МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМЕНИ Н. Э. БАУМАНА
Факультет информатики и систем управления
Кафедра теоретической информатики и компьютерных технологий

Лабораторная работа №10
по курсу «Автоматическая обработка текстов»

«Классификация документов с помощью Байесовского классификатора»

Выполнил:
студент группы ИУ9-11М
Беляев А. В.

Проверила:
Лукашевич Н. В.

Москва 2018

1 Цель работы

Для заданных документов и их классов обучить модель и на ее основе отнести входящий документ к одному из классов. Воспользоваться алгоритмами Multinomial и Bernoulli.

Даны документы и их классы:

- D1=(x1, x2, x3), класс C1
- D2=(x1, x2, x4), класс C1
- D1=(x4, x5, x6), класс C2

Входящий документ: D4=(x1, x4, x5)

2 Вычисления

2.1 Multinomial

TRAINMULTINOMIALNB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c/N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

Рис. 1:

Воспользуемся алгоритмом на Рисунке 1.

$$V = (x1, \cdots x6); N = 3$$

Посчитаем веротяности получения класса C1:

$$N_c = 2$$

$$prior(C1) = \frac{2}{3}$$

$$text = x1x2x3x1x2x4$$

Посчитаем токены:

$$x1, x2 : T = 2$$

$$x3, x4 : T = 1$$
$$x5, x6 : T = 0$$

Посчитаем вероятности термов (признаков) в зависимости от класса для класса C1:

$$cp(x1, C1) = cp(x2, C1) = \frac{2+1}{6+6}$$
$$cp(x3, C1) = cp(x4, C1) = \frac{1+1}{6+6}$$
$$cp(x5, C1) = cp(x6, C1) = \frac{1}{6+6}$$

Проделаем то же самое для класса C2:

$$N_c = 1$$
$$prior(C1) = \frac{1}{3}$$
$$text = x4x5x6$$

Посчитаем токены:

$$x1, x2, x3 : T = 0$$
$$x4, x5, x6 : T = 1$$

Посчитаем вероятности термов (признаков) в зависимости от класса для класса C1:

$$cp(x1, C1) = cp(x2, C1) = cp(x3, C1) = \frac{1}{3+6}$$
$$cp(x4, C1) = cp(x5, C1) = cp(x5, C1) = \frac{1+1}{3+6}$$

```
APPLYMULTINOMIALNB(ℂ, V, prior, condprob, d)
1  W ← EXTRACTTOKENSFROMDOC(V, d)
2  for each c ∈ ℂ
3  do score[c] ← log prior[c]
4     for each t ∈ W
5     do score[c] += log condprob[t][c]
6  return arg max_{c∈ℂ} score[c]
```

Рис. 2:

Выполним алгоритм на Рисунке 2 и посчитаем score для обоих классов:

$$W = (x1, x4, x5)$$
$$score(C1) = \log(\frac{2}{3}) + \log(\frac{2+1}{6+6}) + \log(\frac{1+1}{6+6}) + \log(\frac{1}{6+6}) = -6.068$$
$$score(C2) = \log(\frac{1}{3}) + \log(\frac{1}{3+6}) + \log(\frac{1+1}{3+6}) + \log(\frac{1+1}{3+6}) = -6.303$$

Результирующий класс для документа **4** − **C1**, т.к. его score выше, хотя и незначительно.

2.2 Bernoulli

```
TRAINBERNOULLINB(ℂ, 𝔻)
1  V ← EXTRACTVOCABULARY(𝔻)
2  N ← COUNTDOCS(𝔻)
3  for each c ∈ ℂ
4  do N_c ← COUNTDOCSINCLASS(𝔻, c)
5     prior[c] ← N_c/N
6     for each t ∈ V
7     do N_ct ← COUNTDOCSINCLASSCONTAININGTERM(𝔻, c, t)
8        condprob[t][c] ← (N_ct + 1)/(N_c + 2)
9  return V, prior, condprob
```

Рис. 3:

Воспользуемся алгоритмом `TrainBernoulliNb` на Рисунке 3.

$$V = (x1, \cdots x6); N = 3$$

Посчитаем веротяности для класса C1 по формуле $cp(t, c) = \frac{N_{ct}+1}{N_c+2}$:

$$N_c = 2$$

$$prior(C1) = \frac{2}{3}$$

$$cp(x1, C1) = cp(x2, C1) = \frac{2+1}{2+2} = \frac{3}{4}$$

$$cp(x3, C1) = cp(x4, C1) = \frac{1+1}{2+2} = \frac{2}{4}$$

$$cp(x5, C1) = cp(x6, C1) = \frac{0+1}{2+2} = \frac{1}{4}$$

То же самое для класса C2:

$$N_c = 1$$

$$prior(C2) = \frac{1}{3}$$

$$cp(x1, C1) = cp(x2, C1) = cp(x3, C1) = \frac{0+1}{1+2} = \frac{1}{3}$$

$$cp(x4, C1) = cp(x5, C1) = cp(x6, C1) = \frac{1+1}{1+2} = \frac{2}{3}$$

Выполним процедуру `ApplyBernoulliNB` на Рисунке 4 и посчитаем score для обоих классов:

$$score(C1) = \log\frac{2}{3} + \log\frac{3}{4} + \log(1-\frac{3}{4}) + \log(1-\frac{2}{4}) + \log\frac{1}{4} + \log\frac{1}{4} + \log(1-\frac{1}{4}) = -5.139$$

$$score(C2) = \log\frac{1}{3} + \log\frac{1}{3} + \log(1-\frac{1}{3}) + \log(1-\frac{1}{3}) + \log\frac{2}{3} + \log\frac{2}{3} + \log(1-\frac{2}{3}) = -4.917$$

Наибольший score – у класса **C2**.

```
APPLYBERNOULLINB(ℂ, V, prior, condprob, d)
1  V_d ← EXTRACTTERMSFROMDOC(V, d)
2  for each c ∈ ℂ
3  do score[c] ← log prior[c]
4     for each t ∈ V
5     do if t ∈ V_d
6           then score[c] += log condprob[t][c]
7           else score[c] += log(1 − condprob[t][c])
8  return arg max_{c∈ℂ} score[c]
```

Рис. 4:

3 Выводы

В ходе работы была выполнена классификация документа. Разные алгоритмы дали разные результаты классификации, хотя полученные результаты отличаются незначительно