3.文本情感分析

情感分析简介

情感分析（Sentiment Analysis）是文本挖掘中的经典研究方法，它指对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程。很多文本数据（例如用户评论数据）都**反映了文本作者的情感色彩和情感倾向性，如喜、怒、哀、乐和批评、赞扬**等。通过分析文本中蕴含的感情倾向和主观色彩就可以进一步了解作者对于某一事件或某一主题的具体看法。

常用的文本情感分析方法

当前比较常用的文本情感分析方法主要有两种。

- 一种是**基于机器学习的情感分析方法**，将情感词作为主要特征词，并结合其他特征训练分类器，比较依赖特征选择和数据规模。这类方法需要事先标注一部分文档的情感倾向作为训练集，然后将情感分析问题转化为有监督学习问题，通过采用机器学习方法进行建模，然后预测大规模未标记文档的情感得分。
- 另一种常用方法则是**基于情感词典的文本分析方法**。该方法通过比较已有词典中的词语与需要评分的词语，根据语料库中已有词语的评分和相关规则来给出目标文本的评分。本节将主要介绍如何基于词典法进行情感分析。

3.1、情感词典法

基于情感词典的分析方法

- 通过词典法，我们通常可以识别出一个文本中哪些词表示正向情感倾向，哪些词表示负向情感倾向。
- 假设N=负向情感倾向词的个数，P=正向情感倾向词的个数。如果只需要笼统的判断一个文档是正向还是负向，则可以假设：
  - ✓ 如果N-P>0，则文本表示负向情感；
  - ✓ 如果N-P<0，则文本表示正向情感；
  - ✓ 如果N-P=0，则文本表示中性情感。
- 如果希望进一步计算文本具体的情感得分，则可以利用N和P的取值来进行计算。例如，可以定义**V=(P-N)/(P+N)，V则表示情感得分。**由公式可知，V的取值在-1~1之间，越大表示越偏正向，越小表示越偏负向。

01

中文文本词典

中文文本词典

目前中文常用的词典有知网情感分析用语词集、中文版本的Loughran and McDonald（简称LM）词典、中国金融情绪词典、中国财经媒体领域的正负面词库、金融科技领域的情感词词典等。除了使用特定的词典外，研究者也可以根据自己的研究目的自行构造词典。下面对几个常用的中文词典进行介绍。

中文文本词典

（1）中文版本的LM词典

Loughran and McDonald (2011) 从上市公司的10-K文件中人工收集并整理构造出来，他们的实证结果表明其有着良好的效果。LM词典的应用在英文词典中将会有详细的介绍。**一些中文词典研究者将LM词典直接翻译过来，并进行相关检查，构建了中文版的LM词典。**

因为该词典是直接翻译过来的，所以某些词语的情感态度在中文语境下并不是非常准确。有很多研究者还构造了其他更加适用于中文的词典，并经常有研究者拿自己构建的词典与中文版本的LM词典进行比较。

中文文本词典

（2）情感分析用词语集

中国知网（https://www.cnki.net/）于2007年发布了“情感分析用词语集”。 该文本词典包含中文和英文情感词典；中文词典和英文词典各包含6个子文件，分别为“正面情感”、“负面情感”、“正面评价”、“负面评价”、“程度级别”和“主张”六个词集。该词典的优势在于**便于获得，而且普适性较强**，但是如果需要对某一领域的词语进行分析，则该词典尚还**不够精细**。

下载地址：http://www.keenage.com/html/c_bulletin_2007.htm

中文文本词典

（3）中国金融情绪词典

Li et al. (2019) 通过下载2008年至2018年股票论坛的数据，基于传统的词典方法以及支持向量机、卷积神经网络等机器学习方法，构造了中国金融情绪词典。**相比于中文版本的LM词典，使用中国金融情绪词典进行情感分析的准确率提高了约30%。**

北京大学国家发展研究院依托该词典和中文版本的LM词典，定期发布**“中国投资者情绪指数”**。该词典的详细构造方法可以在北京大学国家发展研究院的网站获取。

网址：https://www.nsd.pku.edu.cn/xzyj/zsfb/zgtzzqxzs/250262.htm

中文文本词典

（4）其他财经类词典

汪昌云等（2015）借鉴LM字典等构造的方法，结合《现代汉语词典》等，并借鉴知网-中文信息结构库提供的正负面词汇进行匹配处理，形成了**中国财经媒体领域的正负面词库**。该研究主要是将情感词用于IPO定价的影响分析与预测，并验证了相关预测的可靠性。

王靖一等（2018）利用和讯网上的新闻，**构建了三个词典，分别为“专有名词词典”、“关键词词典”以及适用于金融科技领域的“情感词词典”**。其中，“专有名词词典”中包含了一些金融科技领域的专有名词，例如“蚂蚁金服”等，以帮助正确分词。而其构建的金融科技领域的“情感词词典”在中文金融科技领域对情感的判别有很好的效果。

中文文本词典

中国人民大学
RENMIN UNIVERSITY OF CHINA

（5）搜狗词典

搜狗细胞词库也是一个非常有力的词典工具。搜狗有着**不同领域词语的细胞词库**，可以在搜狗官网下载。例如，可以在搜狗网站下载到《上市公司新闻报道正负面词库》

该词库包含常见的新闻用词，但是词典中没有具体进行正向和负向的分类。

下载地址：https://pinyin.sogou.com/dict/detail/index/77950

中文文本词典

（6）中文情感记性词典（NTUSD）

台湾中央研究院（Academia SINICA）自然语言处理与情感分析实验室（NLPSA）基于朴素贝叶斯、支持向量机和深度学习模型，开发了中文情感记性词典（NTUSD）。该词典包含有约八千余个正面词语和约两千多个负面词语。该词典的原始版本是繁体中文，不过同样有简化版可供下载使用。**该词典将词语划分成了正向和负向两类，不过有一些用词更符合台湾的习惯。**该实验室还有很多最新的自然语言处理、情感分析成果，读者可以去官网查看，或者去下载该词典。

下载地址：http://academiasinicanlplab.github.io/

中文文本词典

（7）中文褒贬义词典

清华大学自然语言处理与社会人文计算实验室李军在该实验室从事研究时，基于其他学者的工作，开发了中文褒贬义词典，该词典中包含四千多个正面词语和五千多个负面词语，该词典集合了当时其他学者的工作，具有一定的代表性。**不过这些词典都是普适性的词典，而学者进行研究时，一般会对自己的研究领域构建一个更合适的词典。**

下载地址：http://nlp.csai.tsinghua.edu.cn/site2/index.php/13-sms

02

英文文本词典

英文文本词典

（1）英文版本的LM词典

Loughran and McDonald词典**主要用于金融和财务领域中的文本分析**，包括消极情感、积极情感、中性情感、与法律诉讼有关、情感强烈、情感较弱等七类词语。

下载地址：https://sraf.nd.edu/textual-analysis/resources/

英文文本词典

（2）The General Inquirer词典

由Philip J. Stone等（1966）发表，可以免费下载。该词典包括4206个词语，其中有包括积极情感的词语共1915个，消极情感的词语共2291个。

词典给出了每个词语的非常全面的信息，如情感方面、情感的强烈程度、是否表示主动或被动等。组织结构如下：

下载地址：http://www.wjh.harvard.edu/~inquirer

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S
1	Entry	Source	Positiv	Negativ	Pstv	Affil	Ngtv	Hostile	Strong	Power	Weak	Submit	Active	Passive	Pleasur	Pain	Feel	Arousal	EMOT
2	A	H4Lvd																	
3	ABANDON	H4Lvd		Negativ			Ngtv				Weak								
4	ABANDONMENT	H4		Negativ							Weak								
5	ABATE	H4Lvd		Negativ										Passive					
6	ABATEMENT	Lvd																	
7	ABDICATE	H4		Negativ							Weak	Submit		Passive					
8	ABHOR	H4		Negativ				Hostile						Passive				Arousal	
9	ABIDE	H4	Positiv			Affil							Active						
10	ABILITY	H4Lvd	Positiv						Strong										
11	ABJECT	H4		Negativ							Weak	Submit		Passive					
12	ABLE	H4Lvd	Positiv		Pstv				Strong										
13	ABNORMAL	H4Lvd		Negativ			Ngtv												
14	ABOARD	H4Lvd																	
15	ABOLISH	H4Lvd		Negativ			Ngtv	Hostile	Strong	Power			Active						
16	ABOLITION	Lvd																	

英文文本词典

（3）LIWC 词典

LIWC词典旨在对文本内容进行量化分析，并将导入的文本文件的不同类别的词语**（尤其是心理学类词语）**加以计算，如因果词、情绪词、认知词等心理词类在整个文本中的占比。LIWC经过十余年的发展、修改与扩充，日益稳定，目前最新的版本是由Pennebaker, J.W.等（2007）发表的LIWC2007。该词典包括2300个词语，超过70种分类，其类别体系与GI（The General Inquirer）词典基本一致，主要有以下分类：

情感方面，包括消极情感的词语如bad, weird, hate, problem, tough等，积极情感的词语如love, nice, sweet等；**认知方面**，包括表示可能性的词语如maybe, perhaps, guess等，表示拒绝和限制的词语如block, constraint等；**代词方面**，包括表示否定的代词如no, never等，表示数量的代词如few, many等。下载地址：http://www.liwc.net/

英文文本词典

(4) Bing Liu Opinion Lexicon

Bing Liu等（2004）在分析消费者评论时开发了Bing Liu Opinion Lexicon，主要用于社交媒体言论的分析。该词典包括6786个词语，其中积极情感和消极情感的词语分别有2006个和4783个。**需要特别说明的是，词典不但包含通常情况下的用词，还包含了拼写错误、语法变形、俚语以及社交媒体标记等。**

下载地址：http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

英文文本词典

（5）SentiWordNet情感词典

Stefano Baccianella等（2010）对WordNet中的词条进行情感分类，开发了SentiWordNet情感词典。该词典中，每个词语均**以百分比形式标注了积极情感、消极情感与客观性的强弱程度**。

下载地址：http://sentiwordnet.isti.cnr.it/

英文文本词典

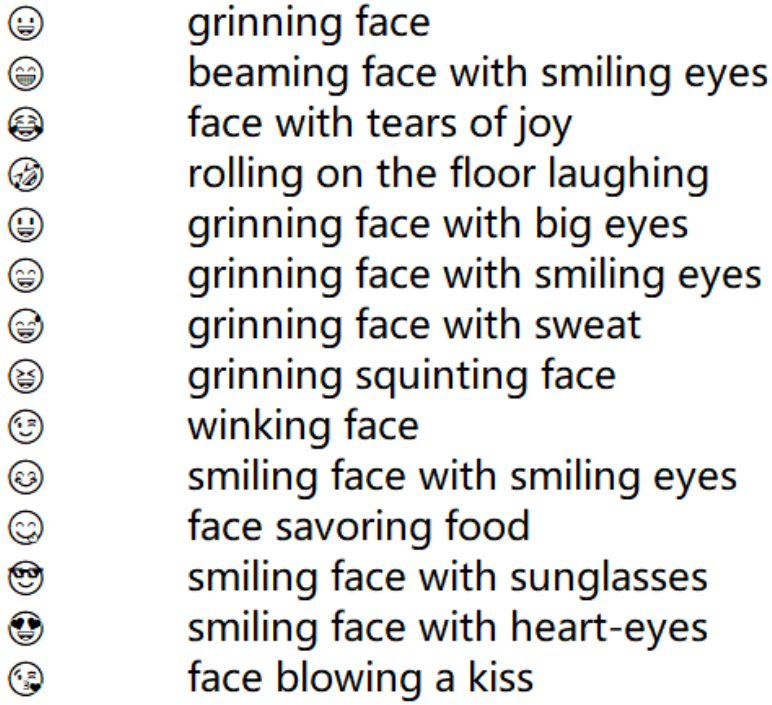
（6）Valence Aware Dictionary for Sentiment Reasoning (简称Vader)词典

