知乎 LIVE

机器学习入门之特征工程


王晋东不在家

http://jd92.wang

20:30-21:30 2017.03.19

知乎 Live

目录

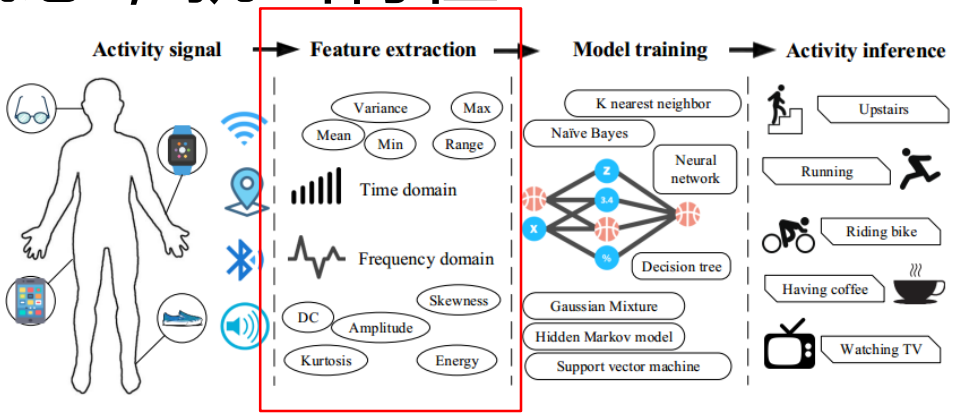
特征工程简介（1）

- 什么是特征工程？
  - 小李相亲回来,别人问：对方有什么特点？
  - 小白兔，白又白，两只耳朵竖起来
  - “你化成灰我也认识你！”

共有的　　能进行概括

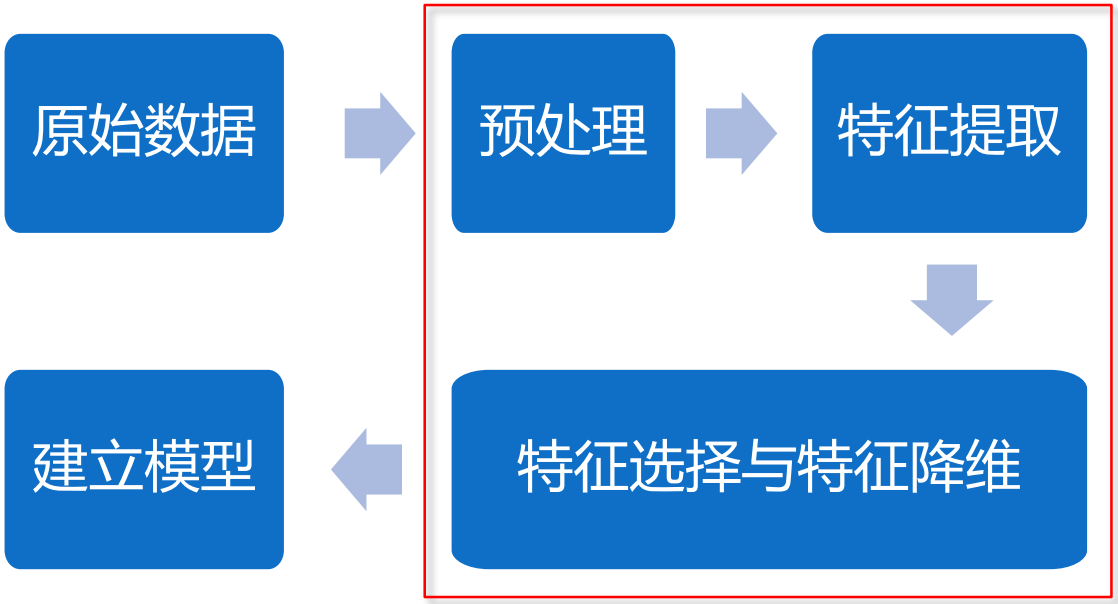
- 我们根据事物所具有的共性所抽象出来的能代表这一事物的概念，就叫特征

能代表这一类事物



特征工程简介（2）

- 为什么特征工程是重要的？
  - “数据和特征决定了机器学习的上限，而模型和算法只是逼近这个上限而已。”
  - 除了数据和模型外，特征是影响学习效果的重要因素。



