约束而再次进行数据分析。

数据挖掘与机器学习的关系



[1] 机器学习与数据挖掘，周志华，
https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/cccf07.pdf

AI 知识图鉴



[1] AI 知识图鉴：机器学习、深度学习、数据分析、数据挖掘, http://wemedia.ifeng.com/89898794/wemedia.shtml

基于 Pandas 数据分析常用函数功能 Summary

- Syntax–Creating DataFrames
- Reshaping Data –Change the layout of a data set
- Summarize Data
- Handing Missing Data
- Make New Variables
- Subset Observations (Rows)
- Subset Variables (Columns)
- Combine Data Sets
- Group Data
- Windows
- Plotting

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.html#pandas.DataFrame

美国大选

熟悉数据来源

This dataset is a collection of state and national polls conducted from November 2015-November 2016 on the 2016 presidential election. Data on the raw and weighted poll results by state, date, pollster, and pollster ratings are included.

该数据集是 2015 年 11 月至 2016 年 11 月 2016 年总统大选期间进行的州和全国民意调查的集合。包括州，日期，民意测验机构和民意测验机构的评级维度的原始和加权民意调查结果数据。

熟悉数据内容和格式

- type：We're forecasting the election with three models
  - Polls-plus forecast：What polls, the economy and historical data tell us about Nov. 8
  - Polls-only forecast：What polls alone tell us about Nov. 8
  - Now-cast：Who would win the election if it were held today
- population
  - A = ADULTS
  - RV = REGISTERED VOTERS
  - V = VOTERS
  - LV = LIKELY VOTERS

熟悉数据内容

读入数据

data = pd.read_csv('presidential_polls.csv')

查看数据

```
data.head()
```

美国有多少个州？

data.groupby(“state”).groups.keys()

常用统计量

describe([percentiles, include, exclude])
Generate descriptive statistics that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values.

data.describe()

什么是分位数？

query 操作

取出评级为 C-的数据

data.query(“grade == ‘%s’ ” % ‘C-’)

日期数据处理

建议阅读此网页了解 python 日期的基本操作：
https://docs.python.org/2/library/datetime.html

字符串转化为 python 日期对象

from datetime import datetime, date, time

t1 = datetime.strptime("21/1/2006 11:54:05", "%d/%m/%Y %H:%M:%S")

python 日期对象转换为字符串

datetime.strftime(t1, "%Y-%m-%d")

列数据自定义函数处理

Series.apply(func, convert_dtype=True, args=(), **kwds)[source]

日志字符串转化为 date 对象

data.startdate.apply(lambda x: datetime.strptime(x, "%m/%d/%Y"))

增加一列

原始日期字符串转为 date 对象

```
data[“new_startdate”] = data.startdate.apply(lambda x:
datetime.strptime(x, “%m/%d/%Y”))
```

增加一列带权的 poll 字段

```
data[‘w_rawpoll_trump’] = data[‘poll_wt’] *
data[‘rawpoll_trump’]
```

指定 value 排序

data.sort_values([“startdate”])

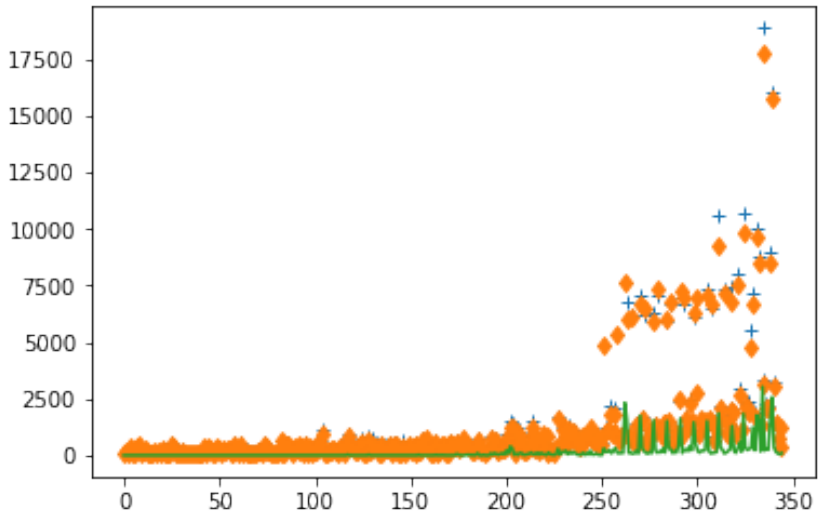
按 index 排序

t = data.groupby("startdate").sum()

t.sort_index()