Vader词典主要用于社交媒体内容的情感分析。该词典包括常用词的词典和emoji表情的词典。其中，常用词词典采用人工标注的方法为7000多个常用情感词（包括形容词，名词，副词等）进行了情感极性及强度判定，从-4到+4表示从极度负面和极度正面情感，共10人进行标注，结果位于第四列。**不同于其他情感词典，Vader词典还考虑了常用颜文字，如“:)”，以应对twitter等网络环境下非标准句子的情感判别；同时，Vader词典考虑了WTF, LOL等常用缩写词和 nah，giggly等常用俚语的情感。**

英文文本词典

(6) Valence Aware Dictionary for Sentiment Reasoning (简称Vader)词典

emoji词典则对解释了常见的3570个emoji表情的含义，组织结构如下：



下载地址：https://pypi.org/project/vaderSentiment/#files

英文文本词典

（7）SentiStrength情感词典

SentiStrength情感词典以得分的形式标注了每个词语的正面情感和负面情感的强弱程度，从-1到-5表示从不具有负面情感到极度负面的情感，从1到5表示从不具有正面情感到极度正面的情感。组织结构如下：

```
abandon*        -2      liwc uness specified otherwise
abate   -2      General Inquirer Feb 2010
abdicate*       -2      General Inquirer Feb 2010
abhor*  -4      General Inquirer Feb 2010
abject  -2      General Inquirer Feb 2010
abnormal*       -2      General Inquirer Feb 2010
abolish*        -2      General Inquirer Feb 2010
abomina*        -3      General Inquirer Feb 2010
abrasive*       -2      General Inquirer Feb 2010
abrupt  -2      General Inquirer Feb 2010
abscond*        -2      General Inquirer Feb 2010
absence -2      General Inquirer Feb 2010
absent* -2      General Inquirer Feb 2010
absurd* -2      Feb-11
```

英文文本词典

（7）SentiStrength情感词典

另外，该词典也对常见颜文字的情感进行了标注，组织结构如下：

%-(	-1
%-)	1
(-:	1
(:	1
(^ ^)	1
(^-^)	1
(^.^)	1
(^_^)	1
(o:	1
(o;	0
)-:	-1
): | -1

SentiStrength情感词典同时有免费软件版本，可以直接在相应软件中进行文本分析。

英文文本词典

中国人民大学
RENMIN UNIVERSITY OF CHINA

（8）其他软件形式的词典

主要包括**Evaluative Lexicon**和**Diction文本情感分析软件**，均为免费软件，可以直接在软件中上传需要分析的文本文件。Evaluative Lexicon软件的下载地址为http://www.evaluativelexicon.com/。Diction软件的下载地址为https://dictionsoftware.com/。

03

情感分析示例

中文情感分析示例

中国人民大學
RENMIN UNIVERSITY OF CHINA

示例1：直接通过Python包内置的词典进行情感判断

python中有一个包**SnowNLP**自带一些训练好的字典，可以用来判断语句的极性，我们通过部分爬取到的和讯网新闻的数据来进行文本极性的判断。操作代码如下：

```
1.  from snownlp import SnowNLP
2.  import csv
3.  import pandas as pd
4.  data = pd.read_csv("和讯网新闻-0.csv")
5.  n = len(data)
6.  for i in range(0, n):
7.      s = SnowNLP(data["title"][i])
8.      print("分词情况：\n",s.words)
9.      print("词性评分：\n",s.sentiments)
```

中文情感分析示例

示例2：使用已有词典进行文本极性判断

除了使用内置的Python包以外，也可以使用已有的文本词典进行文本极性的判断。下面以R代码作为实例。

首先是进行常规的数据读入、分词、停用词去除操作。

```
# 首先，我们读入数据
data.news <- read.csv("./和讯网新闻-0.csv")
data.polar <- data.frame(data.news$title)

# 与 python 的例子一样，为了简化问题
# 我们只通过对其标题（title）的情况判断该新闻的极性
# 分词章节应该介绍了停用词
stopwords <- read.csv("./hit_stopwords.txt", encoding = "UTF-8", sep = "\n", header = FALSE)
stopwords <- stopwords$V1

# 进行分词
library(jiebaR)
cutter <- worker()
data.polar$cut_result <- apply(data.polar, 1, segment, cutter)

# 去除停用词
# 从结果来看，一些词语例如“关于”、“的”等均被去除
for (i in 1:dim(data.polar)[1]) {
  if (sum(data.polar$cut_result[[i]] %in% stopwords) > 0) {
  data.polar$cut_result[[i]] <- data.polar$cut_result[[i]][-which(data.polar$cut_result[[i]] %in% stopwords)]
  }
}
```

中文情感分析示例

中国人民大学
RENMIN UNIVERSITY OF CHINA

- 导入三个常用的文本词典（NTUSD、清华大学——李军中文褒贬义词典、知网情感词典），将三个词典合并、去除重复词，共获得11935个正面词汇、15303个负面词汇。
- 导入并合并词典以后，通过判断分词结果在正向、负向词典中出现的频率。若正向词出现频率高，则判断为“正向”；若负向词出现频率高，则判断为“负向”；若出现频率相同，则判断为“中性”。

```
for (i in 1:dim(data.polar)[1]) {
  negative.count<-sum(data.polar$cut_result[[i]]%in%negative)
  positive.count<-sum(data.polar$cut_result[[i]]%in%positive)
  if (negative.count > positive.count)
    data.polar$polarity[i] <- "负向"
  if (negative.count < positive.count)
    data.polar$polarity[i] <- "正向"
  else
    data.polar$polarity[i] <- "中性"
}
```

中文情感分析示例

- 这种方法，**由于词典总量有限，且这些词典没有提供具体的正负性大小值，再加上大部分词语并不存在于已有的词典中**（实际上，真正表达情感倾向的词语在标题中也出现量有限），所以很容易出现判定为“中性”情况，**判断结果不太准确**。
- 例如在标题《反垄断成为两会热词 专家建议提高处罚力度》中，“反垄断”和“处罚”被判定为负向词，“成为”和“提高”被判断为正向词，由于正负向词语均为2，故而最终判定为“中性”，判断结果不甚理想。
- 总体来说，词典法是一种**无监督的、偏传统的情感分析方法**。现在也可以采用有监督学习方式进行情感分析的方法，通过对大量已标注数据的训练，这些有监督学习方法也可以得到更好的判断结果。

中文情感分析示例

中国人民大学
RENMIN UNIVERSITY OF CHINA

示例2：使用已有词典进行文本极性判断

为了解决上述问题，一方面**可以扩大文本数量或长度**，例如将使用标题判断改为使用正文进行判断，则可以得到更精确的结果；另一方面，使用示例1中Python的**snownlp包中的内置函数**，也可以有更为精确的判断。

总体来说，词典法是一种**无监督的、偏传统的情感分析方法**。在后面的章节还将介绍采用有监督学习方式进行情感分析的方法，通过对大量已标注数据的训练，这些有监督学习方法也可以得到更好的判断结果。

英文情感分析示例

中国人民大学
RENMIN UNIVERSITY OF CHINA

示例：使用Bing Liu Opinion Lexicon词典进行文本极性判断

- 以IMDB上爬取的电影评论为例，使用适合对社交媒体言论进行分析的Bing Liu Opinion Lexicon词典对评论进行文本极性判断。
- 由于R语言中的**qdap**包中包含了Bing Liu Opinion Lexicon，积极情感和消极情感的词语分别为positive.words和negative.words，所以这里直接调用包中的Bing Liu词典进行文本极性判断。此外，也可以从外部读入文本形式的词典，分析方法类似。

英文情感分析示例

- 首先对文本进行预处理
- 将预处理后每条评论包含的词语保存为列表形式，这里记为review_list。
- 计算出每条评论中积极情感与消极情感的词语频率，即N和P的值，并可以进一步计算得到情感得分V。
- 根据N和P值的大小判断评论的极性为“正向”、“负向”或“中性”

```
#消极词语的词频
N_review <- sapply(review_list, function(x) sum(x %in% negative.words))
#积极词语的词频
P_review <- sapply(review_list, function(x) sum(x %in% positive.words))
#计算情感得分
V_review <- (P_review-N_review)/(P_review+N_review)
polar_review <- NULL
for(i in 1:length(N_review)){
 if(N_review[i] > P_review[i]){
  polar_review[i] <- "负向"
 }
 else if(N_review[i] < P_review[i]){
  polar_review[i] <- "正向"
 }
 else{
  polar_review[i] <- "中性"
 }
}
#文本极性分析的结果，包括每条评论的积极和消极情感的词频、情感得分、文本极性
polar_review_df <- data.frame(review=review_ori, N=N_review, P=P_review, V=V_review,polar = polar_review)

#写入 csv 中
write.csv(polar_review_df, "D:/data/polar_review.csv")
```

英文情感分析示例

中国人民大学
RENMIN UNIVERSITY OF CHINA

以第八条评论为例，评论文本为：

Ant-Man is a small movie but a lot of fun. Paul Rudd is perfect as Ant-Man. The heist story is very good and this movie is very funny. The side characters are great and a lot of fun. The villain is bland and boring. Other than that this is movie and one of the most rewatchable films in the MCU.8.2/10

分析得到，**该评论包含3个消极情感的词语，6个积极情感的词语，因此情感倾向为“正向”，且计算得到情感得分约为0.33。**

英文情感分析示例

可以看到，如果词典只指出了词语的情感倾向为积极或者消极，则我们默认每个词语的情感的强烈程度是相等的，即每个词语在计算情感得分时的权重相等。**因此，这样计算情感得分虽然较为简便，但是存在不合理性。**对于同时存在积极情感词汇和消极情感词汇的文本（例如“The acting was good , but the movie could have been better!”），这一方法可能无法正确判断文本的情感倾向。

3.2、机器学习法

有监督学习

情感分析属于有监督学习。有监督学习任务中既可以观察到数据的标签信息（即因变量Y），也可以观察到用于解释标签信息的其他数据（即解释变量X），有监督学习的实质是通过X对Y进行建模，寻找X和Y之间的关系，从而实现通过X预测Y的目的。

在对文本数据进行分类时，通常会涉及以下几个步骤：

- 文本预处理
- 文本特征提取
- 分类模型构建

数据拆分

- 在建立分类模型之前，通常需要对全部数据进行拆分，比如随机抽取里面的p%数据作为训练数据集，剩下1-p%的数据作为测试数据集。训练数据集通常用于进行模型训练，测试数据集用来测试模型效果。

- 除此之外，为了在建模时可以利用到全部数据集，可以使用交叉验证的方式。比如，五折交叉验证通常是指将全部数据随机拆分为数据量相等的5份，然后每次选择其中一份作为测试数据集，剩下四份作为训练数据集。按照这种方式，所有数据将轮流作为测试数据集。完成数据的划分之后，就可以对训练数据建立分类模型了。

01

朴素贝叶斯

朴素贝叶斯

中国人民大学
RENMIN UNIVERSITY OF CHINA

- 朴素贝叶斯分类器是基于贝叶斯理论进行分类的，主要用到了**贝叶斯定理和特征条件独立性假设**。特征条件独立性假设是指，对某个已知类别，假设所有特征都相互独立，即每个特征独立的对分类结果产生影响。
- 贝叶斯分类器的核心思想是贝叶斯定理，它通过贝叶斯定理计算待分类对象属于第k个类别的后验概率，以此来表示其分到第k个类别的可能性，然后选择使其具有最大后验概率的类别作为最后的分类类别。

