Classifier

2018 年 7 月 5 日

1 單純貝氏 (分類)

1.1 介紹

單純貝氏 (Naive Bayes) 是一個非常經典的機率算法，利用機率來推測出某個特徵值下，是某種分類的機率。

這個方法對於文字的分類非常的有效果，而且可以經由少數的數據就達成一個很棒的效果。

1.2 條件機率

在介紹單純貝氏前，我們一定要先介紹條件機率，條件機率其實是一個非常簡單的概念。

條件機率 = 設定一個條件下，某件事情發生的機率

舉個例子，我們就以文章的例子來說好了

P(這是一篇愛情相關的文章 | 文章中出現 "伴侶" 這詞) 這就是一個所謂的條件機率

整個意思是當我已經設定好 (觀察到) 這篇文章中有出現 "伴侶" 這個詞的時候，文章是愛情相關的機率是多少

1.3 貝氏定理

那這個機率跟我們的分類問題有什麼關係呢?

今天我給你一篇文章，你已經觀察到裡面出現了伴侶這個詞，但是你想預測他是什麼分類的。

那該怎麼做呢? 很簡單，假設我們有三類文章: 宗教，愛情，工作。

你只要計算 $P(\text{宗教} \mid \text{伴侶詞出現}), P(\text{愛情} \mid \text{伴侶詞出現}), P(\text{工作} \mid \text{伴侶詞出現})$

再比較一下三個機率誰大誰小！大的就是我們的預測分類!

問題是怎麼簡易計算這機率呢?

貝氏定理就可以拿出來了，我們用最簡單的方式描述一次貝氏定理

$P(B \text{ 發生}) \times P(A \text{ 發生} \mid B \text{ 發生}) = P(A \text{ 發生}) \times P(B \text{ 發生} \mid A \text{ 發生}) = P(AB \text{ 同時發生}) = P(A B)$

套用到我們上面的問題

假設 A: 宗教或愛情或工作 B: 伴侶詞出現

$P(\text{宗教} \mid \text{伴侶詞出現}) = P(\text{宗教}) \times P(\text{伴侶詞出現} \mid \text{宗教}) \propto P(\text{伴侶詞出現})$

$P(\text{宗教})$: 宗教文章數量 / 我們蒐集的文章總數量

$P(\text{伴侶詞出現} \mid \text{宗教})$: 把你蒐集的宗教文章拿出來看看伴侶出現的機率是多少

$P(\text{伴侶詞出現})$: 把所有文章拿出來看看伴侶出現機率，但事實上你不用計算，因為我們只是要比大小！而不需要真正的值。

你會發現，你想得到的答案，可以藉由你事先蒐集的資料比較好統計的部分推算出來!!!

而且只要比較 $P(\text{類別}) \times P(\text{伴侶詞出現} \mid \text{類別})$ 就可以了!!!!!

1.4 單純貝氏

事實上，我們必須把上面的式子寫個更完整一點

$P(\text{宗教} \mid \text{伴侶詞出現 } n \text{ 次}) = P(\text{宗教}) \times P(\text{伴侶詞出現 } n \text{ 次} \mid \text{宗教}) \propto P(\text{伴侶詞出現 } n \text{ 次})$

你會發現比較上面的 $P(\text{宗教}) \times P(\text{伴侶詞出現 } n \text{ 次} \mid \text{宗教})$ 就可以了！這個值是可以計算出來的 (方法類似丟骰子 5 次，出現 3 次 6 機率是多少)

再擴展一下，

$P(\text{宗教} \mid \text{伴侶詞出現 } n \text{ 次且信仰詞出現 } m \text{ 次}) = P(\text{宗教}) \times P(\text{伴侶詞出現 } n \text{ 次且信仰詞出現 } m \text{ 次} \mid \text{宗教}) \propto P(\text{伴侶詞出現 } n \text{ 次且信仰詞出現 } m \text{ 次})$

還是比較 $P(\text{宗教}) \times P(\text{伴侶詞出現 } n \text{ 次而且信仰詞出現 } m \text{ 次} \mid \text{宗教})$ 就可以了，但你發現，這個是很難計算的!

因為你根本不知道這兩件事情 (伴侶和信仰這兩個詞) 到底有沒有關係!

