文本分类实战

刘春花 姜姗

流程目录-1

1.预备环境

scikit-learn 简介

使用scikit-learn的基本流程

2.基础知识

TF-IDF

SVM

流程目录-2

3. 代码实现（向量空间模型）

Bag-of-words -> SVM

TF-IDF -> SVM

PCA-> SVM

4. Coding Time!

Scikit-learn 简介

Scikit-learn项目最早由数据科学家 David Cournapeau 在 2007 年发起，需要NumPy和SciPy等其他包的支持，是Python语言中专门针对机器学习应用而发展起来的一款开源框架。

Scikit-learn的基本功能主要被分为六大部分：分类，回归，聚类，数据降维，模型选择和数据预处理。

Scikit-learn 简介

重要链接：

- 官网：http://scikit-learn.org/stable/
- Github： https://github.com/scikit-learn/scikit-learn
- 入门指南：https://www.leiphone.com/news/201701/ZJMTak4Y8ch3Nwd0.html

使用scikit-learn的基本流程

1. 获取数据文件

本次用到的数据集被称为“20种新闻组”，是一个经常被用来进行机器学习和自然语言处理的数据集。它包含20种新闻类别的近20000篇新闻，其官方简介可参见 http://qwone.com/~jason/20Newsgroups/ 。

我们使用sclearn的自带函数sklearn.datasets.fetch_20newsgroups。这个函数能自动从网上下载“20种新闻组”的数据并进行读取，为了节省计算和处理的时间，我们仅选取20种分类中的四种进行之后的分析工作。

使用scikit-learn的基本流程

2. 提取文本特征

无论是什么机器学习方法，都只能针对向量特征（也就是一系列的数字组合）进行分析，因此在读取文本之后，我们要将文本转化为数字化的特征向量。

常用模型：Bags of words，TF-IDF

使用scikit-learn的基本流程

3. 估计器（Estimator）

估计器，很多时候可以直接理解成**分类器**，主要包含两个函数：

fit()：训练算法，设置内部参数。接收训练集和类别两个参数。

predict()：预测测试集类别，参数为测试集。

大多数scikit-learn分类器接收和输出的数据格式均为numpy数组或类似格式。

使用scikit-learn的基本流程

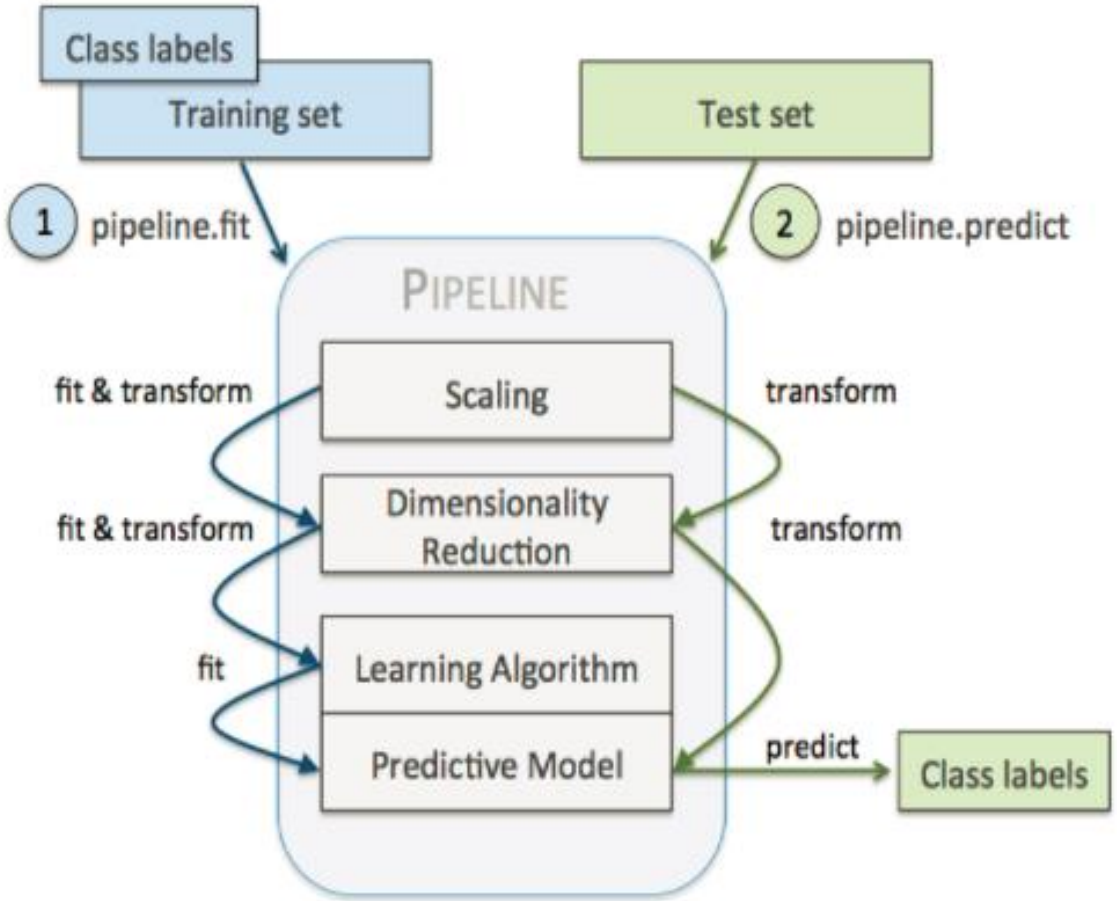
4. 建立管道(Pipeline)

