# Лабораторна робота №4

## Наївний Байєс в Python

***Мета роботи:*** набути навичок працювати з даними і опонувати роботу у Python з використанням теореми Байэса.

***Література***

*Supervised learning -* [*https://scikit-learn.org/stable/user\_guide.html*](https://scikit-learn.org/stable/user_guide.html)

*Naive Bayes Tutorial: Naive Bayes Classifier in Python -*

[*https://dzone.com/articles/naive-bayes-tutorial-naive-bayes-classifier-in-pyt*](https://dzone.com/articles/naive-bayes-tutorial-naive-bayes-classifier-in-pyt)

Naive Bayes classifier *-* [*https://en.wikipedia.org/wiki/Naive\_Bayes\_classifier*](https://en.wikipedia.org/wiki/Naive_Bayes_classifier)

### Зміст роботи

#### **Завдання 1.** Розібрати приклад прогнозування даних з використанням Теореми Байеса

Опис Даних

Вхідні дані включають день, прогноз, вологість і вітрові умови. Останній стовпець (цільова змінна) - «Гра» (play) позначає можливість проведення матчу.



На основі погодних умов необхідно визначити, чи відбудеться матч. Для цього необхідно:

Крок 1. Перетворити набір даних в частотну таблицю (frequency table), використовуючи кожен атрибут набору даних.



Крок 2. Для кожної частотної таблиці створити таблицю правдоподібності (likelihood table), розрахувавши відповідні ймовірності. Наприклад, ймовірність хмарної погоди (overcast) становить 0,29, а ймовірність того, що матч відбудеться (yes) - 0,64.



* Ймовірність «**Yes**’» для «**Sunny**» є:

## P(c|x) = P(Yes|Sunny) = P(Sunny|Yes)\* P(Yes) / P(Sunny) = (0.3 x 0.71) /0.36 = 0.591

* Ймовірність «**No**» для «**Sunny**» складає:

### P(c|x) = P(No|Sunny) = P(Sunny|No)\* P(No) / P(Sunny) = (0.4 x 0.36) /0.36 = 0.40

Таким же чином нам потрібно створити таблицю правдоподібності і для інших атрибутів:



#### **Завдання 2.** Чи відбудеться матч при наступних значеннях:

*Outlook =* Rain (*дощ)*

Humidity (*Вологість) =* High (*висока)*

Wind (*Вітер) =* Weak (*Слабкий)*

Play (*Гра відбудеться) =?*

Ймовірність «**Yes**» в цей день = P(Outlook = Rain|Yes)\*P(Humidity= High|Yes)\* P(Wind= Weak|Yes)\*P(Yes)= 2/9 \* 3/9 \* 6/9 \* 9/14 = 0,0199

Імовірність негативної відповіді **«No»** в цей день = P(Outlook = Rain|No)\*P(Humidity= High|No)\* P(Wind= Weak|No)\*P(No)= 2/5 \* 4/5 \* 2/5 \*

5/14 = 0,0166

Тепер, коли ми нормалізуємо значення, ми отримуємо:

### P(Yes) = 0.0199 / (0.0199+ 0.0166) = 0.55 P(No) = 0.0166 / (0.0199+ 0.0166) = 0.45

Модель передбачає, що ймовірність 55%, що завтра буде гра.

 Використовуючи дані з пункту 2 визначити чи відбудеться матч при наступних погодних умовах:

|  |  |  |
| --- | --- | --- |
| Варіант  |  | Умова  |
| 1, 6, 11  | Outlook = Overcast Humidity = High Wind = Weak  |  | Перспектива = Похмуро Вологість = Висока Вітер = Слабкий  |
| 2, 7, 12  | Outlook = Overcast Humidity = High Wind = Strong  |  | Перспектива = Похмуро Вологість = Висока Вітер = Сильний  |
| 3, 8, 13  | Outlook = Sunny Humidity = High Wind = Weak  |  | Перспектива = Сонячно Вологість = Висока Вітер = Слабкий  |
| 4, 9, 14  | Outlook = Sunny Humidity = Normal Wind = Strong  |  | Перспектива = Сонячно Вологість = Нормальна Вітер = Сильний  |
| 5, 10, 15  | Outlook = Rain Humidity = High Wind = Strong  |  | Outlook = Дощ Вологість = Висока Вітер = Сильний  |

#### **Завдання 3.** Використавши файл з медичною інформацією med.csv (або довільну інформацію яка має подібну структуру, на ваш розсуд) сформуйте власну таблицю з 3ма параметрами в колонках подібну до таблиці футбольних матчів наприклад виберіть значення alco, smoke, active, cardio у вигляді колонок і передбачте значення gender для довільної комбінації значень колонок (1,1,1,1). Можете використати для виконання завдання певну кількість унікальних даних, наприклад 50 рядків по 25 для генедеру 1 і 0.

