Table of Contents

线性文本分类**(Linear text classification)**

   &ensp我们先来看下文本分类**(text classification)**问题：现有一篇文章，给它分配一个独立的标签$$y\in Y$$, $$Y$$代表所有可能的标签集。文本分类有很多的应用，如垃圾邮件的过滤，电子病历的分析等等，同时它也是构成更复杂自然语言处理的基本要素。   &ensp要完成这样一个任务，首先问题是如何表示每一篇文章。一个常见的方法就是把文章中每个词的个数构成一个向量，例如$$x=[0,1,1,0,0,2,0,1,13,0...]\top $$,这里$$x_j$$表示词j的个数，$$x$$的长度为$$V\equiv|\nu| $$，$$\nu$$是所有可能词语的集合   &ensp通常我们把$$x$$这样一个向量称之为词袋**(bag of words)**，因为它所包含的信息只有每个词的个数，不包含每个词的在文章中出现的顺序，而且它抛弃了语法，句子边界，段落等所有信息。尽管这样，词袋模型应用在文本分类上效果还是很好。设想下如果你在一个邮件中看到freee,是否就可以判定为垃圾邮件？如果看到Bayesian这个词呢？很多标签分类问题中，一些单个词就可以起到很大的预测作用。   &ensp通过词袋去预测一个标签，我们可以给词表里的每个词语打分，来评估它们和标签的匹配度。在垃圾邮件分类应用里，当标签为垃圾邮件时，freee这个词的打分是一个正值分数，而Bayesian就是一个负值分数。这些分数称之为权值**(weights)**,通常把它们排成一个列向量表示为$$\theta$$ &ensp有时你需要一个多类别分类器，也就是$$K\equiv|Y|>2$$。例如，我们想把新闻分为体育，名人，音乐和商业等类别。我们已知词袋向量$$x$$，通过权值向量$$\theta$$来预测标签$$y$$。对每个标签$$y\in Y$$，我们计算$$\psi(x,y)$$，这个标量公式衡量了$$x$$和$$y$$之间的匹配程度，在线性词袋分类器中，这个式子就是权值$$\theta$$和特征函数**(feature function)**的内积

$$ \psi(x,y)=\theta \cdot f(x,y)\tag{2.1}

$$ 式子可以看出，函数含有两个参数，词的个数$$x$$和标签$$y$$。函数返回一个特征向量。例如，已知$$x,y$$。特征向量的第j个元素就是，

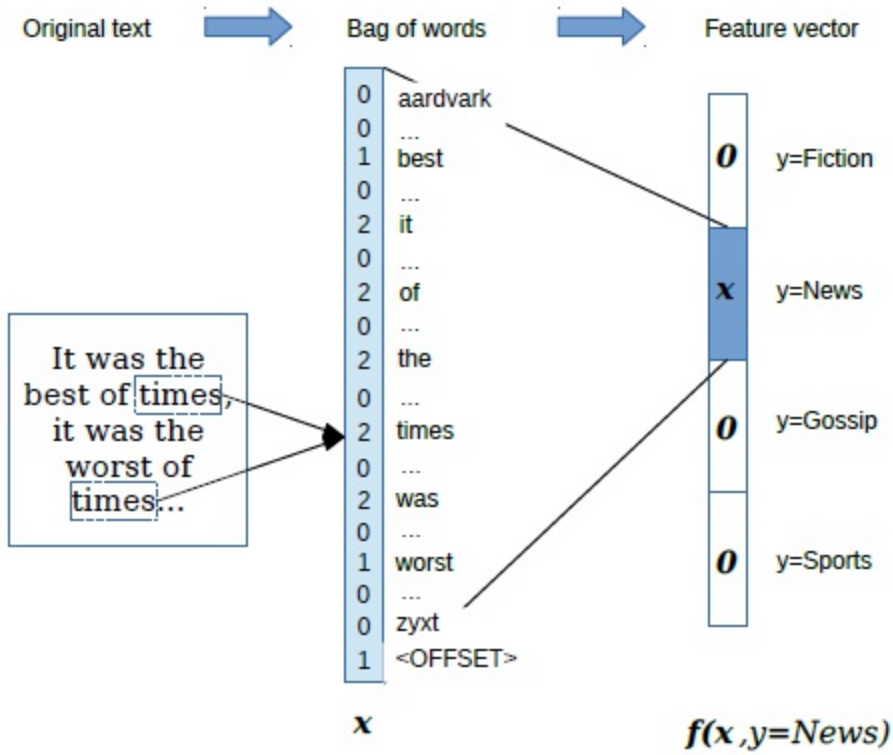
$$ \begin{eqnarray} f*{(x,y)} = *left{ *begin{array}{ll} x*{freee} & \mbox{if } y = Spam \ 0 & otherwise \end{array} \right. \end{eqnarray}\tag{2.2}