为了简化对于原始数据的清洗、特征提取以及分类过程，Scikit-learn提供了Pipeline类来实现一个整合式的分类器建立过程。

分类器可以通过建立一个Pipeline的方式来实现，而各种特征提取、分类方法都可以在建立Pipeline的时候直接指定，从而大大提高编码和调试的效率

使用scikit-learn的基本流程

4. 建立管道(Pipeline)



Bag-of-words

① Bob likes to play basketball, Jim likes too.

② Bob also likes to play football games.

基于这两个文本文档，构造一个**词表**：

Vocabulary =

{1. “Bob”, 2. “like”, 3. “to”, 4. “play”, 5. “basketball”,6. “also”, 7. “football”, 8. “games”, 9. “Jim”, 10. “too”}。

这个词表一共包含10个不同的单词，利用词表的索引号，上面两个**文档**可以用一个**10维向量表示**(向量中元素为词表中单词在文档中出现的**频率**)：

D=[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]

① [1, 2, 1, 1, 1, 0, 0, 0, 1, 1]

② [1, 1, 1, 1 ,0, 1, 1, 1, 0, 0]

TF-IDF 原理

TF-IDF是一种用于资讯检索与资讯探勘的常用加权技术。它是一种统计方法，用以评估一字词对于一个文件集或一个语料库中某一份文件的重要程度。

TF-IDF的含义是***词频逆文档频率***，其含义时：如果某个词或短语在一篇文章中出现的频率高，并且在其他文章中很少出现，则认为此词或者短语具有很好类别区分能力，适合用来分类。

TF-IDF 计算方法

词频(TF)：

某一个给定的词语在该文件中出现的频率。这个数字是对词数的归一化，以防止它偏向长的文件。对于在某一特定文件里的词语来说，它的重要性可以表示为：

$$\mathrm{tf}_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

以上公式中，分子是该词在文件中的出现次数，而分母则是在文件中所有字词出现次数之和。

TF-IDF 计算方法

逆向文件频率(IDF)：

是一个词语普遍重要性的度量。某一特定词语的IDF，可以由总文件数目除以包含该词语的文件的数目，再将得到的商取对数得到。

$$\mathrm{idf_i} = \log \frac{|D|}{|\{j : t_i \in d_j\}|}$$

分子：语料库中的文件总数

分母：包含词语的文件数目。如果该词语不在语料库中，就会导致分母为零，因此一般情况下使用$1+\{\mathrm{j}: t_i \in d_j\}$

TF-IDF 计算方法

然后再计算TF与IDF的乘积。

$$TF - IDF_{\mathrm{ij}} = TF_{ij} \times IDF_i$$

某一特定文件内的高词语频率，以及该词语在整个文件集合中的第文件频率，可以产生高权重的TF-IDF。因此，TF-IDF倾向于过滤掉常见的词语，保留重要的词语。

SVM 原理

首先什么是SVM?