我们本次Live重点

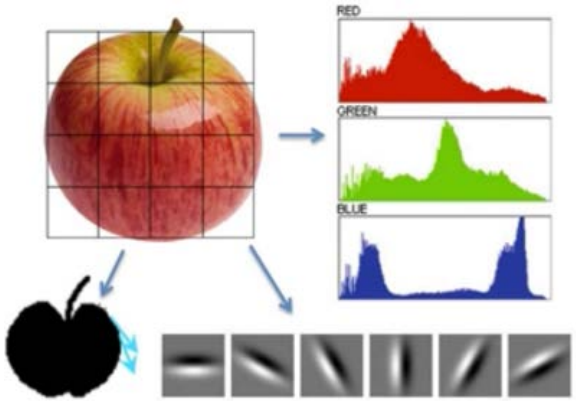
图片：http://www.jianshu.com/p/35135ab0a627

特征工程——预处理

- 预处理：
  - 现实世界数据是“肮脏”的，好数据→好结果
  - 对原始数据的清洗、过滤、缺失值处理、标准化、归一化等，使其更方便做后期的特征处理和机器学习。
- 常用预处理方法：
  - 去重：去掉重复的数据
  - 过滤：把反常值（极高/极低）用平滑的值代替
  - 标准化：把数据放缩到同样的范围（比如0均值、最大最小归一化）
  - 缺失值处理：用已有的数据补全丢掉的数据（如均值）
  - 离散化：把数据按不同区间（箱子）划分（分箱）

特征工程——特征提取

- 如何获取重要特征？
  - 相关领域的专家知识
    - 行为识别领域，加速度提取时域和频率信息
  - 深度学习自动学习特征
    - 人脸识别、图片识别等应用，依赖卷积神经网络
  - 有时原始数据本身是特征
    - 时间序列预测时，原始数据本身就是特征
  - 实验、经验与发现
    - 不断尝试新的特征
    - 深度学习并不是万能的



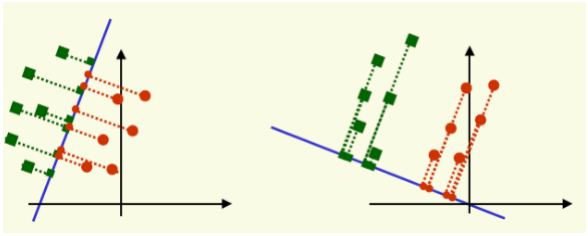
特征工程——特征选择

- 特征选择方法：
  - 过滤法：
    - 方差选择法：计算各个特征方差，选择方差大于阈值的特征
    - 相关系数法：计算各个特征的Pearson相关系数
    - 互信息法：计算各个特征的信息增益
  - 封装法：
    - 递归消除法：使用基模型(如LR)在训练中进行迭代，选择不同特征
  - 嵌入法：
    - 使用带惩罚项的基模型进行特征选择(比如LR加入正则)
    - 树模型的特征选择（随机森林、决策树）

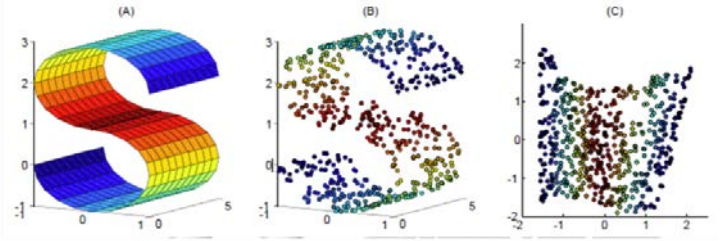
参考：https://www.zhihu.com/question/29316149

特征工程——降维

- 常用特征降维方法：
  - 线性降维
    - 主成分分析（PCA）：选择方差最大的K个特征[无监督]
    - 线性判别分析（LDA）：选择分类性能最好的特征[有监督]
  - 非线性降维（大多是流形学习）
    - 核主成分分析（KPCA）：带核函数的PCA
    - 局部线性嵌入（LLE）：利用流形结构进行降维
    - 还有拉普拉斯图、MDS等
  - 迁移成分分析（TCA）：不同领域之间迁移学习降维



