评估

数据

训练模型

70% 30%

80% 20%

75% 25%

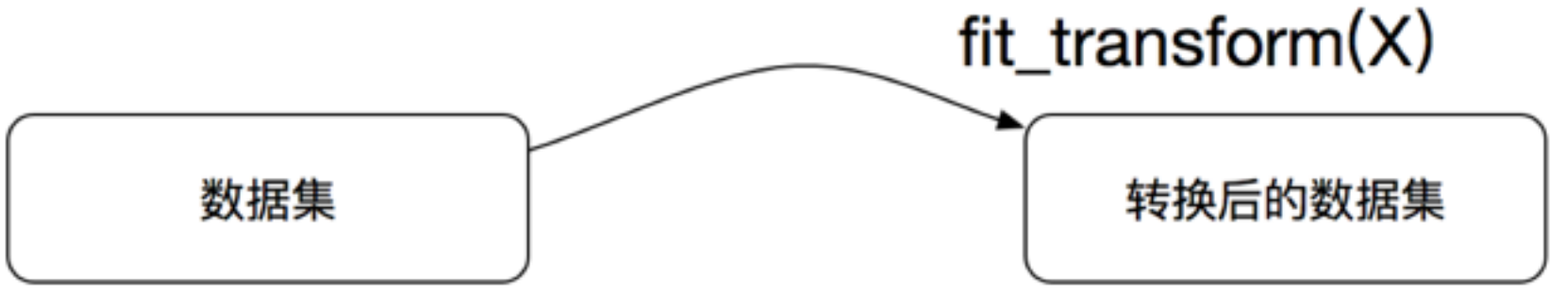
建立模型

评估模型

分类

回归

聚类



文章1，文章2　　文章3，文章4

fit_transform():输入数据直接转换

fit():输入数据，但不做事情　[[23,34,5],[20,19,45]]　计算平均值，方差等等

+transform(): 进行数据的转换

[[23,34,5],[20,19,45]]

[[1,2,3],
[4,5,6]]

[[2,3,4],
[4,5,7]]

2.5,3.5,4.5

3,4, 5.5

每个算法API当中的参数

estimator

训练集　　测试集

x_train,
y_train

1、调用fit

fit(x_train, y_train)

估计器　estimator

1、y_predict = predict(x_test)

x_test,
y_test

2、预测的准确率：score(x_test, y_test)

2、输入与测试集的数据

电影名称	打斗镜头	接吻镜头	电影类型	与未知电影的距离
California Man	3	104	爱情片	20.5
He's not Really into dues	2	100	爱情片	18.7
Beautiful Woman	1	81	爱情片	19.2
Kevin Longblade	101	10	动作片	115.3
Robo Slayer 3000	99	5	动作片	117.4
Amped II	98	2	动作片	118.9
?	18	90	未知	

如何求距离?

根号((18-3)^2 +(90-104)^2) = 20.5

相似的样本，特征之间的值应该都是相近的

两个样本的距离可以通过如下公式计算，又叫欧式距离

比如说，a(a1,a2,a3),b(b1,b2,b3)

$$\sqrt{(a1-b1)^2+(a2-b2)^2+(a3-b3)^2}$$

K-近邻算法：需要做标准化处理

K 取值：1 , 3, 5, 6

分类

特征值：x, y 坐标， 定位准确性，年，日，时，周　　目标值：入住位置的id

处理：　　0<x<10　0<y<10

1、由于数据量大，节省时间 x, y缩小

2、时间戳进行（年，月，日，周，时分秒）， 当做新的特征

3、几千~几万，少于指定签到人数的位置删除

pd.to_datatime

样本数	职业	体型	女神是否喜欢
1	程序员	超重	不喜欢
2	产品	匀称	喜欢
3	程序员	匀称	喜欢
4	程序员	超重	喜欢
5	美工	匀称	不喜欢
6	美工	超重	不喜欢
7	产品	匀称	喜欢

条件：所有特征之间是条件独立

自然语言处理（不独立）

记作：$P(A,B)$

$P(A, B) = P(A)P(B)$

问题

1、女神喜欢的概率？ **4/7**

错的

2、职业是程序员并且体型匀称的概率？ **P(程序员，匀称) = 3/7 * 4/7 = 12/49**

3、在女神喜欢的条件下，职业是程序员的概率？ **2/4 = 1/2**

4、在女神喜欢的条件下，职业是产品，体重是超重的概率？

P(产品，超重|喜欢) = P(产品|喜欢)P(超重|喜欢) = 1/2*(1/4) = 1/8

记作：$P(A|B)$

特性：$P(A1,A2|B) = P(A1|B)P(A2|B)$

”朴素“贝叶斯　　特征独立

P(科技 |词1，词2，词3…)　　文档1：词1，词2，词3……

P(娱乐 |词a,词b….)