朴素贝叶斯

定义$\mathrm{x} = (x_1, \ldots, x_V)'$为某个文本特征向量，其中V为筛选后的特征词个数。

假设一共有K个类别，对于$1 \le \mathrm{k} \le \mathrm{K}$，定义$y_k = 1$表示x属于第k类，$y_k = 0$表示x不属于第k类。之前提到在贝叶斯分类器中，假设各个特征对分类的影响是相互独立的，因此在给定类别y_k下，x出现的概率可以表示为：

$$\mathrm{P}(\mathrm{x}|y_k) = \prod_{i=1}^{V} P(x_i|y_k) = P(x_1|y_k) \times \cdots \times P(x_V|y_k)$$

值得注意的是，在拿到训练集数据后，各个文档的标签是已知的，因此可以比较方便的计算当前各个类别下文档出现的概率。

朴素贝叶斯

但是如果想知道x所属的类别，其实质应该是计算$\mathrm{P}(y_k|\mathrm{x})$，即在给定x的条件下各个可能类别$y_k$的出现概率。根据贝叶斯定理可知$\mathrm{P}(y_k|\mathrm{x})$的计算公式如下：

$$\mathrm{P}(y_k|\mathrm{x}) = \mathrm{P}(\mathrm{x}|y_k)\mathrm{P}(y_k)/\mathrm{P}(\mathrm{x}) \propto \mathrm{P}(\mathrm{x}|y_k)\mathrm{P}(y_k)$$

其中，$\mathrm{P}(y_k)$表示类别k出现的概率，通常可以用其在训练集中的频率表示，$\mathrm{P}(\mathrm{x})$表示不考虑任何类别时文本x出现的概率。对于给定文本x，$\mathrm{P}(\mathrm{x})$会出现在$\mathrm{P}(y_1|\mathrm{x}),\ldots\mathrm{P}(y_K|\mathrm{x})$的计算中，因此实质上并不对类别选取产生影响。

最后，选取使得$\mathrm{P}(y_k|\mathrm{x})$取值最大的k值，即是x对应的类别。

01

逻辑回归模型

线性回归模型

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

此时Y是连续型变量，当Y变为分类型变量时，怎么办？

线性回归模型

如果用线性回归：

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon = X'\beta + \varepsilon$$

产生的后果：

- 右边→连续型，左边→离散型，几乎永远不可能相等
- 用线性回归也可以进行估计，但是估计出的Y不是0或1

线性回归模型

是否可以把0,1型的Y变为连续型呢?

Y=0 or 1

离散型 →连续型

Y的**发生概率**
p ∈ [0,1]

X与p的关系

$p = X'\beta$

对p进行单调变换

可能性指标：odds ratio
p/(1-p) ∈ [0, +∞])

对数变化

Log-odds
Log{p/(1-p)} ∈ [−∞, +∞]

$\log\{\frac{p}{1-p}\} = X'\beta$

逻辑回归模型

1. 对Y=1的概率进行建模，即：p=P(Y=1)=E(Y)
2. 对p进行logit变换，即：$Z = \log\{\frac{p}{1-p}\}$
3. 对Z建立线性回归模型，即$Z = X'\beta$

等价形式：$$\mathrm{P}(Y = 1) = p = \frac{\exp(Z)}{1 + \exp(Z)} = \frac{\exp(X'\beta)}{1 + \exp(X'\beta)}$$

逻辑回归模型

☆R中的广义线性回归语句*glm*

☆语法为：*glm(formula, **family = gaussian**, data, weights, subset, na.action, start = NULL, etastart, mustart, offset, control = glm.control(...), model = TRUE, method = "glm.fit", x = FALSE, y = TRUE, contrasts = NULL, ...)*

☆与lm不同之处就在于参数***family***

☆这个参数的作用在于定义一个族以及连接函数，使用该连接函数将因变量的期望与自变量联系起来

☆family= binomial(link=logit)表示引用了二项分布族binomial中的logit连接函数

02

决策树

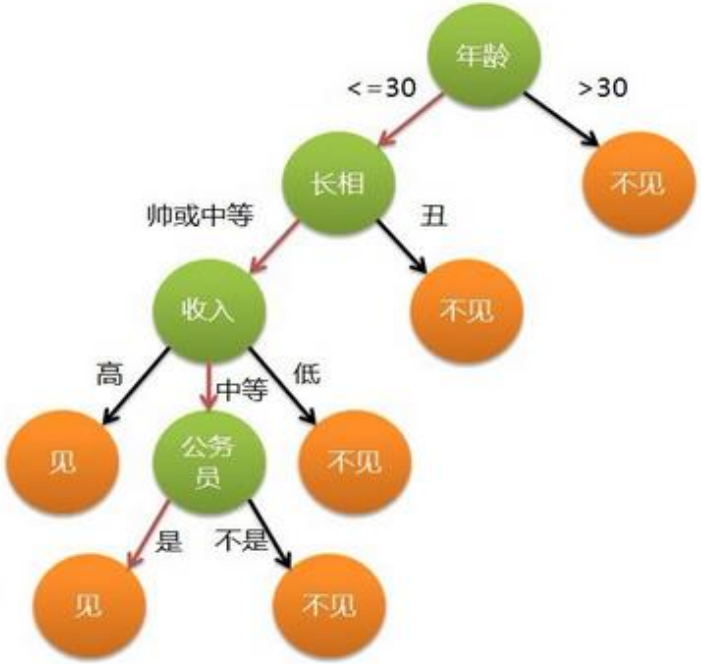
决策树

- 决策树是一种根据自变量的值进行递归划分以预测因变量的方法。决策树通常有两类：当因变量为分类变量时，称相应的决策树为**分类树**；若因变量为连续变量时，称相应的决策树为**回归树**。
- 决策树的整个决策过程形如一棵倒立的树。该树的根节点包含整个训练数据集，从根节点出发进行数据分割，每次分割对应一个节点。在每个节点上，决策树会选择对当前数据集分类效果最好的自变量，然后将该节点分裂成两个或多个子节点，最后直到分类完毕或所有自变量都已被使用，最后一层节点被称为叶节点。

决策树

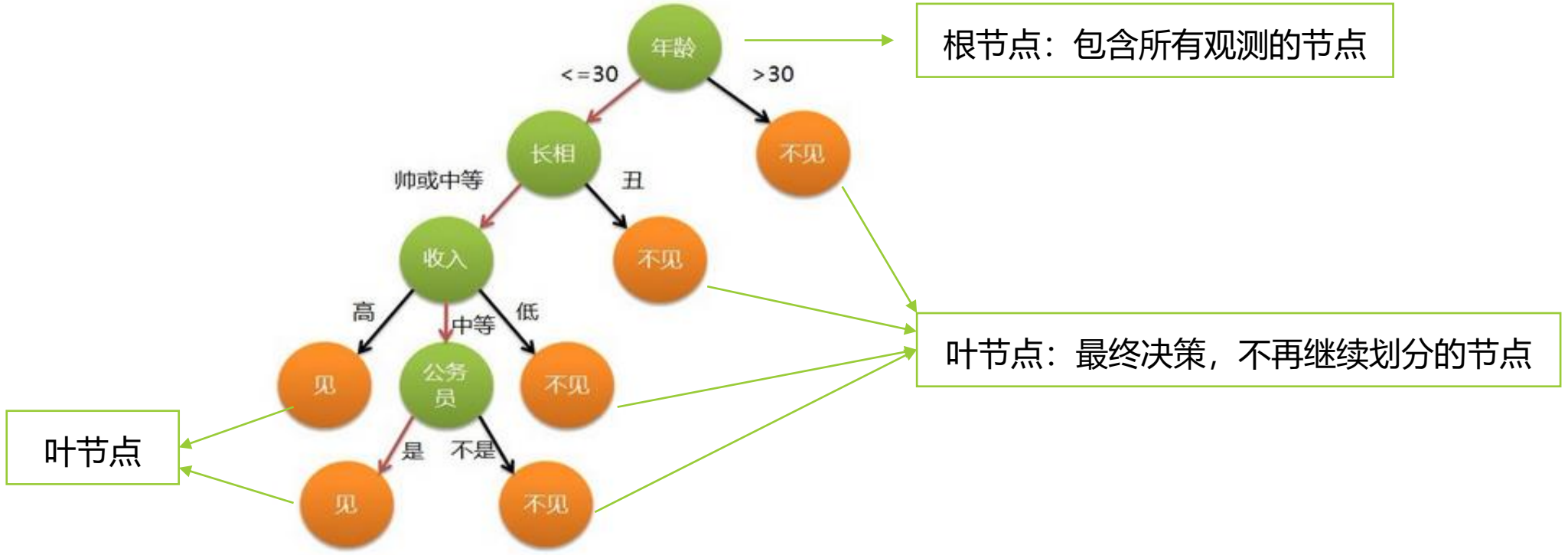
决策树分类的思想类似于我们做决策的过程。

女儿：多大年纪了？ (年龄)
母亲：26。
女儿：长的帅不帅？ (长相)
母亲：挺帅的。
女儿：收入高不？ (收入情况)
母亲：不算很高，中等情况。
女儿：是公务员不？ (是否公务员)
母亲：是，在税务局上班呢。
女儿：那好，我去见见。



决策树

整个决策过程形如一棵**倒生长的树**，因此叫决策**树**。



决策树

- 使用决策树进行决策的过程就是从根节点开始，测试某个特征属性，并按照其取值选择输出分支，直到到达叶子节点，将叶子节点存放的类别作为决策结果。

- 这种逐级判断的流程在现实业务中很常见。如医生阅读检验报告时通常先看某个指标是否超出某个范围，然后再看其他指标。

- 早期的人工智能领域中，有一类专家系统就是这个思路：由人类专家来制定规则并通过算法构造复杂的树，帮助人们来进行决策。

决策树

- 在决策树的生长过程中，需要解决这几个问题：
  - 选择哪个特征属性进行划分？年龄？长相？收入？
  - 在选定的特征属性上如何进行分割？划分标准是什么？

- 归结于一个问题：如何评估每次划分的效果

决策树的生长

- 每次划分的目标：使得同一个分支里的样本有尽可能一致的类别标签，即尽可能“纯”

- 纯度的度量通常有两个指标：
  - 熵
  - 基尼系数

熵

- 设数据集为D，|D|为样本个数。
- 设有K个分类Ck， k=1，2...K，|Ck|表示属于Ck分类的样本个数，因此有：

$$\sum_{k}|C_k| = |D|$$

- 计算数据集D的经验熵为：

$$H(D) = -\sum_{k=1}^{K}\frac{|C_k|}{|D|}\log\frac{|C_k|}{|D|}$$

- 熵的取值越小，表明节点越纯

基尼系数

- 基尼系数的定义：

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^2$$

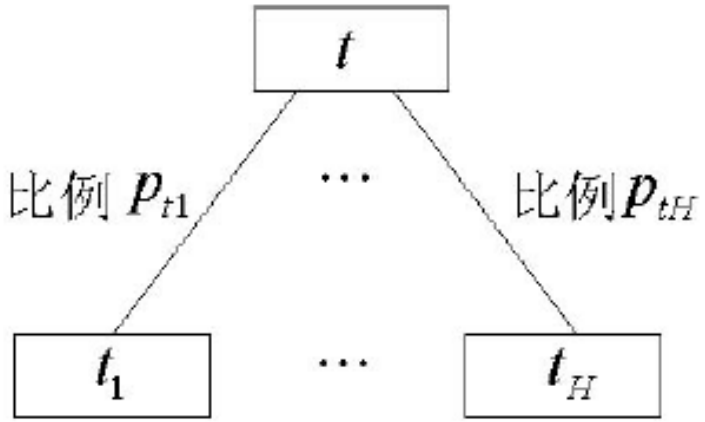
$$=1-\sum_{k=1}^{K}\left(\frac{|C_k|}{|D|}\right)^2$$

- 基尼系数可以看成熵的一半。基尼系数越小，节点越纯

如何在某个变量进行划分?

