KU x depa
MACHINE LEARNING
Training Course 2022

Naïve Bayes

DR. SETHAVIDH GERTPHOL

KU KASETSART UNIVERSITY depa

Today's Outline

- Naïve Bayes
- Bernoulli Naïve Bayes Example
- Using Sklearn

Naïve Bayes

- ใช้หลักการทางหลักสถิติเพื่อคำนวณความน่าจะเป็นที่ตัวอย่างใหม่จะอยู่ในแต่ละ Class จากโอกาสที่ feature แต่ละตัวจะปรากฏใน Class นั้น และจึงเลือกผลลัพธ์ที่มีความน่าจะเป็นสูงที่สุด
- สมมติฐาน: feature แต่ละตัวเป็นอิสระต่อกัน (ไม่เกี่ยวข้องกัน)
- คล้ายกับการ"เหมารวม"ของคน คือใช้"ตัวอย่างเก่า"กับ"ความน่าจะเป็นที่องค์ประกอบจะเกิดขึ้นในตัวอย่างเก่า" เพื่อคาดการตัวอย่างใหม่เมื่อเห็นองค์ประกอบเดิมเกิดขึ้นอีก
    - เช่น เราสังเกตว่าตอนฝนตกท้องฟ้าส่วนมากจะมืดครึ้มและมีฟ้าฝ่าบ้าง ในอนาคตถ้าท้องฟ้ามืดครึ้ม เราก็จะคาดการณ์ว่าโอกาสที่ฝนจะตกนั้นมีสูงกว่าถ้าท้องฟ้าไม่มืดแต่ฟ้าผ่าอย่างเดียว

ประเภทของ Naive Bayes

1. Gaussian Naive Bayes
   - ใช้ในกรณีที่ feature เป็นค่าตัวเลข เช่น ส่วนสูง น้ำหนัก
2. Multinomial Naive Bayes
   - ใช้ในกรณีที่ feature เป็นการนับ เช่นจำนวนคำที่ปรากฏในเอกสาร
3. Bernoulli Naive Bayes
   - ใช้ในกรณีที่ feature เป็นค่า category เช่นอยู่หรือไม่อยู่ใน EU, มีหรือไม่มีคำนี้ในเอกสาร

Multinomial กับ Bernoulli Naive Bayes มักใช้กับการค้นคืนเอกสาร

Bernoulli Naïve Bayes Example

- ทำนายกลุ่มของอุณหภูมิของแต่ละประเทศโดยพิจารณาจากประเทศนั้นอยู่ใน European Union หรือไม่ และประเทศนั้นติดกับแนวฝั่งทะเลหรือไม่
- **Features**
  - Coastline: [yes, no]
  - EU: [yes, no]
- **Labels**
  - Cold
  - Cool
  - Warm
  - Hot

Naïve Bayes Example (con't)

1. คำนวณความน่าจะเป็นของแต่ละผลเฉลยจากทุกตัวอย่างในชุดข้อมูลฝึก

Cold	0.18
Cool	0.38
Warm	0.24
Hot	0.20

1.0

Naïve Bayes Example (con't)

2. หลังจากได้ค่าความน่าจะเป็นของแต่ละผลเฉลยแล้วให้คำนวณค่าความน่าจะเป็นของแต่ละคุณลักษณะและแต่ละผลเฉลย

Cold (0.18)	coastline = yes	0.83
	coastline = no	0.17
	EU = yes	0.67
	EU = no	0.33

Warm (0.24)	coastline = yes	0.50
	coastline = no	0.50
	EU = yes	0.50
	EU = no	0.50

Cool (0.38)	coastline = yes	0.69
	coastline = no	0.31
	EU = yes	0.77
	EU = no	0.23

Hot (0.20)	coastline = yes	1.0
	coastline = no	0.0
	EU = yes	0.71
	EU = no	0.29

Naïve Bayes Example (con't)

3. คำนวณความน่าจะเป็นของตัวอย่างโดยนำค่าคุณลักษณะมาคูณกัน เพื่อให้ได้เป็นค่าความน่าจะเป็นของผลเฉลยนั้น

- New Instance
  - France, coastline=yes, EU=yes

Category	Prob.	coastline = yes	EU = yes	Score
Cold	0.18	0.83	0.67	0.10
Cool	0.38	0.69	0.77	0.20
Warm	0.24	0.50	0.50	0.06
Hot	0.20	1.0	0.71	0.14