公式可以理解为：

$$P(C|F1,F2,\dots)=\frac{P(F1,F2,\dots|C)P(C)}{P(F1,F2,\dots)}$$

其中c可以是不同类别

P(科技 |词1，词2，词3…) = P(F1,f2,f3|科技)P(科技)

P(娱乐 |词1,词2….) = P(F1,f2,f3|娱乐)P(娱乐)

训练集很多文档　词的列表　　重要的词　　训练集误差大，结果肯定不好

不许要调参

训练集统计结果(指定统计词频)：

特征\统计	科技(30篇)	娱乐(60篇)	汇总（90篇）
“商场”	9	51	60
“影院”	8	56	64
“支付宝”	20	15	35
“云计算”	63	0	63
汇总(求和)	100	121	221

现有一篇被预测文档：出现了影院，支付宝，云计算，计算属于科技、娱乐的类别概率？

预测文档?

P(科技| 影院，支付宝，云计算) = P(影院,支付宝,云计算|科技)P(科技)

=(8+1/100+1*4)(20/100)(63/100)(30/90) = 0.00456109

P(娱乐| 影院，支付宝，云计算)=P(影院,支付宝,云计算|娱乐)P(娱乐)

= (56+1/121+1*4)(15+1/121+1*4)(0+1/121+1*4)(60/90) = 0.001

假设了 文章当中一些词语另外一些是独立没关系　　不太靠谱

训练集当中去进行统计词这些工作　　会对结果造成干扰

朴素贝叶斯：文本分类
神经网络 效果要好

混淆矩阵

二分类

评估标准：准确率

精确率和召回率

- 在分类任务下，预测结果(Predicted Condition)与正确标记(True Condition)之间存在四种不同的组合，构成混淆矩阵(适用于多分类)

是猫 预测结果 不是猫

真实结果	正例	假例
正例	true positive 真正例TP	false negative 伪反例FN
假例	false positive 伪正例FP	true negative 真反例TN

正例：猫

反例：不是猫

猫 混淆矩阵

狗 混淆矩阵

癌症检测

20 癌症

80 非癌症

100

100 癌症

10 癌症

90 非癌症

交叉验证： 所有数据分成n等分

k =1,　 5,　7, 10

4折交叉验证

验证集	训练集	训练集	训练集

得出一个准确率 模型1

训练集	验证集	训练集	训练集

得出一个准确率 模型2

求平均值模型结果8

训练集	训练集	验证集	训练集

得出一个准确率 模型3

训练集	训练集	训练集	验证集

得出一个准确率 模型4

网格搜索：调参数　K-近邻：超参数K

a　[2,3,5,8,10]　　b　[20,70,80]　两两组合　　15

1~32

1~16　　17~32

1个字节　8bit

5次

如果：不知道任何一个球队的信息的话，5bit　1/32　1/32　…….

5 = -(1/32log1/32 + 1/32log1/32 + ……..)

开放一些数据信息　　5 > -(1/4log1/4 + 1/4log1/4 + ……..)　信息熵

1/6　　1/6　　1/10

德国　　巴西　　中国

ID	年龄	有工作	有自己的房子	信贷情况	类别
1	青年	否	否	一般	否
2	青年	否	否	好	否
3	青年	是	否	好	是
4	青年	是	是	一般	是
5	青年	否	否	一般	否
6	中年	否	否	一般	否
7	中年	否	否	好	否
8	中年	是	是	好	是
9	中年	否	是	非常好	是
10	中年	否	是	非常好	是
11	老年	否	是	非常好	是
12	老年	否	是	好	是
13	老年	是	否	好	是
14	老年	是	否	非常好	是
15	老年	否	否	一般	否

一个特征条件之后，减少的信息熵的大小

$$H(D) = -(9/15\log 9/15 + 6/15\log(6/15))$$

$$g(D, 年龄) = H(D) - H(D'|年龄) = 0.971 - [1/3H(青年)+1/3H(中年)+1/3H(老年)] =$$

$$H(青年) = -(2/5\log(2/5)+ 3/5\log(3/5))$$

$$H(中年) = -(2/5\log(2/5)+ 3/5\log(3/5))$$

$$H(老年) = -(4/5\log(4/5)+ 1/5\log(1/5))$$

决策树的分类依据之一：信息增益

基尼系数：划分更加仔细

随机森林建立多个决策树的过程：

N个样本， M个特征

单个树建立过程：

1、随机在N个样本当中选择一个样本，重复N次　样本有可能重复

2、随机在M个特征当中选出m个特征　　m取值

建立10颗决策树，样本，特征大多不一样

随机又放回的抽样　bootstrap

随机森林：n_estimator决策树的数量

maxx_depth:每颗树的深度限制