- 对于某个自变量X，构造其所有可能的划分集合
  - 当X是连续变量，将其在数据中的所有取值从小到大排，然后进行**分箱**处理
  - 当X是离散型，按照其不同水平的取值划分

- 对任意一种划分方式，计算：
  - 划分前节点的纯净度
  - 划分后几个分支的平均纯净度
  - 纯净度改变的大小
- 选择纯净度提升最大的划分作为最优划分



叶节点的确定

- 随着划分的持续进行，树持续生长，直到下列情况发生其一，则停止生长，使目前的节点作为叶节点

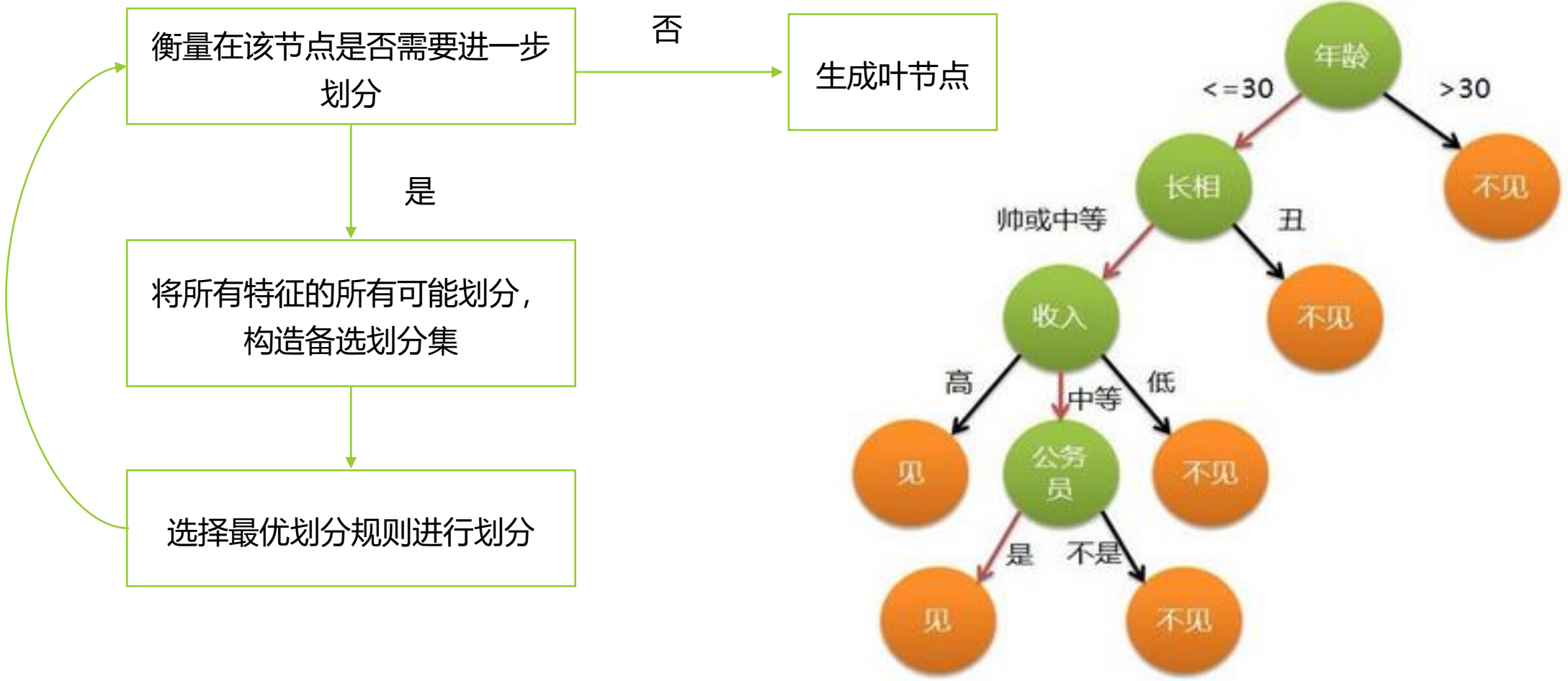
  - 节点内训练数据的观测数达到某个最小值
  - 树的深度达到一定限制
  - 没有哪种划分准则可以使纯净度显著提升

决策树的生长过程

- 将所有变量的所有可能划分进行汇总，得到备选划分集合

- 在所有备选集合中，选择纯净度提升最大的划分作为最优划分

生长过程示例


衡量在该节点是否需要进一步划分
否
生成叶节点
是
将所有特征的所有可能划分，构造备选划分集
选择最优划分规则进行划分
年龄
<=30
>30
长相
不见
帅或中等
丑
收入
不见
高
中等
低
见
公务员
不见
是
不是
见
不见

决策树的修剪

中国人民大学
RENMIN UNIVERSITY OF CHINA

- 决策树是根据训练数据集生长而成，叶节点越多，对训练数据的拟合效果越好，但叶节点过多会将训练数据集的噪音也学习进来，从而造成过拟合的现象
  - 想象每个数据点作为一个叶节点

- 因此需要使用修正数据集对决策树进行剪枝。根据各个子树在修正数据集的预测效果来选择最优的子树

随机森林

中国人民大学
RENMIN UNIVERSITY OF CHINA

- 随机森林是Breiman提出的算法，该方法是Bagging算法的一个变体，是一种基于决策树的组合分类器算法。

- 随时森林的生成过程：
  1. 样本的随机：从样本集中用Bootstrap随机选取n个样本
  2. 特征的随机：从所有特征中随机选取K个特征，在这K个特征中选择最佳分隔特征及其划分规则，建立决策树
  3. 重复以上两步m次，即建立了m棵决策树
  4. 这m个决策树形成随机森林，通过投票表决结果，决定数据属于哪一类（投票机制有一票否决制、少数服从多数、加权多数）

随机森林

- 随机森林算法在决策树的训练过程中引入了**随机特征选择**：传统决策树在选择划分特征时是在当前节点特征集合中选一个最优集合，而随机森林则是先使用Bagging方法生成不同训练集，然后基于决策树的每个节点，在该节点的特征集合中得到一个包含数个特征的子集，再从子集中选择最优特征用于划分。
- 同时，每棵决策树的训练样本由**随机采样**获得，且生成决策树时各节点分裂所选择的特征也是随机的。通过两种随机性的结合，可以降低不同决策树的相似性，从而使得该算法面对噪声时的鲁棒性更好，对非平衡数据处理得到的结果也更加稳健，并进一步提升随机森林的分类精度。

03

支持向量机

支持向量机

支持向量机（SVM，Support Vector Machine）是基于VC维（Vapnik-Chervonenkis dimension）和结构风险最小化等统计学习理论提出的机器学习算法。支持向量机在最开始提出时只是用于解决二分类问题，但是实际应用中，支持向量机也可以用来解决多分类问题。在多分类场合下，支持向量机会将所有K个类别的分类问题转化为一系列的二分类问题进行处理。同时，支持向量机也可以支持因变量为连续型变量的情形。

从一个故事开始

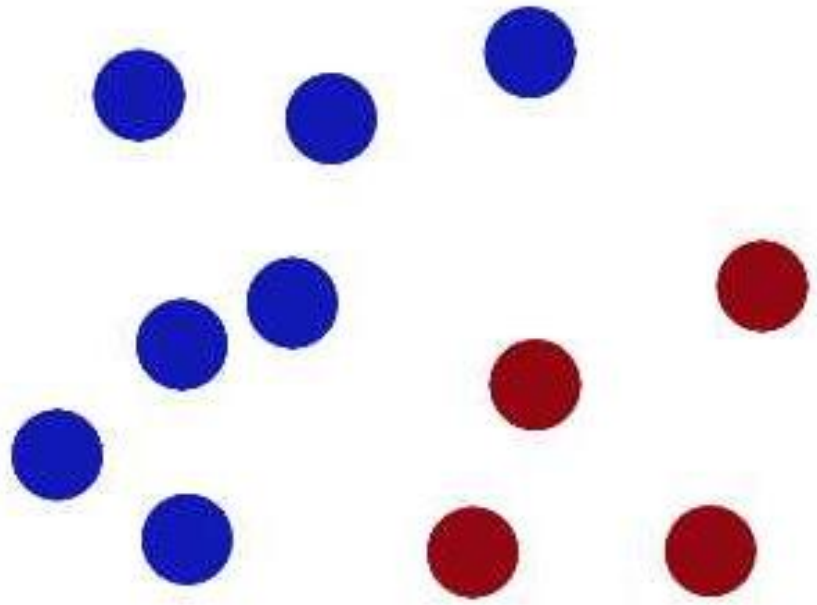
在很久以前的情人节

大侠要去救他的公主

看守公主的魔鬼和大侠玩了一个游戏

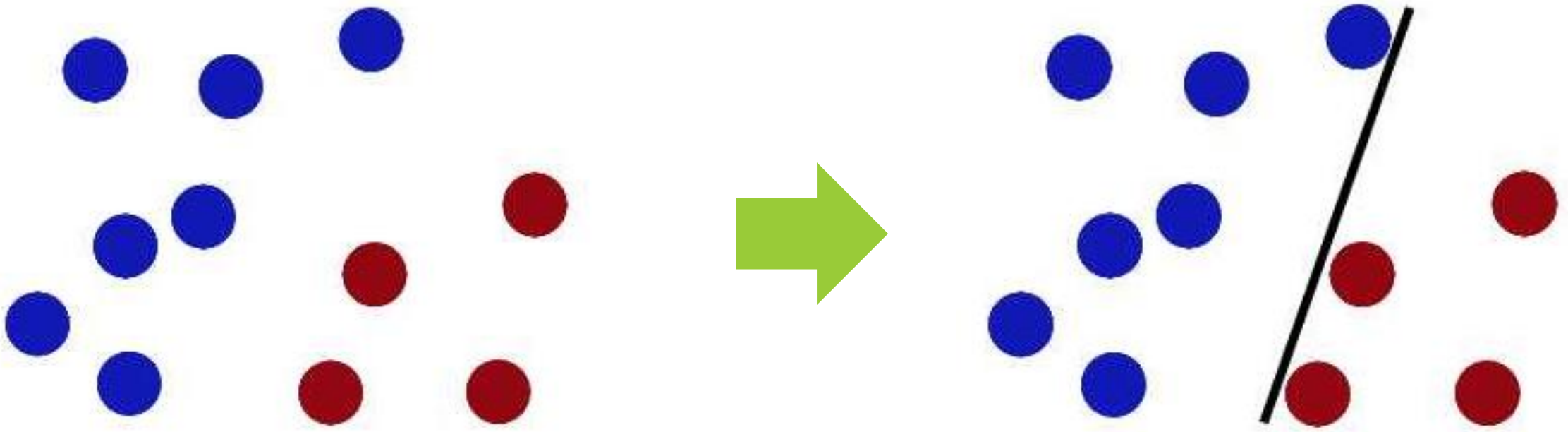
魔鬼在桌子上似乎有规律放了两种颜色的球，说：

“你用一根棍分开它们？要求：尽量在放更多球之后，仍然适用。”