但我們做了一個很 ‘‘單純’’ 的假設，就假設這兩個詞的出現完全沒有關係 (當然是不符合現實的，不過我們只是要比大小，所以這假設並不會影響結果太多)

假設完以後，我們就可以說:

$P(\text{伴侶詞出現 } n \text{ 次且信仰詞出現 } m \text{ 次} \mid \text{宗教}) = P(\text{伴侶詞出現 } n \text{ 次} \mid \text{宗教}) \times P(\text{信仰詞出現 } m \text{ 次} \mid \text{宗教})$

做完假設以後!! 你發現我們整個東西是可以計算的，就可以開始來比大小，決定我們的分類了!

1.5 文字資料的整理

1.5.1 詞向量

文字最基本的整理方法就是把分詞後的結果，做成一個向量 (有方向的量)，你可以把向量想像成往某個方向走出幾步，再往某個方向走幾步的感覺。

例子: 我們現在有兩篇文章

文章 1: 我喜歡吃牛排，也喜歡吃雞排文章 2: 你喜歡閱讀和旅遊

切割 1: 我/喜歡/吃/牛排/，/也/喜歡/吃/雞排切割 2: 你/喜歡/閱讀/和/旅遊

這裡要注意，由於特徵是針對所有的文章計算，所以你要把所有你要訓練的文章先統計好有幾種不同的詞，這個種數就是你要做出的維度

這上面的例子總共有我/喜歡/吃/牛排/也/雞排/你/閱讀/和/旅遊 10 個維度

特徵	我	喜歡	吃	牛排	也	雞排	你	閱讀	和	旅遊
向量 1	1	2	1	1	1	1	0	0	0	0
向量 2	0	1	0	0	0	0	1	1	1	1

1.5.2 TF-IDF 方法

事實上，如果要實作單純貝氏，上面的數量向量已經可以拿來當我們的資料了

不過其實我還是會習慣把他化成 tf-idf 量度

什麼是 tf-idf 呢?

tf: **這個詞出現在整篇文章的次數，出現越多次，代表這個詞越能代表整篇文章**

idf: **這個詞在我蒐集的全部文章出現過的文章數，出現越多次，代表這個詞是一個慣用詞，重要性下降。**

上面兩個係數成起來就是我們的語言分析的基礎，衡量這篇文章的關鍵詞是什麼的方法

1.5.3 預先處理

中文的預先處理只有一個最高守則！請記得都轉成簡體或者繁體，不然會被算成不同兩個字!

1.6 資料集

整理過後的新聞語料集，簡體已都轉成繁體了!

https://drive.google.com/open?id=1zbVKIHMUugqXkKDc4q7CNaWXcwQ7pzFE

請到上面把整個資料下載

共分三個

chinese_news: 原本的新聞

chinese_trans: 翻譯成繁體的新聞

chinese_tests: 我從原本的新聞每個分類擷取出 10 篇當作測試文章

1.7 預測目標

用單純貝氏預測一篇沒分類過的文章是什麼主題

1. 分類問題
2. 監督式演算法

1.8 需要函式庫

jieba 函式庫: 請使用 pip 或利用 PyCharm 安裝, 可以幫我們做完分詞
https://github.com/fxsjy/jieba
可以根據下面的說明用繁體字典替換

1.9 資料預處理

可以根據下面的說明用繁體字典替換

1.9.1 讀入我們的訓練資料

這裡跟我們之前做的比較不一樣，是讀入整個資料夾的一個一個檔案內容
讀者在這裡可以在中間多把東西 (dir_path, file_names 印出來) 來熟悉讀取檔案流程