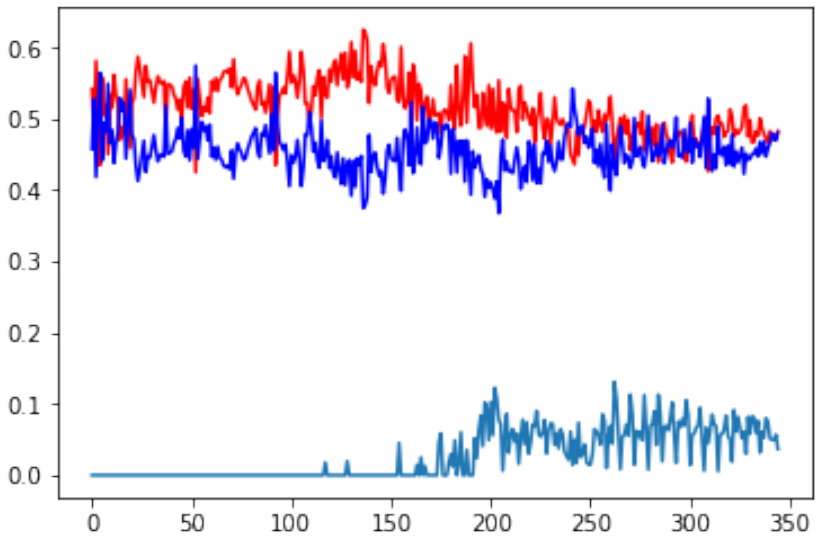
What are the trends of the polls over time by day?

分别统计每个候选人不同 startdate 的投票数之和？



看不出趋势？

归一化



Trump 后段发力？

代码

```
data = pd.read_csv('presidential_polls.csv')
data["new_startdate"] = data.startdate.apply(
  lambda x: datetime.strptime(x, "%m/%d/%Y"))
t = data.groupby("new_startdate").sum()
t = t.sort_index()
trump = t.rawpoll_trump.values
clinton = t.rawpoll_clinton.values
johnson = t.rawpoll_johnson.values
all_poll = clinton + trump + johnson
plt.plot(clinton/all_poll, '+')
plt.plot(trump/all_poll,'d')
plt.plot(johnson/all_poll)
```

week trend?

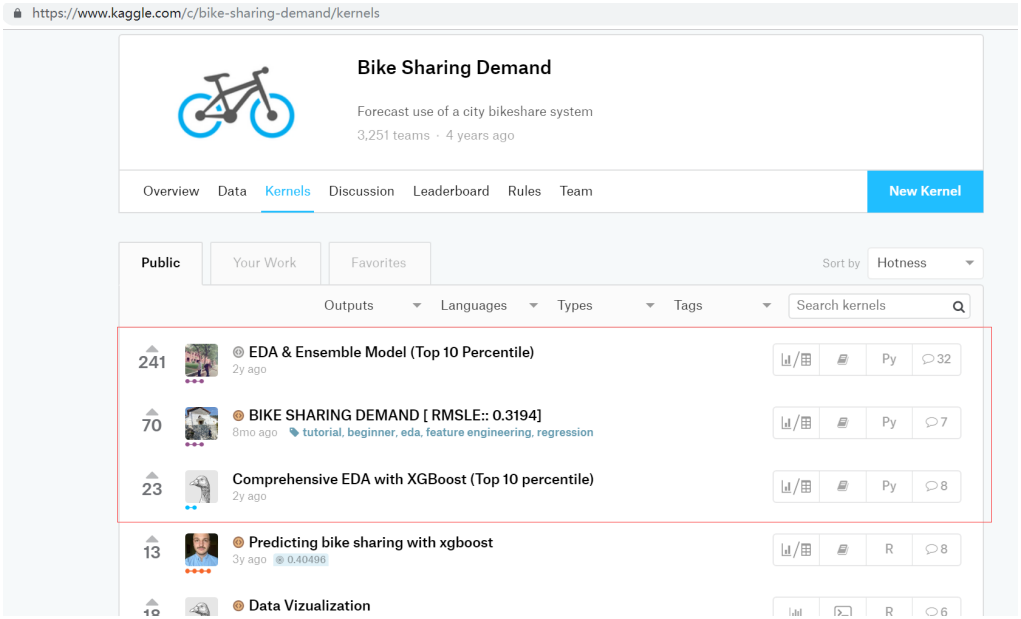
```
days = data.new_startdate.values
days.sort()
data["week_index"] = data.new_startdate.apply(
  lambda x: int((x - days[0]).days / 7))
t = data.groupby("week_index").sum()
```

month trend?

```
data["month_index"] = data.new_startdate.apply(
  lambda x:  datetime.strftime(x, "%Y%m"))
t = data.groupby("month_index").sum()
```

共享单车

参考公开的代码学习



下次课会详细讲解