从一个故事开始

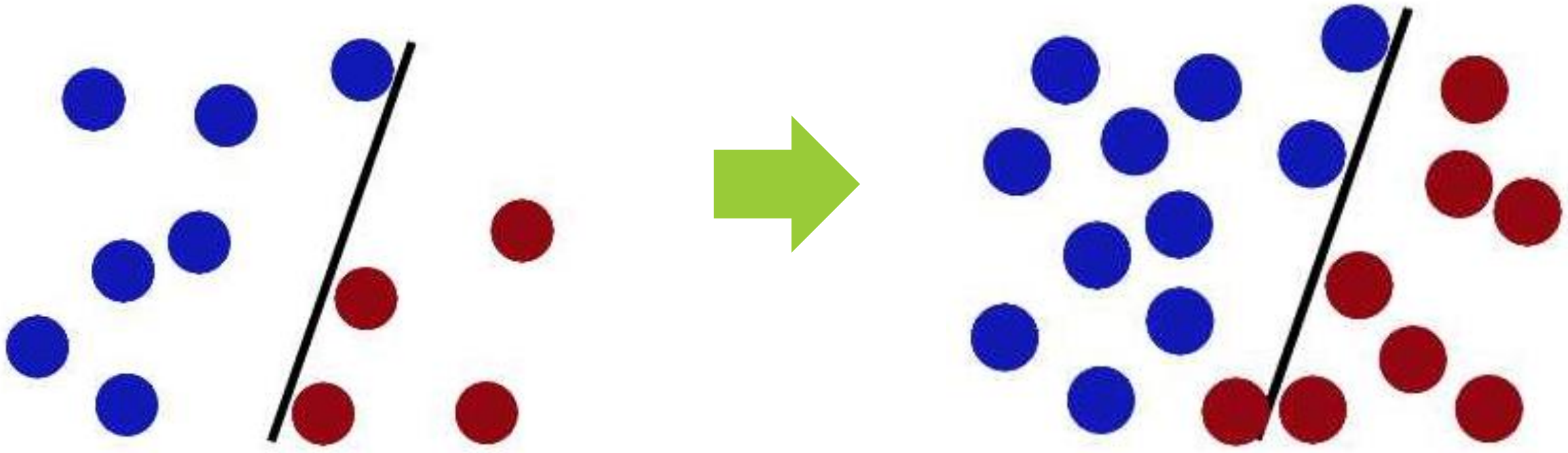
大侠进行了如下分隔，似乎干的不错



从一个故事开始

然而，魔鬼又在桌上放了更多的球

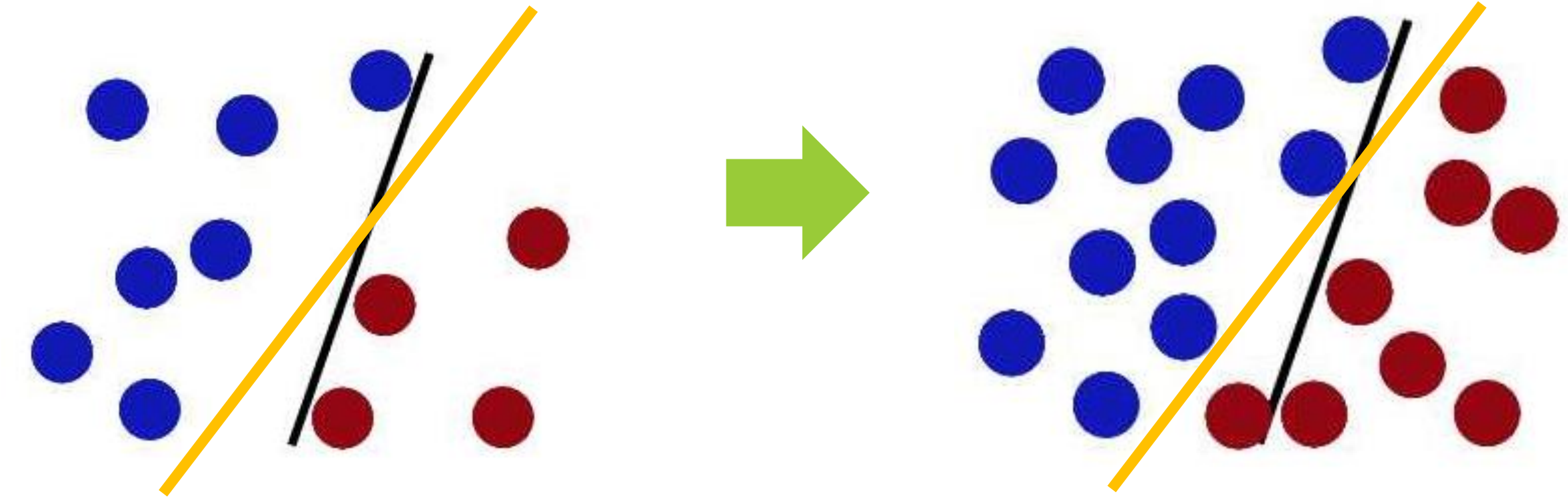
有一个球站错了阵营。。。



从一个故事开始

如果，大侠给出的是另一种划分

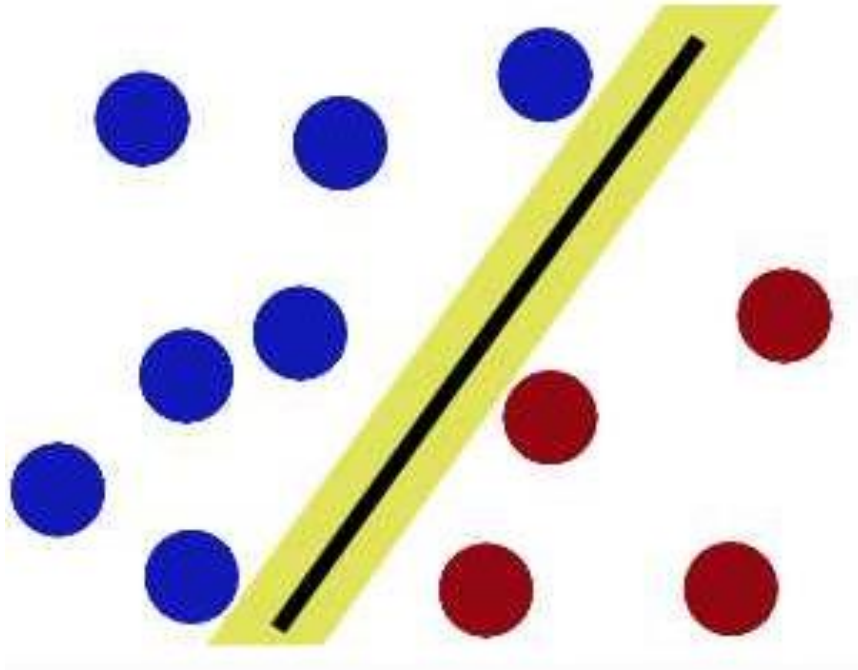
似乎就不会被魔鬼难倒了。。。



从一个故事开始

如何选择划分方式，才能在两边放置更高的球之后，仍然能很好的分开？

划分的最佳位置：划分后，直线两边有尽可能大的间隙，从而最大限度的容纳新的球

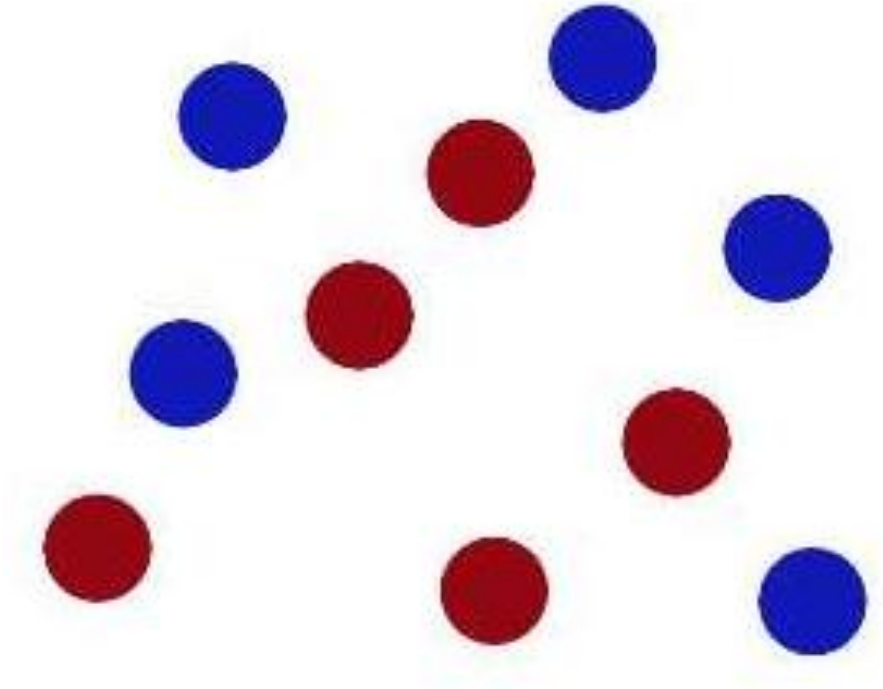


从一个故事开始

眼看大侠已经完美的完成的任务

魔鬼很生气，后果很严重

愤怒的魔鬼打乱了桌子上的球。。。



从一个故事开始

大侠没有直线可以很好帮他分开两种球了

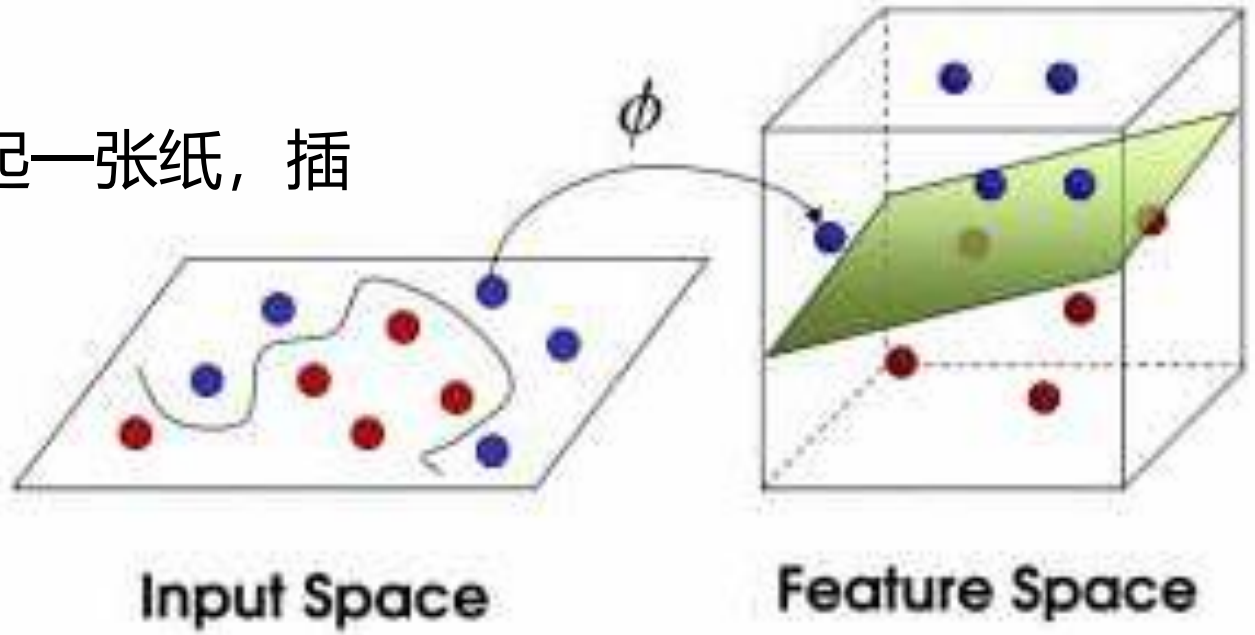
怎么办呢?