$$ 当label为SPAM时，函数返回词freee的个数，否则返回0.对应的权值$$\theta_j$$ 就是衡量freee和SPAM之间的匹配程度。一个正值分数表示出现这个词的文章很有可能分成这个标签。   &ensp为更好地形式化这样一个特征函数，我们定义列向量$$f(x,y)$$，

$$ \begin{align} f(x,y=1)=&[x;\underbrace{0;0;...;0}*{(K-1)\times V}]\tag{2.3}\ f(x,y=2)=&*{[\underbrace{0;0;...;0}_V;x;\underbrace{0;0;...;0}*{(K-2)\times V}]}\tag{2.4}\ f(x,y=K)=&[\underbrace{0;0;...;0}_{(K-1)\times V};x]\tag{2.5} \end{align}

$$

其中$$[\underbrace{0;0;...;0}_{(K-1)\times V}]$$表示长度为$$(K-1)\times V$$的零向量，分号表示垂直连结。从图2.1可以更形象看出这点。



       &ensp图2.1:词袋和特征向量表示，文本分类任务
   &ensp如果已知了权值向量$$\theta\in R^{V\times K}$$，我们就可以计算$$\psi(x,y)$$了,通过内积操作标量化目标对象$$x$$和标签$$y$$之间匹配度。 $$\overline y= arg \underbrace{ min }_{y\in Y}\psi(x,y)\tag{2.6}$$$$\psi(x,y)=\theta \cdot f(x,y)\tag{2.7}$$ 这个内积式子清楚地把data($$x和y$$)与参数$$\theta$$分割开，并且它可以很自然地推广到**结构化预测（structured prediction）**，Y的空间比较大，我们目的就是把标签之间的共享子结构模型化。   &ensp通常我们在词袋向量$$x$$后面加一个**偏移特征(offset feature)**:1。另外，为了统一向量的长度，向量的其他维度填充为0。因此，这个特征向量$$f(x,y)$$的长度就是$$(V+1)\times K$$。偏移特征的权值就可以看是作被分为对应标签的偏向或逆向程度。举个例子，如果我们希望大部分文章为垃圾邮件，那对应y=SPAM的偏移特征的权值就应该比y=HAM的权值大。
   &ensp那么问题来了，权值向量$$\theta$$从哪里来？一个可行的办法就是人为手动设置。如果我们想区分一段文本是英语还是西班牙语，我们可以用英语和西班牙的词典，为每一个单词出现在相关的词典里设置一个权值。例如，

$$\theta_{(E,bicycle)}=1 \qquad \theta_{(S,bicycle)}=0$$
$$\theta_{(E,bicicleta)}=0 \qquad \theta_{(S,bicicleta)}=1$$
$$\theta_{(E,con)}=1 \qquad \theta_{(S,con)}=1$$
$$\theta_{(E,ordinateur)}=0 \qquad \theta_{(S,ordinateur)}=0.$$

同样，如果我们想做正负情感分析，我们可以利用正负情感词典**(sentiment lexicons)**   &ensp但是由于词典的庞大和难以选择合适的权值使得手动设置权值变得相当困难。因此，我们要从数据中学习这些权值。电子邮件用户手动地给垃圾邮件标记为SPAM；新闻工作者把他们自己的文章标记为商业类或者时尚类。利用这些实例标签**(instance labels)**，我们可以通过监督式机器学习**(supervised machine learning)**自动获得这些权值。这章将讨论几种用于分类的机器学习方法。第一个是基于概率计算的。**

2.1朴素贝叶斯(Naive Bayes)

fdsf