让我们来看一下wiki.

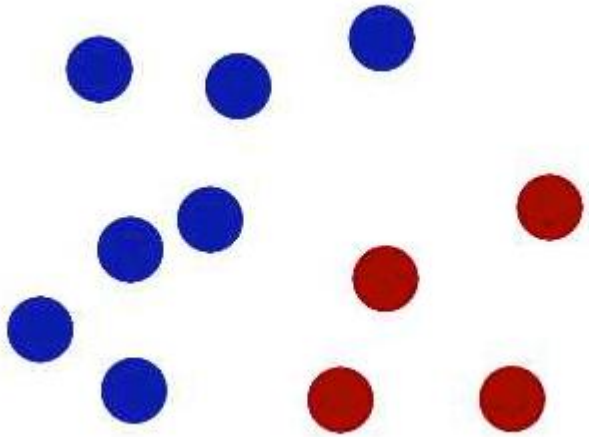
In machine learning, **support vector machines** (**SVMs**, also **support vector networks**[1]) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting). An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall.

如何向吃瓜群众解释支持向量机(SVM)

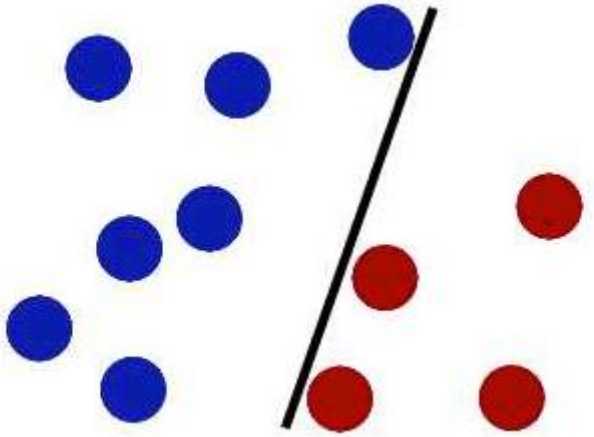
用人类的语言解释SVM

在很久以前的情人节，大侠要去救他的爱人，但魔鬼和他玩了一个游戏。

魔鬼在桌子上似乎有规律放了两种颜色的球，说：“你要用一根木棍分开它们，要求尽量在放更多球之后，仍然适用。”