大侠不愧武功盖世

只见大侠一拍桌子，球飞到空中

然后，凭借大侠的轻功，只见他抓起一张纸，插到了两种球的中间

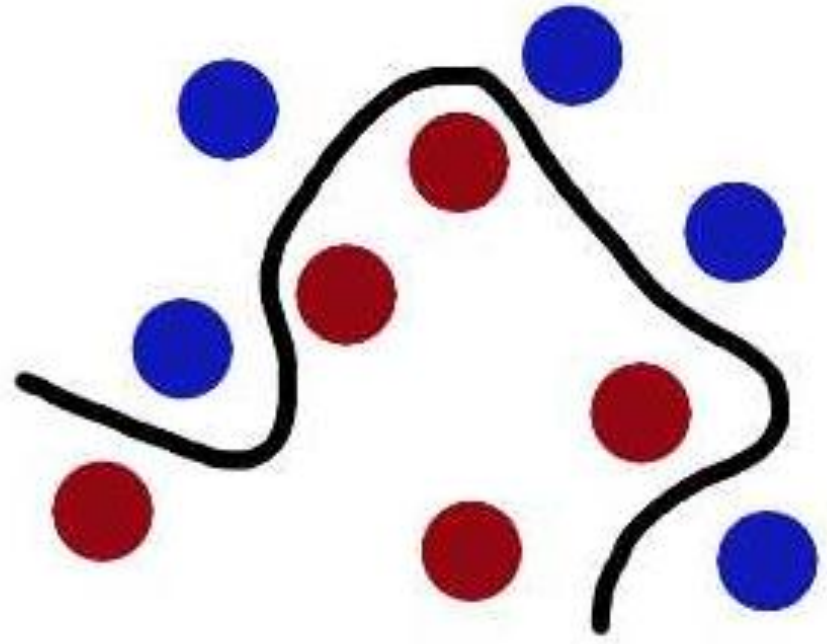
于是，两种球神奇的分开了



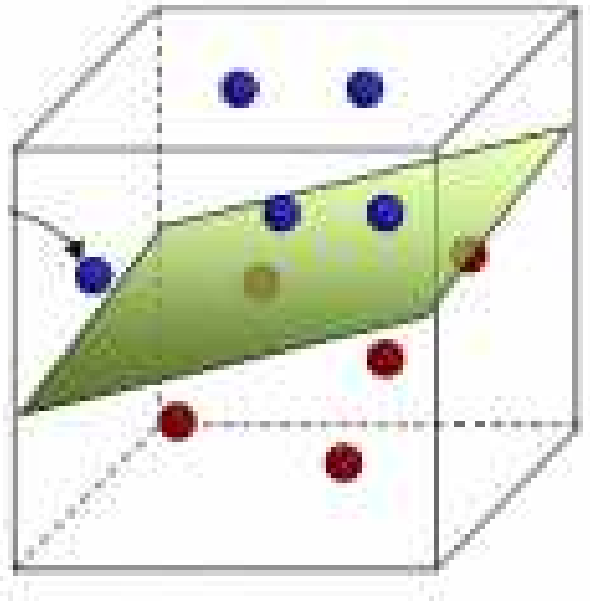
从一个故事开始

中国人民大学
RENMIN UNIVERSITY OF CHINA

从魔鬼的角度来看，这些球是被一条曲线分开的



从大侠的角度看，这些球是被某个平面分开的



从一个故事开始

再之后，无聊的大人们

把这些球叫做 「data」

把分开球的直线叫做 「classifier」

得到最大间隙trick叫做「optimization」

拍桌子叫做「kernelling」

那张纸叫做「hyperplane」

故事来源：https://www.zhihu.com/question/21094489

支持向量机

支持向量机（Support Vector Machine, SVM）

对于线性可分问题，SVM就是一条直线，可以将两类物体**线性分开**

但它又不是一条普通的直线，它是无数条可以分类的直线当中**最完美的**

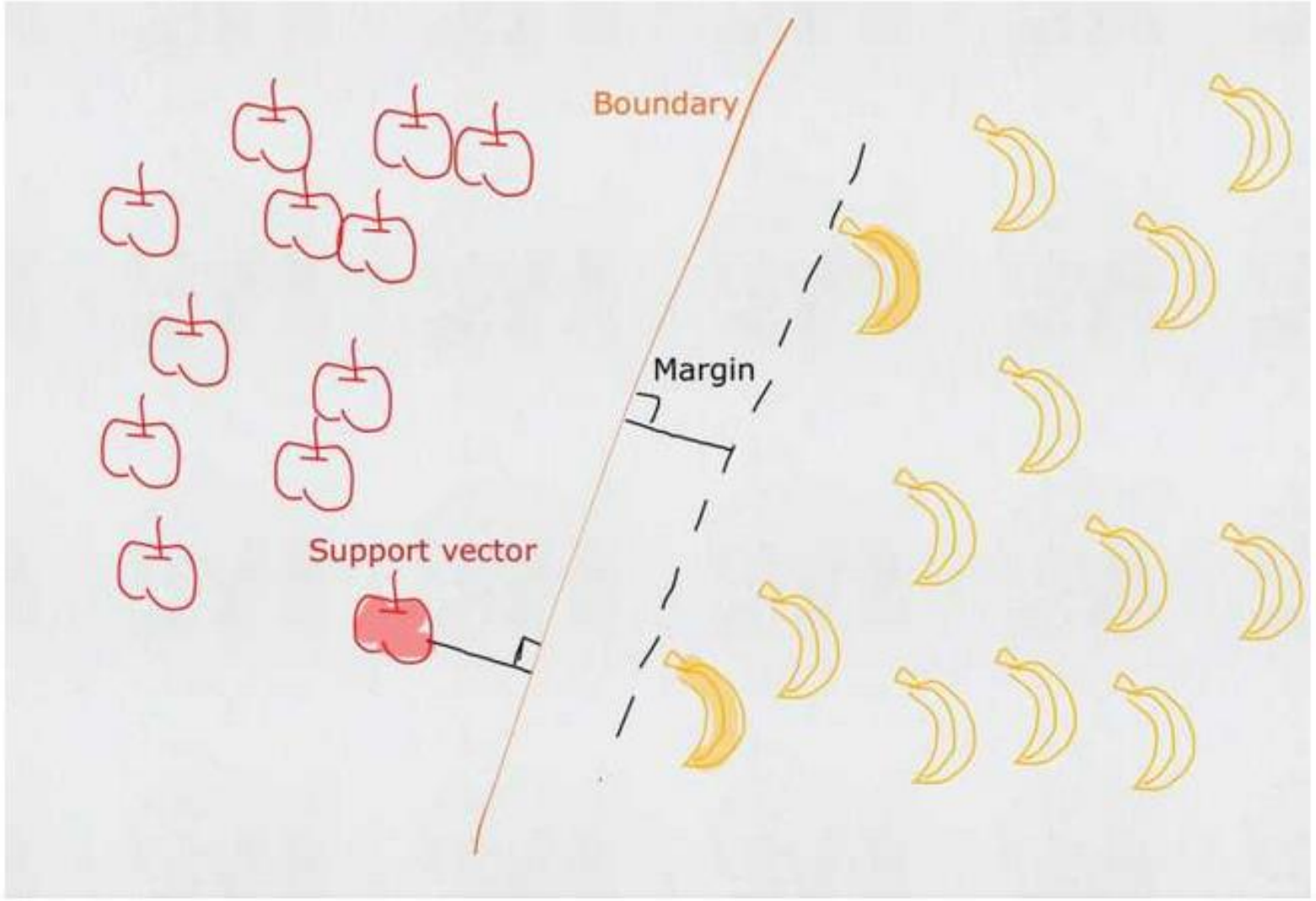
因为它恰好在两个类的中间，距离两个类的点都一样远

支持向量机

Support vector就是这些离分界线最近的『点』

Boundary表示划分的直线

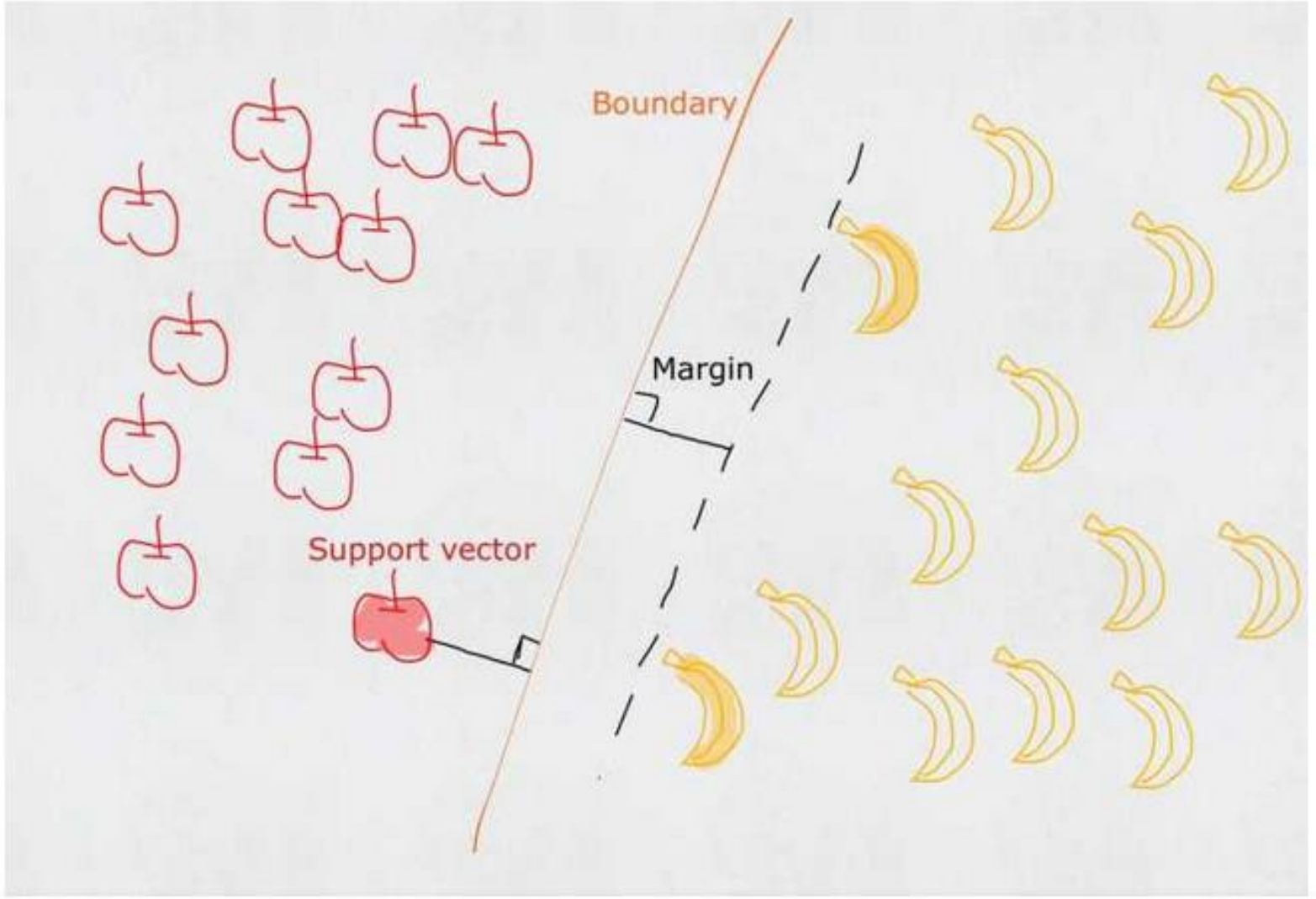
Margin表示最近的点离boundary的距离



支持向量机

支持向量机的目标：

选择一条最优的boundary，使得两边的margin到boundary的距离最大



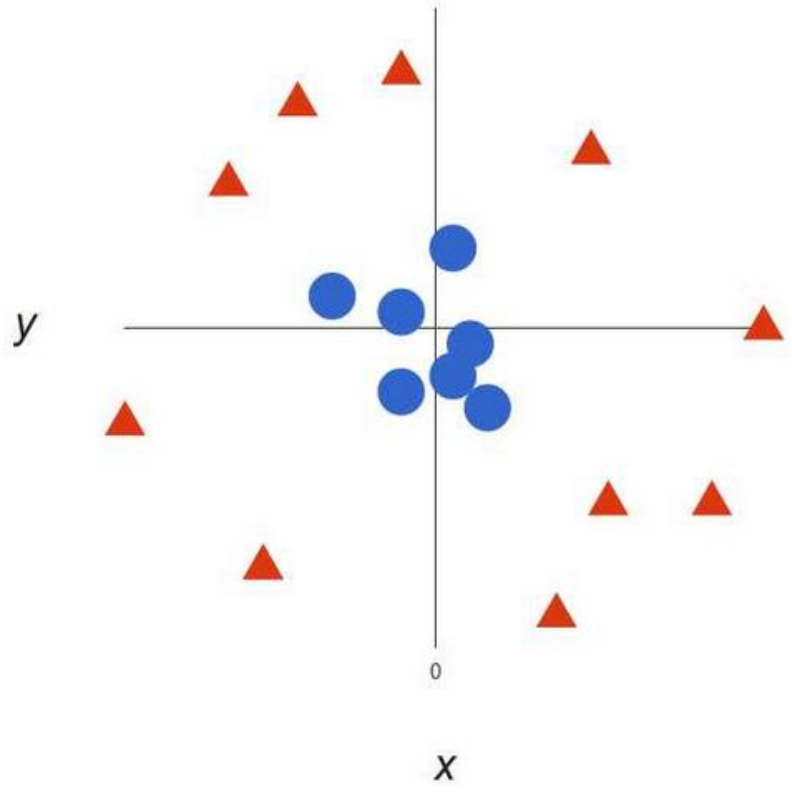
支持向量机

如果苹果和香蕉不是线性可分的呢?