In [1]:
```
import os
import jieba
import pandas as pd


# 為了顯示的漂亮, 我刻意的把印出來的 row 只顯示 15 個和 column 只顯示十個
# 大家練習的時候可以去掉下面兩行
pd.set_option('display.max_rows', 15)
pd.set_option('display.max_columns', 10)

# scikit-learn 會有些 deprecation warning, 為了顯示漂亮, 我刻意地忽略掉
import warnings
warnings.filterwarnings('ignore')

base_dir = "chinese_news_trans"
test_dir = "chinese_news_test"
```

```
    # 因為要處理資料夾很多次，所以定義成函式
    def process_dirs(base_dir):
        df = pd.DataFrame(columns = ["類別", "內容"])

        # os.walk 會走到檔案才停下來
        for dir_path, dir_names, file_names in os.walk(base_dir):
            for single_file in file_names:
                if not single_file.startswith("."):
                    f = open(os.path.join(dir_path, single_file), "r", encoding = "utf-8")
                    content = f.read()
                    # 讀完黨以後做出第一步處理，先把換行都去掉
                    content = content.replace("\r", "").replace("\n", "")
                    split_word = jieba.cut(content)
                    # 分詞
                    content = " ".join(split_word)
                    s = pd.Series([dir_path.split("/")[-1], content], index = ["類別", "內
                    df = df.append(s, ignore_index = True)
        df['類別'] = df['類別'].astype('category')
        return df
```

```
In [12]: df = process_dirs(base_dir)
         df
```

Out[12]:

	類別	內容
0	交通	【 日 期 】 19960104 【 版 號 】 1 【 ...
1	交通	【 日 期 】 19960226 【 版 號 】 5 【 ...
2	交通	大 秦鐵路 萬噸 列車 試運 成功 新華社 北京...
3	交通	遼寧省 檯 安縣 村村 都 通 柏油路 鄉村 公...
4	交通	北京 —烏蘭 巴托 —莫斯料 3 ／ 4 次...
5	交通	福建 將建 第二 條出 省鐵路 新華社 福州 5...
6	交通	大秦 二期工程 中 最長 的 平市 東河 特大 ...
...	..	...
2630	體育	參加 首屆 世界 盃 乒乓球 團體賽 的 中國 ...
2631	體育	全國 健美 賽 和 健美操 賽決 出 6 項 第...
2632	體育	馬 玉芹 破 女子 4 0 0 米 跑 全國 青...
2633	體育	國際 奧委會 中國 臺 北委員 吳經國 訪問 北...
2634	體育	亞奧 理事 會 3 9 個 成員 組織 全部 以...

```
2635   體育                世界 盃 乒乓球 團體賽 男子 團體 採用 新賽...
2636   體育                登頂 隊員 簡介 新華社 珠穆 朗瑪峰 5 月 ...

[2637 rows x 2 columns]
```

1.9.2 類別處理

由於 scikit-learn 不接受字串，所以我們一定要把類別轉換成整數
你可以使用 cat.categories 得到所有類別
再使用 cat.codes 轉換成整數

```
In [3]: # 把類別替換成 code, 並且記錄起來
        # 走過 categories 同時順便把字典創造起來
        saved_map = { cat:df['類別'].cat.categories.get_loc(cat)
                     for cat in df['類別'].cat.categories }
        saved_map

Out[3]: {' 交通': 0,
         ' 政治': 1,
         ' 教育': 2,
         ' 環境': 3,
         ' 經濟': 4,
         ' 藝術': 5,
         ' 計算機': 6,
         ' 軍事': 7,
         ' 醫藥': 8,
         ' 體育': 9}

In [4]: df['類別'] = df['類別'].cat.codes
        df

Out[4]:      類別                                                  內容
        0     0  【 日 期 】 19960104 【 版 號 】 1 【 ...
        1     0  【 日 期 】 19960226 【 版 號 】 5 【 ...
        2     0                  大 秦鐵路 萬噸 列車 試運 成功 新華社 北京...
        3     0                  遼寧省 檯 安縣 村村 都 通 柏油路 鄉村 公...
        4     0                  北京 —烏蘭 巴托 —莫斯科 3 ／ 4 次...
        5     0                  福建 將建 第二 條出 省鐵路 新華社 福州 5...
        6     0                  大秦 二期工程 中 最長 的 平市 東河 特大 ...
```

```
...   ..                                                  ...
2630   9                    參加 首屆 世界 盃 乒乓球 團體賽 的 中國 ...
2631   9                    全國 健美 賽 和 健美操 賽決 出 6 項 第...
2632   9                    馬 玉芹 破 女子 4 0 0 米 跑 全國 青...
2633   9                    國際 奧委會 中國 臺 北委員 吳經國 訪問 北...
2634   9                    亞奧 理事 會 3 9 個 成員 組織 全部 以...
2635   9                    世界 盃 乒乓球 團體賽 男子 團體 採用 新賽...
2636   9                    登頂 隊員 簡介 新華社 珠穆 朗瑪峰 5 月 ...

[2637 rows x 2 columns]
```