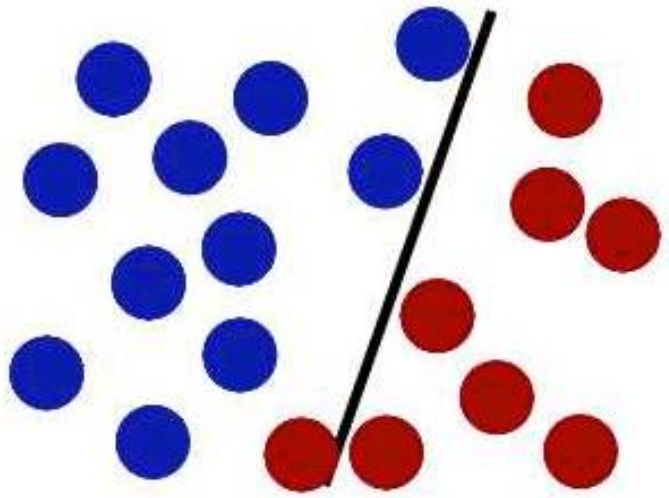
最初



大侠的方案

用人类的语言解释SVM

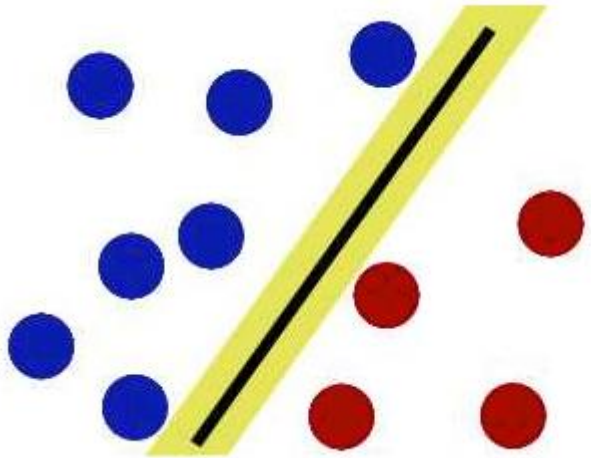
然后魔鬼，又在桌上放了更多的球，似乎有一个球站错了阵营。

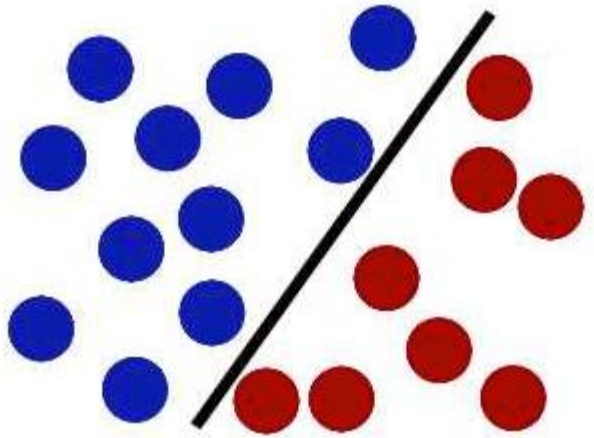


嗯，大侠这时候决定施展武功--SVM！

用人类的语言解释SVM

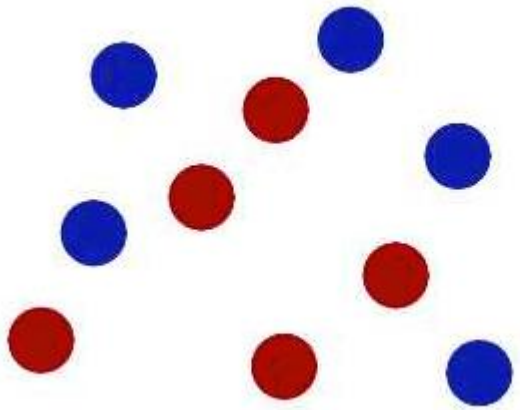
SVM就是试图把棍放在最佳位置，好让在木棍的两边有尽可能大的间隙。现在即使魔鬼放了更多的球，木棍仍然是一个好的分界线。