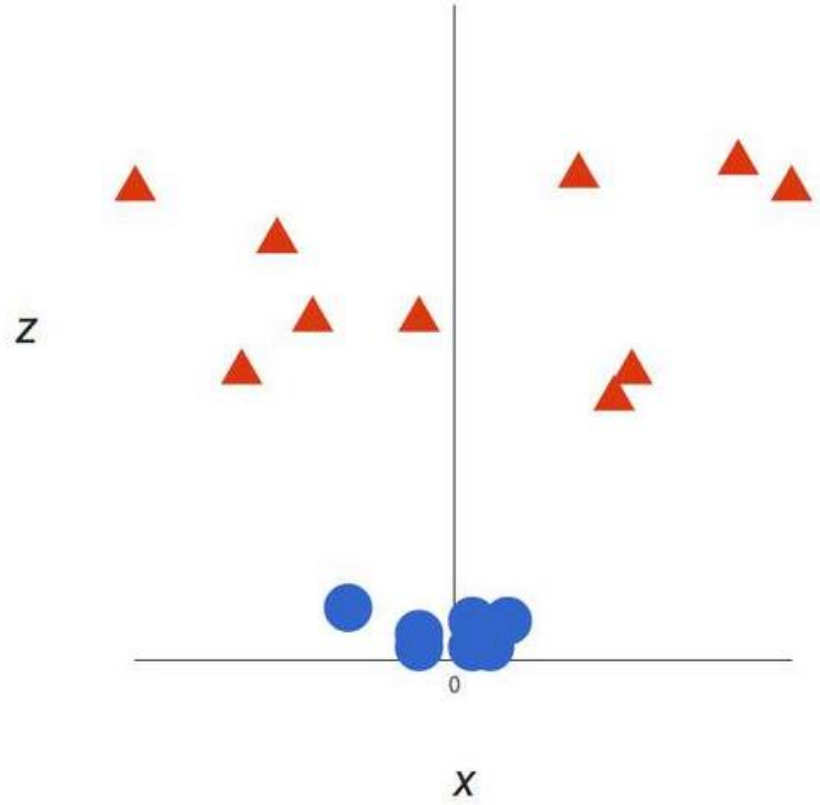
使用核函数（kernel），把这些点映射到更高维的空间中

支持向量机

使用核函数（kernel），把这些点映射到更高维的空间中



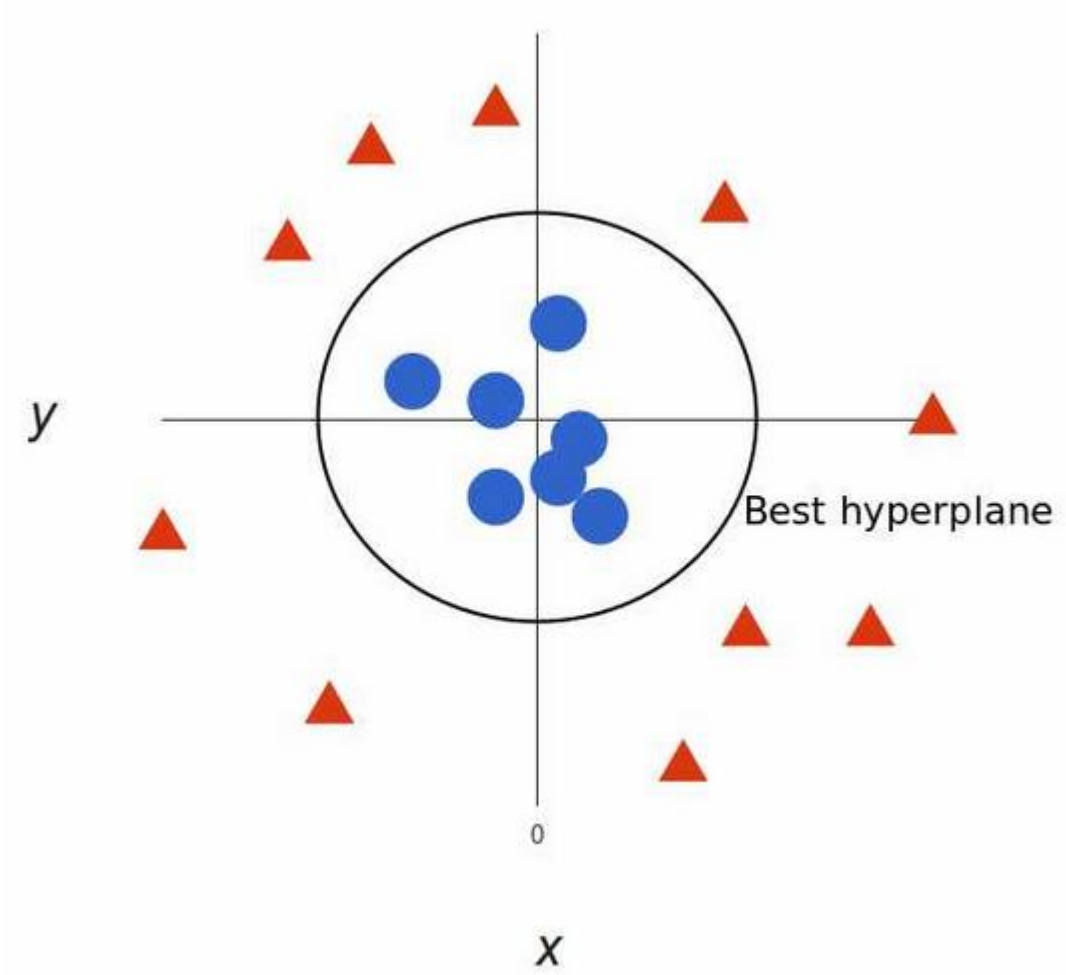
引入第三个维度
令z=x*x+y*y



支持向量机

F(x,y)=x*x+y*y就是一个核函数

使用核函数的方式，通过非线性映射将样本映射到高维特征空间中，从而实现在高维空间中线性可分的目的。它具有能够避免过拟合、保证局部最优解为全局最优解和较好的泛化性等优点。

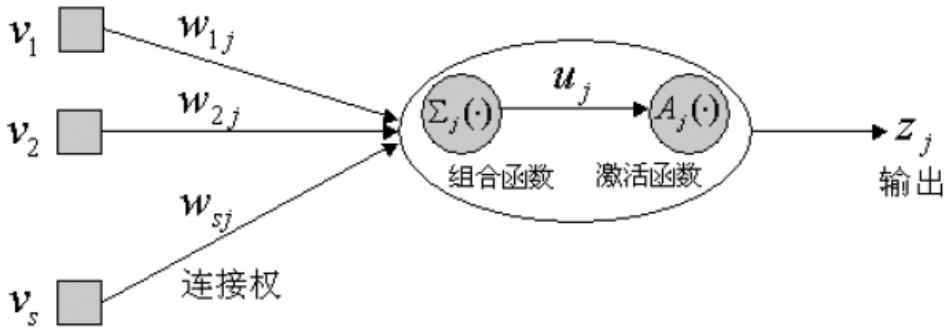


04

神经网络

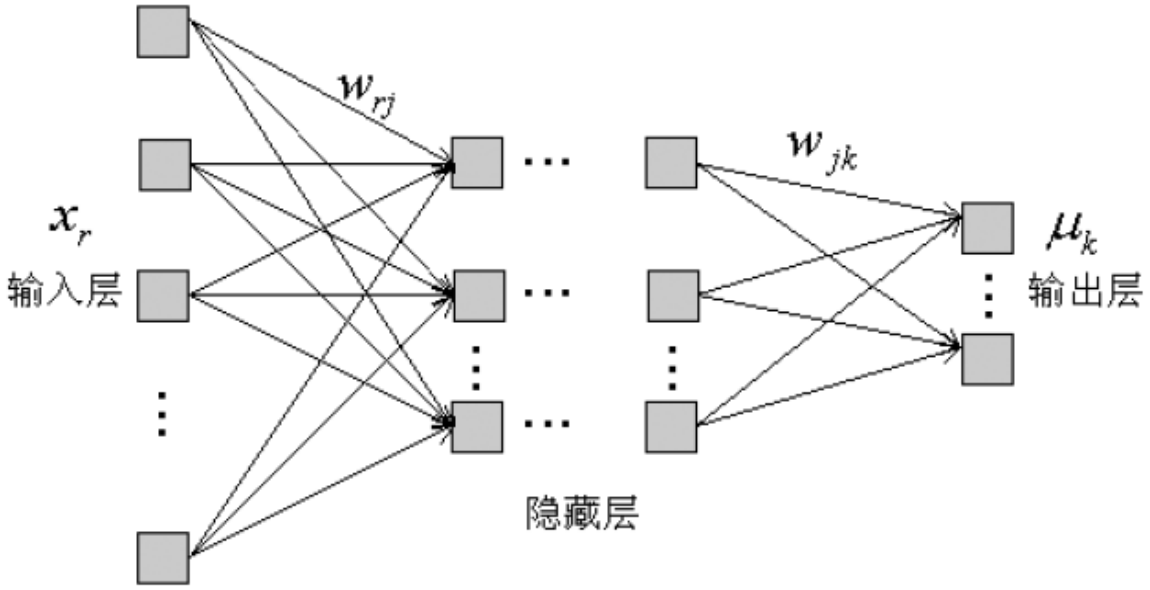
神经网络

- 神经网络模型是模拟大脑中神经元结构和信息传递方式而产生的一类模型。
- 一个典型的人工神经元的结构如下图所示：$v_1, \ldots, v_S$表示输入信号，每个输入信号都对应一个连接权重$w_{1j}, \ldots, w_{Sj}$。这些输入信号叠加相应的权重之后，通过神经元内的组合函数进行汇总，定义汇总后的信号为u_j，该信号会进一步通过神经元内的激活函数进行处理，最后输出结果z_j。信号z_j则会进一步传送给其他的神经元进行后续处理。



神经网络

- 多个人工神经元连接在一起就形成了神经网络。最常用的神经网络是多层感知机模型（形式见右图）。
- 在多层感知机模型中，第一层表示输入层，即代表信息的输入；最后一层是输出层，代表信息的输出；中间各层是隐藏层，由各个神经元组成，是对输入信息的逐层处理。