1.9.3 讀入測試資料

把我們的測試資料讀取，並且使用剛剛存起來的 category 來 map

```
In [5]: test_df = process_dirs(test_dir)
        test_df
```

```
Out[5]:      類別                                                  內容
        0    交通  日 月光 華   - -   Traffic _ Info 精華區 文章 閱讀 - - - ...
        1    交通  日 月光 華   - -   Traffic _ Info 精華區 文章 閱讀 - - - ...
        2    交通  日 月光 華   - -   Traffic _ Info 精華區 文章 閱讀 - - - ...
        3    交通  三趟 火車 停開   乘客 可 全額 退票   瀏覽次數 : 1180          ...
        4    交通  日 月光 華   - -   Traffic _ Info 精華區 文章 閱讀 - - - ...
        5    交通  日 月光 華   - -   Traffic _ Info 精華區 文章 閱讀 - - - ...
        6    交通  日 月光 華   - -   Traffic _ Info 精華區 文章 閱讀 - - - ...
        ..   ..                                                ...
        94   體育  中國 青島 一 韓國 大邱 健美 表演 賽 和 對 抗賽 將舉 辦              ...
        95   體育  男子 健美 初登亞 運會   中國 猛男 直指 前 三                在 即將...
        96   體育  最 優秀 選手 無緣 亞運會 健美 賽                健美 在 亞洲 運動會 ...
        97   體育    各國 記者 眼中 的 羽毛球 世錦賽   - - - - - - - - - - - -...
        98   體育  友好 運動會 第五天   東道 主選手 大顯 神威 2001 年 09 月 03 日 02 ...
        99   體育  不靠 技術 比運氣       第二 屆 奧運會 在 巴黎 舉行 , 同時 這裡 也 正在 舉行...
        100  體育  帆   板   運   動   簡   介 (二 )       我國 在 79 年 由 國家...

        [101 rows x 2 columns]
```

1.9.4 替換測試類別

這邊必須使用剛剛存起來的字典來替換
因為如果直接使用 code 可能會發生沒對照到的事故

```
In [6]: test_df['類別'] = test_df['類別'].replace(saved_map)
        test_df
```

Out[6]:

	類別	內容
0	0	日 月光 華 - - Traffic _ Info 精華區 文章 閱讀 - - - ...
1	0	日 月光 華 - - Traffic _ Info 精華區 文章 閱讀 - - - ...
2	0	日 月光 華 - - Traffic _ Info 精華區 文章 閱讀 - - - ...
3	0	三越 火車 停開 乘客 可 全額 退票 瀏覽次數 : 1180 ...
4	0	日 月光 華 - - Traffic _ Info 精華區 文章 閱讀 - - - ...
5	0	日 月光 華 - - Traffic _ Info 精華區 文章 閱讀 - - - ...
6	0	日 月光 華 - - Traffic _ Info 精華區 文章 閱讀 - - - ...
..	..	...
94	9	中國 青島 一 韓國 大邱 健美 表演 賽 和 對 抗賽 將舉 辦 ...
95	9	男子 健美 初登亞 運會 中國 猛男 直指 前 三 在 即將...
96	9	最 優秀 選手 無緣 亞運會 健美 賽 健美 在 亞洲 運動會 ...
97	9	各國 記者 眼中 的 羽毛球 世錦賽 - - - - - - - - - - - -...
98	9	友好 運動會 第五天 東道 主選手 大顯 神威 2001 年 09 月 03 日 02 ...
99	9	不靠 技術 比運氣 第二 屆 奧運會 在 巴黎 舉行 , 同時 這裡 也 正在 舉行...
100	9	帆 板 運 動 簡 介 (二) 我國 在 79 年 由 國家...

