CSCB705 Класификация и разпознаване на образи

Кует Хъу Нгуен - F89497

Проект: **Спам имейл филтър**

Здравейте, Аз съм Кует Нгуен, факутет номер F89497. Аз съм студент по Информатика. В този проект аз ще направя един спам имейл филтър от "Бейсови класификатори"

Проект: **Спам имейл филтър**

От **lecture_4.pdf** имаме тази формула

$$\sum_i e^{w_i x + v_i}$$

В този модел параметите отново могат да се оценят чрез метода на максимално правдоподобие. Нека разползгаме с две множества $(X_1^{(1)}, \cdots, X_{n_1}^{(1)})$ и $(X_1^{(2)}, \cdots, X_{n_2}^{(2)})$ от наблюдения със всеки един от двата класа. Функцията на правдоподобие ще има вида

$$L = \prod_{i=1,2} \prod_{j=1}^{n_i} P(X_j^{(i)} \mid w_i) P(w_i) = \prod_{i=1,2} \prod_{j=1}^{n_i} P(w_i \mid X_j^{(i)}) P(X_j^{(i)}).$$

6

Проект: **Спам имейл филтър**

От него преобразувах в по-разбираема формула за лесно програмиране

$$P(Spam|w1, w2, w3 \ldots\ldots wn) \infty P(Spam) * \prod_{i=1}^{n} P(wi|Spam)$$

$$P(Non\_Spam|w1, w2, w3 \ldots\ldots wn) \infty P(Non\_Spam) * \prod_{i=1}^{n} P(wi|Non\_Spam)$$

Проект: **Спам имейл филтър**

За да знае на $P(w_i|Spam)$ и $P(w_i|Non\_Spam)$ във формула по-горе, ще трябва да използвам тези.

$$P(w_i|Spam) = \frac{N_{w_i|Spam} + \alpha}{N_{Spam} + \alpha . N_{vocabulary}}$$

$$P(w_i|Non\_Spam) = \frac{N_{w_i|Non\_Spam} + \alpha}{N_{Non\_Spam} + \alpha . N_{vocabulary}}$$

Кодирането

$P(Spam)$ $P(Non\_Spam)$

```
# P(Spam) and P(Non_Spam)
p_spam = len(spam_emails) / len(training_set_clean)
p_non_spam = len(non_spam_emails) / len(training_set_clean)
```

Кодирането

N_{Spam}

```
# N_Spam
n_words_per_spam_email = spam_emails["Email"].apply(len)
n_spam = n_words_per_spam_email.sum()
```

N_{Non_Spam}

```
# N_Non_Spam
n_words_per_non_spam_email = non_spam_emails["Email"].apply(len)
n_non_spam = n_words_per_non_spam_email.sum()
```

$N_{vocabulary}$

```
# N_Vocabulary
n_vocabulary = len(vocabulary)
```

Кодирането

α

```
# Laplace smoothing
alpha = 1
```

$N_{w_i|Spam}$

```
n_word_given_spam = spam_emails[word].sum()  # spam_emails already defined
```

$N_{w_i|Non\_Spam}$

```
p_word_given_non_spam = (n_word_given_non_spam + alpha) / (
    n_non_spam + alpha * n_vocabulary
)
```

Кодирането

$P(w_i|Spam)$

```
p_word_given_spam = (n_word_given_spam + alpha) / (n_spam + alpha * n_vocabulary)
```

$P(w_i|Non\_Spam)$

```
p_word_given_non_spam = (n_word_given_non_spam + alpha) / (
    n_non_spam + alpha * n_vocabulary
)
```

cscb705_train_data.py

с този код

- импортиране на данни от файл **emails_short.csv**

- Реформиране на имейлите: премахване на връзки, специални знаци, превръщане на текста в малки букви

- изчисляване на $P(Spam)$, $P(Non\_Spam)$

и запазвам във файл **coef_values.txt**

- изчисляване на параметрите на **Spam Emails** и **Non_spam Emails** и запазване във файла **parameters_spam.csv** и **parameters_non_spam.csv**

cscb705_train_data.py

И след това ще имам данни за да проверя един емайл дали спам или не спам !!

cscb705_check_spam_email.py

с този код

- импортиране на данни от файл **email.txt**
- проверя емайл

cscb705_check_accuracy.py

с този код

- импортиране на данни от файл **emails_test_set.csv**
- Ще проверя за всеки емайл
- Ще показа точност на алгоритъма