Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 9

The Naïve Bayes Classifier

Naïve Bayes Classifier

- *Naïve Bayes* is a probabilistic classification method based on Bayes' theorem (or Bayes' law) with a few tweaks
- Bayes' theorem gives the relationship between the probabilities of two events and their conditional probabilities.
- Last week, we have looked at one example of applying Bayes' theorem in medical testing

Classification methods: Decision Trees

Source: *The Straits Times*

- We will look at another example of applying Bayes' theorem in email spam filtering

Naïve Bayes Classifier: Bayes' Theorem

- Suppose that 5% of all emails are spams, and that the phrase "you are a winner" occurs in 50% of spam emails, and in 10% of non-spam emails.
- Given that we received an email with the phrase "you are a winner" in it, what is the conditional probability that it is a spam email?

Naïve Bayes Classifier: Bayes' Theorem

- Define the events $C = \{\text{email is spam}\}$ and $A = \{\text{contains the phrase ``you are a winner''}\}$
- Let $\neg\mathcal{M}$ denote the negation of the event $\mathcal{M}$
- Based on the problem description, we have $P(C) = 0.05$, $P(\neg C) = 0.95$, $P(A|C) = 0.50$ and $P(A|\neg C) = 0.10$
- We wish to compute the conditional probability $P(C|A)$.

Naïve Bayes Classifier: Bayes' Theorem

- We wish to compute the conditional probability $P(C|A)$.
- Based on the problem description, we have $P(C) = 0.05$, $P(\neg C) = 0.95$, $P(A|C) = 0.50$ and $P(A|\neg C) = 0.10$

$$
\begin{aligned}
P(C|A) &= \frac{P(A|C)P(C)}{P(A)} \\
&= \frac{P(A|C)P(C)}{P(A \cap C) + P(A \cap \neg C)} \\
&= \frac{P(A|C)P(C)}{P(C) \times P(A|C) + P(\neg C)P(A|\neg C)} \\
&= \frac{0.50 \times 0.05}{0.05 \times 0.50 + 0.95 \times 0.10} \approx 0.208
\end{aligned}
$$

Naïve Bayes Classifier: Bayes' Theorem

- That means that the probability of the email being a spam given that it contains the phrase "you are a winner" is about 20.8%
- Without any knowledge of occurrence of the phrase, the probability of the email being a spam is only 5%
- The probability of of the email being labelled a spam (Y) increases after incorporating the feature variable of phrase occurrence (X)

Naïve Bayes Classifier: Bayes' Theorem

- Note that in the previous lectures on classification methods, we often use more than one feature variable X in making predictions.
- For example, the occurrence of the phrase "transfer bank account" can be another feature in predicting spam.
- The more general form of Bayes' theorem allows us to incorporate multiple feature variables or attributes.

Naïve Bayes Classifier: Bayes' Theorem

- Suppose the categorical outcome variable Y takes on values in the set $\{y_1, y_2, ..., y_k\}$. For example, binary Y takes on values in $\{0, 1\}$.
- A more general form of Bayes' theorem assigns a classified label to an object with m feature variables $X = \{X_1, X_2, ..., X_m\}$ such that the predicted label corresponds to the largest value of $P(Y = y_j|X)$, $j = 1, 2, ..., k$.
- The value $P(Y = y_j|X)$ is given by

  for each value of Y,

$$P(Y = y_j|X) = \frac{P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m|Y = y_j) \times P(Y = y_j)}{P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m)},$$

for $j = 1, 2, ..., k$

for k values of y:

calculate max{P(Y=y(i)|X)}

binary

P(Y=0|X), P(Y=1|X)

find max{P(Y=0|X), P(Y=1|X)}

- Note by Bayes' theorem,

$$P(Y = y_j|X) = \frac{P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m|Y = y_j) \times P(Y = y_j)}{P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m)},$$

for $j = 1, 2, ..., k$

- With two simplifications, Bayes' theorem can be extended to become a naïve Bayes classifier.

Naïve Bayes Classifier: Bayes' Theorem

- The first simplification is to use the conditional independence assumption which simplifies the computation of the numerator term,

$$P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m | Y = y_j)$$
$$= P(X_1 = x_1 | Y = y_j) P(X_2 = x_2 | Y = y_j) ... P(X_m = x_m | Y = y_j)$$
$$= \prod_{i=1}^{m} P(X_i = x_i | Y = y_j).$$

Naïve Bayes Classifier: Bayes' Theorem

- The second simplification is to ignore the term in the denominator,

$$P(X_1 = x_1, X_2 = x_2, ..., X_m = x_m)$$

since it is constant for all values in the set $\{y_1, y_2, ..., y_k\}$ of the outcome variable Y.

because we just need to compare the P(Y=y(i)|X) for i=1->k

- With these two simplifications,

$$P(Y = y_j|X) \propto P(Y = y_j) \times \prod_{i=1}^{m} P(X_i = x_i|Y = y_j),$$

for $j = 1, 2, ..., k$. The symbol $\propto$ means "proportional to".