[101 rows x 2 columns]

```
In [7]: from sklearn.feature_extraction.text import TfidfVectorizer

        vec = TfidfVectorizer()
        # 注意一定要使用 fit_transform, 才會幫你轉換成詞向量
        bag = vec.fit_transform(df['內容'])
        print("總共維度:", len(vec.get_feature_names()))
        # 讀者可以把註解拿掉, 就可以看到 transform 後的 matrix
        #for entry in bag:
        #    print(entry)
```

總共維度: 96042

1.10 scikit-learn 單純貝氏

1.10.1 合適的選擇

scikit-learn 裡的單純貝氏有對你的資料類型做過優化

sklearn.naive_bayes : Naive Bayes

The `sklearn.naive_bayes` module implements Naive Bayes algorithms. These are supervised learning methods based on applying Bayes' theorem with strong (naive) feature independence assumptions.

User guide: See the Naive Bayes section for further details.

`naive_bayes.BernoulliNB` ([alpha, binarize, ...])	Naive Bayes classifier for multivariate Bernoulli models.
`naive_bayes.GaussianNB` ([priors])	Gaussian Naive Bayes (GaussianNB)
`naive_bayes.MultinomialNB` ([alpha, ...])	Naive Bayes classifier for multinomial models

1. BernoulliNB: 對於特徵是 True 和 False 二分法優化
2. GaussianNB: 對於特徵是高斯分布的連續數字優化
3. MultinomialNB: 對於特徵是整數，而且是數幾次的分布優化，不過使用就算使用 tf-idf 分數一樣可以達到很好效果

1.10.2 Alpha 的選擇

Alpha 的用意是 Laplace smoothing 的意思
我們現在矩陣裡面很多元素都是等於 0
但等於 0 會造成一個現象，就是你發生的機率會是 0
這時候由於我們的資料並不是完整的資料，所以你如果說 P(某個字出現 | 某種分類) = 0
反而會造成過度武斷，而導致預測的行為變差
所以 smoothing 最主要的用意是讓 0 機率不要是 0，而是一個接近 0 的微小機率 (有機率跟沒機率差很多！)

$$P(X_j|Y_k)\frac{Count(X_j,Y_k)+1}{\sum_j^V (Count(X_j,Y_k)+1)}.$$

我們如果不設定 alpha 的話，預設值就是如同公式的 1
如果你使用的是 CountVectorizer，那沒有問題，就照著預設值設置
但如果你使用的是 TfidfVectorizer，tf-idf 分數算出來都大概在 0.x 而已
你加 1 就會變得不太實在，所以如果使用 tf-idf 的時候，我們通常會設置 alpha = 0.001

Smoothing 的白話文理解: 假設你把出現 "收益" 這詞的信 100% 當成不是垃圾信，那會有一個很嚴重的後果，只要在垃圾信裡加上收益這詞，就永遠不會被你檢測出來，因為 0 乘上去以後，整體機率就是 0，你永遠不會去考慮任何其他特徵，所以你不應該讓機率是 0%，而是用一個極小接近 0 的機率來算，這就是 Smoothing

```
In [8]: # 只要是文字，我們通常就會選擇 MultinomialNB
        from sklearn.naive_bayes import MultinomialNB

        clf = MultinomialNB(alpha = 0.001).fit(bag, df['類別'])
```

```
In [9]: # 由於我們用剛剛的 vec 訓練他，所以維度會保持跟剛剛一樣
        test_bag = vec.transform(test_df['內容'])
        print("維度:", len(vec.get_feature_names()))
```

```
維度: 96042
```

```
In [10]: from sklearn.metrics import accuracy_score
         predict = clf.predict(test_bag)
         print("預測:", list(predict))
         print("正確標籤:", list(test_df['類別']))
         print("Naive-Bayes 正確率: ", accuracy_score(test_df['類別'], predict) * 100, "%")
```

```
預測: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2
正確標籤: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2,
Naive-Bayes 正確率:  100.0 %
```

```
In [11]: from sklearn import neighbors

         clf = neighbors.KNeighborsClassifier(n_neighbors=8)
         clf = clf.fit(bag, df['類別'])
         predict = clf.predict(test_bag)

         print("預測:", list(predict))
         print("正確標籤:", list(test_df['類別']))
         print("kNN 正確率: ", accuracy_score(test_df['類別'], predict) * 100, "%")
```

```
預測: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2
正確標籤: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2,
kNN 正確率:  100.0 %
```

1.10.3 結論

由於訓練資料非常完善，而且目標相對單純，所以你發現不管單純貝氏或者 kNN 都可以順利達到高正確率