- New Instance
  - Serbia, coastline=no, EU=no

Category	Prob.	coastline = no	EU = no	Score
Cold	0.18	0.17	0.33	0.01
Cool	0.38	0.31	0.23	0.03
Warm	0.24	0.50	0.5	0.06
Hot	0.20	0.0	0.29	0.00

Sklearn

```
[1] from google.colab import drive
    drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

เลือกแต่คอลัมน์ที่เป็นหมวดหมู่

```
[6] import pandas as pd
```

```
[4] cols = ['education', 'marital status','occupation', 'race', 'sex']
```

```
[18] df = pd.read_csv("./drive/MyDrive/Datasets/01-census-income.csv")
     X_with_cat = df[ cols ]
     y = df[ 'label' ]

     df2 = pd.read_csv("/content/drive/MyDrive/Datasets/02-future-census.csv")
     unseen_X = df2[ cols ]
     unseen_y = df2[ ['label'] ]
```

One-hot encoding

```
[10] from sklearn.preprocessing import OneHotEncoder
     from sklearn.compose import make_column_transformer
```

```
[11] transformer = OneHotEncoder()
     X_transformed = transformer.fit_transform(X_with_cat)
```

```
[9]  transformer.get_feature_names_out()
```

```
array(['education_ 10th', 'education_ 11th', 'education_ 12th',
       'education_ 1st-4th', 'education_ 5th-6th', 'education_ 7th-8th',
       'education_ 9th', 'education_ Assoc-acdm', 'education_ Assoc-voc',
       'education_ Bachelors', 'education_ Doctorate',
       'education_ HS-grad', 'education_ Masters', 'education_ Preschool',
       'education_ Prof-school', 'education_ Some-college',
       'marital status_ Divorced', 'marital status_ Married-AF-spouse',
       'marital status_ Married-civ-spouse',
       'marital status_ Married-spouse-absent',
       'marital status_ Never-married', 'marital status_ Separated',
       'marital status_ Widowed', 'occupation_ ?',
       'occupation_ Adm-clerical', 'occupation_ Armed-Forces',
```

BernoulliNB

import

```
[12] from sklearn.naive_bayes import BernoulliNB
```

```
[20] bnb = BernoulliNB()
     bnb.fit(X_transformed, y)

     unseen_X_transformed = transformer.transform(unseen_X)
     bnb.score(unseen_X_transformed, unseen_y)
```

```
     0.7923832923832924
```

Transform test data

Test accuracy

```
[21] bnb.score(X_transformed, y)
```

```
     0.7936175237645392
```

Train accuracy

Prediction Probability

```
[24] bnb.predict_proba(unseen_X_transformed)
```

```
array([[0.99697579, 0.00302421],
       [0.02825647, 0.97174353],
       [0.99339486, 0.00660514],
       ...,
       [0.14315683, 0.85684317],
       [0.40913915, 0.59086085],
       [0.91397258, 0.08602742]])
```

```
[25] bnb.classes_
```

```
array(['No', 'Yes'], dtype='<U3')
```

```
[26] unseen_y
```

	label
0	No
1	Yes
2	No

ข้อดีข้อเสียของ Naive Bayes

ข้อดี

- คำนวณความน่าจะเป็นของ feature ต่างๆ ได้ง่าย ทำให้ทำงานกับจำนวน feature ที่เยอะได้
- มักใช้เป็น baseline เพื่อเปรียบเทียบกับวิธีอื่นที่ซับซ้อนกว่า

ข้อเสีย

- สมมติฐานว่า feature แต่ละตัวไม่เกี่ยวข้องกัน นั้นยากที่จะเป็นจริงกับทุกคู่ feature
- ประสิทธิภาพในการ generalize มักจะไม่ค่อยดี
- เซ้นซิทิฟต่อสัดส่วนตัวอย่างของ Class แต่ละ Class ไม่เท่ากัน จึงควรทำให้จำนวนตัวอย่างของแต่ละ Class เท่ากันก่อน
- ถ้า probability ของ feature ใน class ใดเป็น 0 จะทำให้โมเดลไม่มีทางทำนายเป็น class นั้นได้