Naïve Bayes Classifier: Bayes' Theorem

- A Heuristic explanation for ignoring the term in the denominator:
- Suppose we wish to compare $\frac{a}{c}$ versus $\frac{b}{c}$.
- Since c is constant, we just need to compare a versus b.
- Let us look at a simple example for naïve Bayes classifier

Naïve Bayes Classifier: Example

- Suppose we wish to predict the class of a fruit Y that takes on the values $\{banana, orange, other\}$.
- The binary feature variables X are whether the fruit is long, sweet and yellow.
- The tabulation on 1000 pieces of fruit is as follows:

Y	Long	Sweet	Yellow
Banana	200	100	200
Orange	20	100	180
Other	100	50	50

Naïve Bayes Classifier: Example

Y	Long	Sweet	Yellow
Banana	200	100	200
Orange	20	100	180
Other	100	50	50

- We can first compute $P(Y = Banana) = \frac{200+100+200}{1000} = 0.5$, $P(Y = Orange) = \frac{20+100+180}{1000} = 0.3$ and $P(Y = Other) = \frac{100+50+50}{1000} = 0.2$

Naïve Bayes Classifier: Example

Y	Long	Sweet	Yellow	
Banana	200	100	200	500
Orange	20	100	180	300
Other	100	50	50	200

- We can then compute the conditional probabilities

i	x_i	$P(x_i \mid Y = Banana)$	$P(x_i \mid Y = Orange)$	$P(x_i \mid Y = Others)$
1	Long	$\frac{200}{500}$	$\frac{20}{300}$	$\frac{100}{200}$
2	Sweet	$\frac{100}{500}$	$\frac{100}{300}$	$\frac{50}{200}$
3	Yellow	$\frac{200}{500}$	$\frac{180}{300}$	$\frac{50}{200}$

Naïve Bayes Classifier: Example

- Suppose we want to predict the identity for a new piece of fruit which is long, sweet but not yellow, then

$$
\begin{aligned}
& P(Y = Banana|X) \\
& \propto P(Y = Banana) \times P(X_1 = Long|Y = Banana) \\
& \times P(X_2 = Sweet|Y = Banana) \times P(X_3 = \neg Yellow|Y = Banana) \\
& = 0.5 \times \frac{200}{500} \times \frac{100}{500} \times \left(1 - \frac{200}{500}\right) \\
& = 0.024
\end{aligned}
$$

Naïve Bayes Classifier: Example

- We can similarly calculate

$$
\begin{aligned}
&P(Y = Orange|X) \\
&\propto P(Y = Orange) \times P(X_1 = Long|Y = Orange) \\
&\times P(X_2 = Sweet|Y = Orange) \times P(X_3 = \neg Yellow|Y = Orange) \\
&= 0.3 \times \frac{20}{300} \times \frac{100}{300} \times \left(1 - \frac{180}{300}\right) \\
&\approx 0.0027
\end{aligned}
$$

$$
\begin{aligned}
&P(Y = Others|X) \\
&\propto P(Y = Others) \times P(X_1 = Long|Y = Others) \\
&\times P(X_2 = Sweet|Y = Others) \times P(X_3 = \neg Yellow|Y = Others) \\
&= 0.2 \times \frac{100}{200} \times \frac{50}{200} \times \left(1 - \frac{50}{200}\right) \\
&\approx 0.0188
\end{aligned}
$$

Naïve Bayes Classifier: Example

- Since the maximum probability score is $P(Y = Banana|X) = 0.024$, we predict the fruit to be a banana.

Naïve Bayes Classifier: Example

- When looking at problems with a large number of feature values, or attributes with a high number of levels, the probability score values can become very small in magnitude (close to zero).
- This is the problem of numerical underflow, caused by multiplying several probability values that are close to zero.
- A way to alleviate the problem is to compute the logarithm of the probability scores:

$$\log P(Y = y_j) + \sum_{i=1}^{m} \log P(X_i = x_i | Y = y_j),$$

for $j = 1, 2, ..., k$.

Naïve Bayes Classifier: Example in R

- We will use the `naiveBayes` function in the R package 'e1071' for fitting the naïve Bayes Classifier in our fruit example
- The dataset 'fruit.csv' has been posted to the course website under `LectureNotes/DataSets/`

Naïve Bayes Classifier: Example in R

- We will use the `naiveBayes` function in the R package 'e1071' for fitting the naïve Bayes Classifier in our fruit example
- The dataset 'fruit.csv' has been posted to the course website under `LectureNotes/DataSets/`

Naïve Bayes Classifier: Example in R

- Read in dataset 'fruit.csv'
-

```
> fruit.dat= read.csv("fruit.csv")
> fruit.dat=fruit.dat[,-1]
> fruit.dat<- data.frame(lapply(fruit.dat, as.
    factor))
> head(fruit.dat)
   Fruit Long Sweet Yellow
1 Banana    1     1      1
2 Banana    1     1      1
3 Banana    1     1      1
4 Banana    1     1      1
5 Banana    1     1      1
6 Banana    1     1      1
```

change to level

Naïve Bayes Classifier: Example in R

- Load the R package 'e1071' and fit naïve Bayes Classifier
- ```
  library(e1071)

  model <- naiveBayes(Fruit ~ Long+Yellow+Sweet,
  fruit.dat)
  ```
  ```

Naïve Bayes Classifier: Example in R

- Predict the identity for a new piece of fruit which is long, sweet but not yellow

```
> newdata <- data.frame(Long=1,Sweet=1, Yellow=0)
> newdata <- data.frame(lapply(newdata, as.factor))
>
> results <- predict (model,newdata,"raw")
> results
        Banana    Orange    Other
[1,] 0.5284404 0.0587156 0.412844
```

conditional probability (banana | new_data)

```
>
> results <- predict (model,newdata,"class")
> results
[1] Banana
```

if there are many levels

```
Levels: Banana Orange Other
```