PCA和LDA



LLE

实际应用——自行车租赁（1）

- Kaggle自行车租赁预测
  - 给定不同的条件（天气、假期、温度、湿度等）,预测未来租赁自行车的数量
  - https://www.kaggle.com/c/bike-sharing-demand

datetime	season	holiday	workingday	weather	temp	atemp	humidity	windspeed	casual	registered	count
2011-01-01 00:00:00	1	0	0	1	9.84	14.395	81	0.0	3	13	16
2011-01-01 01:00:00	1	0	0	1	9.02	13.635	80	0.0	8	32	40
2011-01-01 02:00:00	1	0	0	1	9.02	13.635	80	0.0	5	27	32
2011-01-01 03:00:00	1	0	0	1	9.84	14.395	75	0.0	3	10	13
2011-01-01 04:00:00	1	0	0	1	9.84	14.395	75	0.0	0	1	1

实际应用——自行车租赁（2）

知乎 Live

- 特征构建：

原始数据处理方式

时间格式

租赁日期与时间 → 转成单独年、月、日

数值格式

温度、湿度、风速
注册人数 → 直接进行使用

类别格式

是否假期、工作日
季节 → 用数值代替类别
One-hot编码

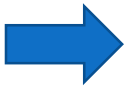
自己提取特征

- 平均一周（月）的人数、温度等
- 对数据做不同阶的差分
- 自己爬取空气质量信息（如果有）
- 节假日详细信息
- 节假日数据的统计信息

- 然后进行特征选择与降维

实际应用——自行车租赁（3）

- 模型构建：
  - SVM for regression
  - GBR
  - XGBOOST
  - Random Forest
  - KNN
  - ……



模型stack

- Linear regression
- Ridge regression
- Lasso
- SVR
- KNN
- ……

- 训练多模型时，先选大步长大范围调参，第二遍再选小范围小步长调一次
- 重视Stack的作用

实际应用——豆瓣电影评分（1）

知乎 Live

- 豆瓣电影评分预测：
  - 根据之前已有电影数据信息,预测即将上映电影的豆瓣评分

一条狗的使命 A Dog's Purpose (2017)

导演: 拉斯·霍尔斯道姆
编剧: W·布鲁斯·卡梅伦 / 凯瑟琳·迈克 / 奥黛丽·威尔斯 / 玛雅·福布斯 / 沃利·亚历斯戴尔
主演: 布丽特·罗伯森 / 丹尼斯·奎德 / 佩吉·利普顿 / 乔什·加德 / K·J·阿帕 / 更多...
类型: 剧情 / 喜剧 / 冒险
官方网站: www.adogspurposemovie.com
制片国家/地区: 美国