神经网络

中国人民大学
RENMIN UNIVERSITY OF CHINA

- 具体来说，训练集中的数据信息会通过输入层的神经元输入到网络，输入层的各个神经元和第一个隐藏层的各个神经元连接，每一层隐藏层的各个神经元和下一层（可能是隐藏层或输出层）的各个神经元相连接。最后，输入的数据信息会通过各个隐藏层的神经元进行转换后，在输出层形成输出值作为对因变量的预测值。
- 通过隐藏层的不断叠加，神经网络实际上是构建了一个高度非线性的模型结构来拟合自变量和因变量之间的关系，其复杂的模型形式和较低的可解释性也是神经网络经常被诟病的一点。

神经网络

- **在神经网络模型中，因变量既可以是连续型变量，也可以是分类型变量。**同时，神经网络可以对一个因变量或者多个因变量同时进行建模。模型会根据因变量的具体形式来定义目标函数，并通过调整连接权重等参数来最小化目标函数进行求解。
- 神经网络模型的历史最早可以追溯到1943年。近十年来对神经网络的讨论又达到了高潮，这其实得益于深度学习模型取得的巨大成功。从本质上来看，深度学习模型其实就是神经网络模型，但是是层数更多、深度更大的神经网络，并且在深度学习模型的发展中也产生了很多新的技术和方法。

05

模型评价

混淆矩阵

- 混淆矩阵能够比较全面地反映模型的性能，很多衡量分类准确率的指标都可以通过混淆矩阵衍生而来。

- 以二分类问题为例，记因变量的两个类别标签为“正”和“负”，因此测试数据集的预测结果可以被划分为4类：真正类（真实类别为正，预测类别为正），真负类（真实类别为负，预测类别为负），假正类（真实类别为正，预测类别为负），假负类（真实类别为负，预测类别为正）。

混淆矩阵

真实类别	预测类别	
	正例	负例
正例	TP(真正类)	FN(假负类)
负例	FP(假正类)	TN(真负类)

TP（true positive）表示真实类别标签和预测类别标签均为正的样本数

FP（false positive）表示真实类别标签为负但是预测类别标签均为正的样本数

FN（false negative）表示真实类别标签为正但是预测类别标签均为负的样本数

TN（true negative）表示真实类别标签和预测类别标签均为负的样本数

混淆矩阵

从混淆矩阵中可以计算得到如下各个评价指标：

$$准确率：Accuracy = \frac{TP+TN}{TP+TN+FP+FN};$$

$$精确率：Precision = \frac{TP}{TP+FP};$$

$$召回率：Recall = \frac{TP}{TP+FN};$$

$$F1值：F1\ score = \frac{2\times Recall\times Precision}{Recall+Precision};$$

在上述指标中，准确率表示测试集所有样本被正确预测的比例，精确率表示模型预测为正的样本中有多大比例是正确的，召回率表示测试集所有为正的样本中有多大比例可以被模型正确预测，最后F1值是精确率和召回率的综合。

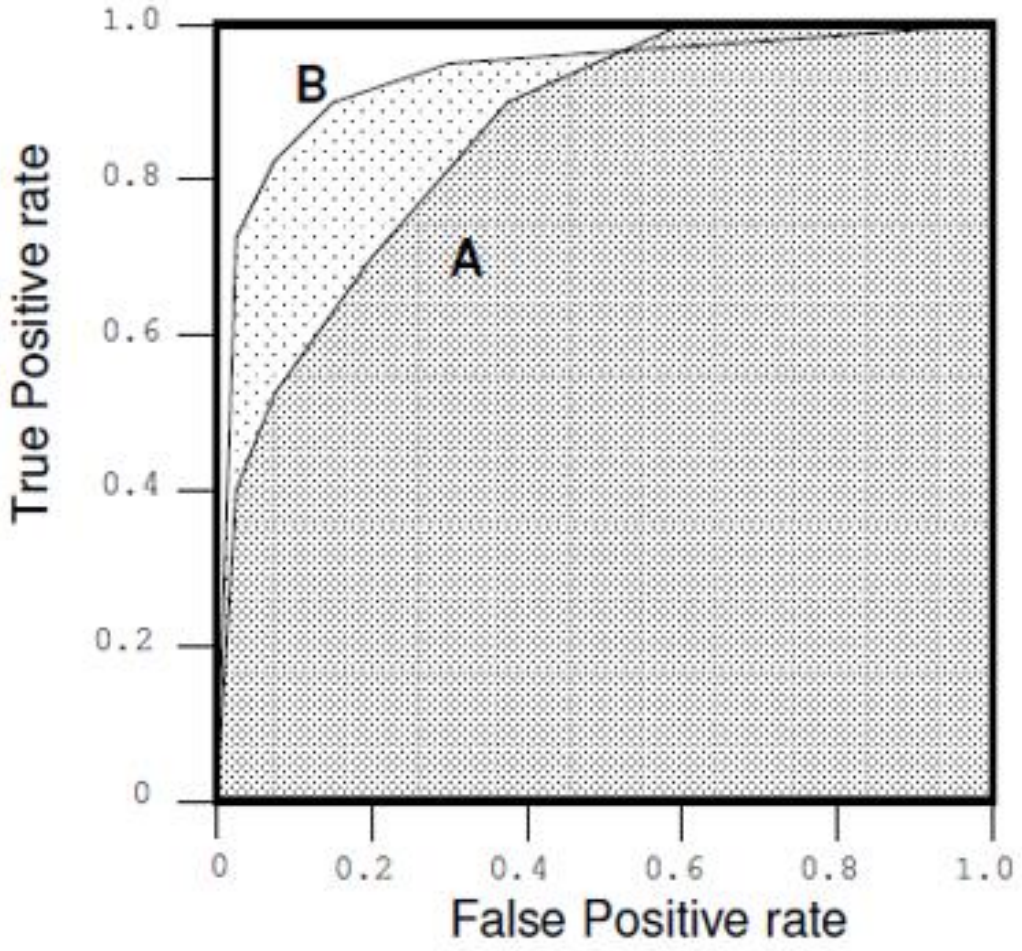
ROC曲线

- 在二分类问题中，在使用模型进行预测时，实际上得到的是样本被预测为正类的概率，然后通过将该概率与某个阈值进行比较来判断样本的类别，比如当预测概率大于阈值时，预测该样本为正，反之为负。选取的阈值不同，样本的预测类别也会不同，所以会产生不同的混淆矩阵。混淆矩阵不同，准确率、精确率、召回率等的取值也会不同。那么如何综合考察呢？可以用ROC曲线以及AUC值。

- 将阈值从0变化到1，在每个阈值下，分别计算精确率，即TPR（true positive rate，TP/(TP+FN)）和FPR（false positive rate，FP/(TN+FP)）。然后以TPR作为纵轴，以FPR作为横轴，即可以得到一条曲线，该条曲线即为ROC（receiver operating characteristic）曲线。

AUC

- 为了进一步衡量ROC曲线代表的模型预测情况，可以计算AUC（Area Under the ROC Curve）取值。AUC表示ROC曲线下方的面积。
- 右图展示了两个模型（A和B）对应的ROC曲线，其阴影部分的面积就是对应的AUC取值。AUC的取值越高表示模型的预测效果越好，例如模型B的预测效果优于模型A。

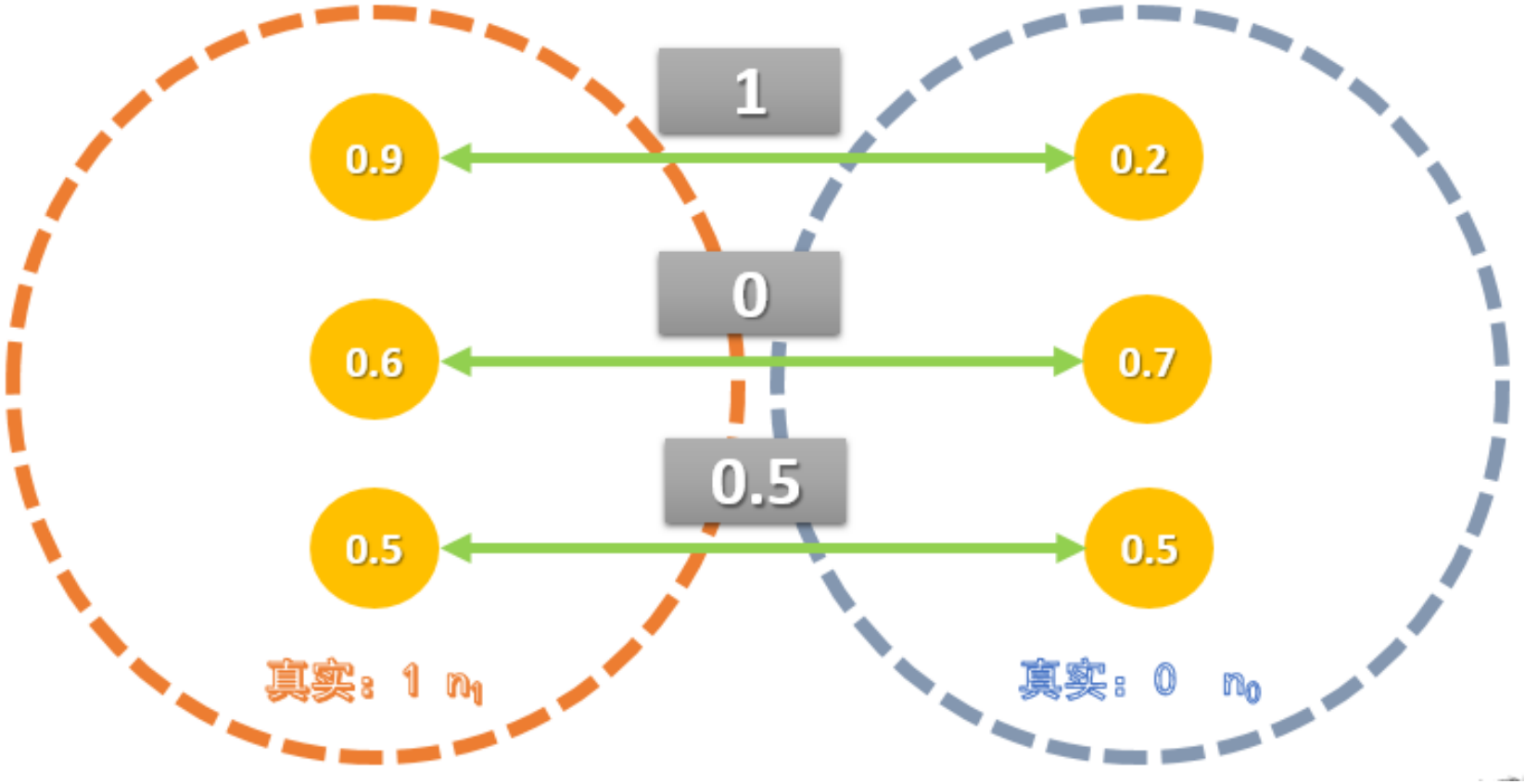


AUC

AUC指标在二分类问题中经常被用来衡量模型的稳定性。另外一种理解AUC取值的角度是：假如从数据集中随机抽取一个正例，一个负例，通过分类器可以得到两个样本被预测为正例的概率，分别记作P(正)和P(负)，那么通过分类器获得P(正)大于P(负)的概率为AUC。

$$\widehat{AUC} = \frac{\sum_{i\in D_1}\sum_{j\in D_0}\left\{I\left[p\left(X_i^*\hat{\beta}\right) > p\left(X_j^*\hat{\beta}\right)\right] + 0.5I\left[p\left(X_i^*\hat{\beta}\right) = p\left(X_j^*\hat{\beta}\right)\right]\right\}}{n_1 n_0}$$

AUC



END