用人类的语言解释SVM

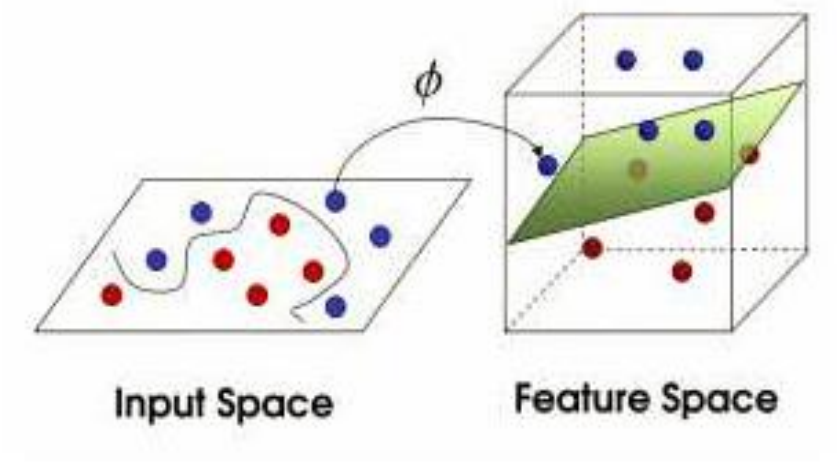
魔鬼看到大侠已经学会了一个trick，于是魔鬼给了大侠一个新的挑战。

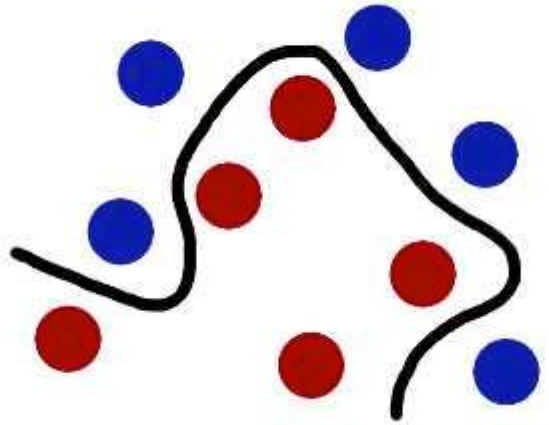


显然，大侠现在没有办法用木棍帮他分开两种球了，现在怎么办呢？

用人类的语言解释SVM

当然像所有武侠片演的那样，大侠一拍桌子，球飞到空中。然后，大侠抓起一张纸，插到了两种球的中间。





现在，从魔鬼的角度看这些球，这些球看起来像是被一条曲线分开了

用人类的语言解释SVM

再之后，无聊的大人们闲得慌，把这些:

球叫做「**data**」

木棍叫做「**classifier**」

最大间隙trick 叫「**optimization**」

拍桌子叫做「**kernelling**」

那张纸叫做「**hyperplane**」

离纸最近的那些球叫做「**support vector**」

用人类的语言解释SVM

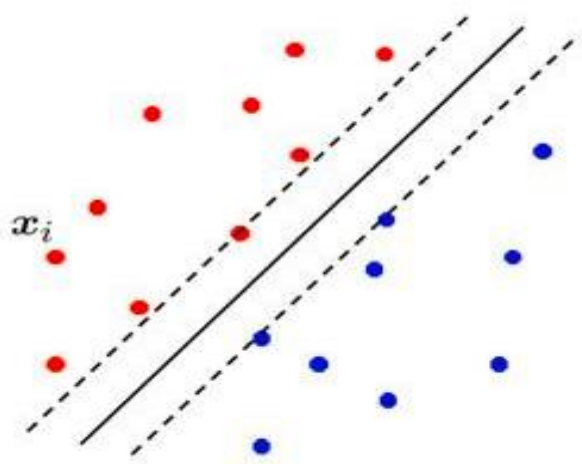
对于很多分类问题，例如最简单的，一个平面上的两类不同的点，如何将它用一条直线分开？在平面上我们可能无法实现。

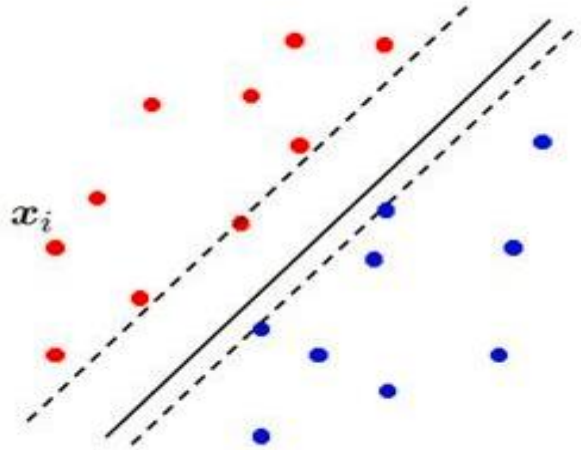
但是如果通过某种映射，将这些点映射到其它空间，我们就有可能在另外一个空间中容易得找到这样一条所谓的“分隔线”，将这些点分开，这就是SVM的基本原理。

用人类的语言解释SVM

但是仅仅找到超平面是不够的，因为在通常的情况下，满足条件的“超平面”的个数不是唯一的。SVM 需要的是利用这些超平面，找到这两类点之间的“最大间隔”。

为什么要找到最大间隔呢？因为分类间隔越大，对于未知点的判断会越准确，也可以说是“最大分类间隔”决定了“期望风险”。





用人类的语言解释SVM

小结：

SVM是一种有监督的学习方法，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合等其他机器学习问题中。

用人类的语言解释SVM

彩蛋

代码实现

Coding Time!!!

Thank You！！！