豆瓣评分

54262人评价

5星	26.0%
4星	42.0%
3星	27.7%
2星	3.7%
1星	0.6%

  - 如何爬取合适的信息进行预测？

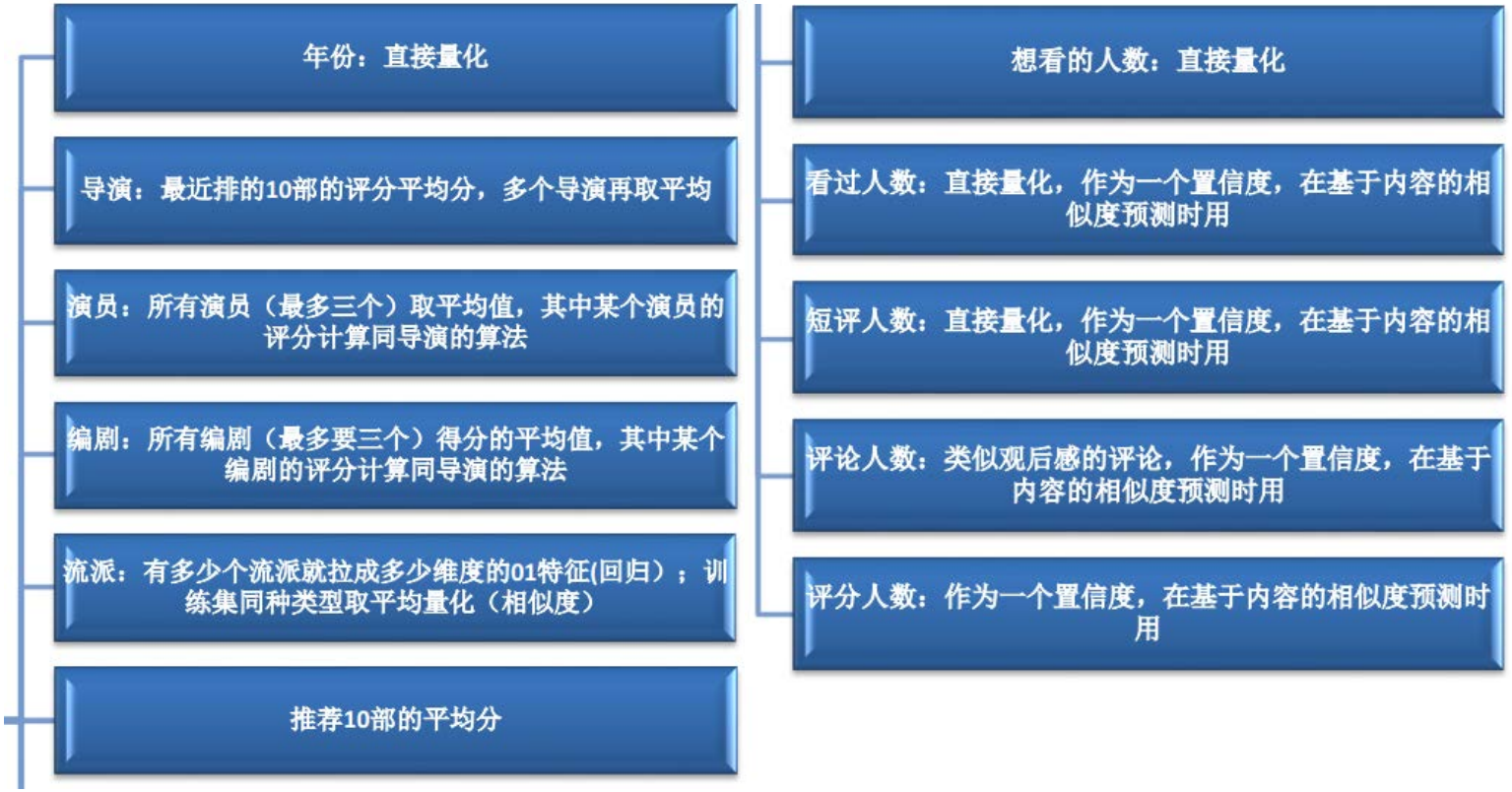
图片来源：https://movie.douban.com/subject/6873143/

实际应用——豆瓣电影评分（2）

- 分析：

实际应用——豆瓣电影评分（3）

- 我们提取的特征：



实际应用——豆瓣电影评分（4）

- 处理流程：
  - 提取上述特征并做规范化；
  - 建立2个模型分别进行评测：
    - SVM for regression
    - 余弦相似度模型
  - 最后的结果较为满意

```
train500 set rmse:
    0.3607
verify30 set rmse:
    0.3173
test20 set rmse:
    0.3937
```

```
<movie>
    <id>6875263</id>
    <name>灰姑娘 Cinderella</name>
    <rating>6.5</rating>
</movie>
<movie>
    <id>11026735</id>
    <name>超能陆战队 Big Hero 6</name>
    <rating>7.6</rating>
</movie>
<movie>
    <id>5154799</id>
    <name>木星上行 Jupiter Ascending</name>
    <rating>5.7</rating>
</movie>
<movie>
    <id>3993588</id>
    <name>狼图腾</name>
    <rating>7.7</rating>
</movie>
```

- 总结：提取适合的特征，很重要！

实际应用——推荐工具和资源

- 进一步了解特征工程：
  - http://www.cnblogs.com/jasonfreak/p/5448385.html
  - http://www.csuldw.com/2015/10/24/2015-10-24%20feature%20engineering/
- 工具
  - Python中的Scikit-learn，可以特征选择、降维
  - Matlab下的各种降维函数的使用
- 实战
  - Kaggle竞赛入门：https://www.kaggle.com/competitions?sortBy=prize&group=active&page=1&segment=gettingStarted

知乎 LIVE

机器学习中的特征工程


王晋东不在家

http://jd92.wang

2017.03.19 20:30-21:30