### Методичні рекомендації

*Теорема Байєса.*

У статистиці і теорії ймовірностей теорема Байєса описує ймовірність події, гуртуючись на попередньому знанні умов, які можуть бути пов'язані з подією, тобто служить способом визначення умовної ймовірності.

З огляду на гіпотезу (H) і доказ (E), теорема Байєса стверджує, що зв'язок між ймовірністю гіпотези до отримання доказу - P(H), і ймовірністю гіпотези після отримання підтвердження - P(H|Е), це:



З цієї причини P(H) називається *апріорною ймовірністю*, а P(H|E) називається *апостеріорною ймовірністю*. Коефіцієнт, який пов'язує P(H|E)/P(E), називається *відношенням правдоподібності*. Використовуючи ці терміни, теорема Байєса може бути перефразована наступним чином:

*«Апостеріорна ймовірність дорівнює попередньої ймовірності, помноженої на відношення правдоподібності».*

##### Приклад

Припустимо, у нас є колода карт, і ми хочемо з'ясувати ймовірність того, що обрана нами карта випадковим чином виявиться королем, враховуючи, що це лицьова карта. Для початку нам потрібно з'ясувати ймовірність:

* P (король) - 4/52, так як в колоді карт 4 короля.
* P (особа| король) дорівнює 1, так як всі королі є лицьовими картами.
* P (особа) одно 12/52, так як в масті з 3 карт 3 карти і всього 4 масті.

Тепер, склавши всі значення в рівнянні Байєса, отримуємо результат

1/3.



Класифікація та прогнозування - два найбільш важливих аспекти машинного навчання, а Naive Bayes - простий, але потужний алгоритм прогнозного моделювання.

Наївний байєсівський класифікатор обчислює ймовірність приналежності об'єкта до якогось класу. Ця ймовірність обчислюється з шансу, що якась подія відбудеться, з опорою на події, які вже відбулися. Кожен параметр об'єкта, що підлягає класифікації, вважається незалежним від інших параметрів.

*Типи наївного байєсівського класифікатора:*

* *поліноміальний*: тут вектори ознак представляють собою значення частотності, тобто частоту, з якою генеруються ті чи інші події за допомогою поліноміального розподілу. Це модель подій, зазвичай використовувана для класифікації документів;
* *Бернуллі*: в багатовимірної моделі подій Бернуллі характеристики є незалежними логічними значеннями (двійковими змінними), якими описуються вхідні дані. Подібно поліноміальний моделі, ця модель широко застосовується в задачах класифікації документів, де використовується не частотність терміну, а бінарні характеристики тієї, термінів що зустрічаються (зустрічається слово в документі так чи ні);
* *Гаусса*: передбачається, що безперервні значення всіх характеристик мають розподіл Гаусса (нормальний розподіл). При нанесенні на графік виходить дзвіноподібна крива.

Наївні байєсовські алгоритми часто використовуються при аналізі емоційного забарвлення текстів, фільтрації спаму, в рекомендаційних системах тощо. Вони легко і швидко впроваджуються, але їх найбільший недолік полягає в складності дотримання вимоги про незалежність предикторов.

*Фільтрація спаму.* Наївні байєсівській класифікатори є популярним статистичним методом фільтрації електронної пошти. Як правило, використовується пакет слів / функцій для ідентифікації спаму в електронній пошті - підхід, що часто використовується в класифікації тексту.

*RSS-канали.* Категоризація новин, наївний байєсівский класифікатор застосовується для класифікації новинного контенту на основі новинного коду. Компанії використовують веб-сканер для вилучення корисного тексту з HTML-сторінок новинних статей для створення повнотекстового RSS. Вміст кожної новинний статті маркується (класифікується).

*Прогноз погоди.* Використовується байєсівська модель для прогнозування погоди, де апостеріорні ймовірності використовуються для обчислення ймовірності кожної мітки класу для примірника вхідних даних, а отримана з максимальною ймовірністю вважається підсумковою.

*Медичний діагноз.* При роботі з медичними даними наївний байєсівский класифікатор враховує докази з багатьох атрибутів, щоб зробити остаточний прогноз, і дає прозорі пояснення своїх рішень, і тому він вважається одним з найбільш корисних класифікаторів для підтримки рішень лікарів.

*Приклад виконання лабораторної роботи.*

import numpy

import pandas

from sklearn import preprocessing from sklearn.naive\_bayes import GaussianNB

 outlook=['Sunny','Sunny','Overcast','Rain','Rain','Rain','Overcast','Sunny

','Sunny','Rain','Sunny','Overcast','Overcast','Rain']

humidity=['High','High','High','High','Normal','Normal','Normal','High','N ormal','Normal','Normal','High','Normal','High']

wind=['Weak','Strong','Weak','Weak','Weak','Strong','Strong','Weak','Weak' ,'Weak','Strong','Strong','Weak','Strong']

 play=['No','No','Yes','Yes','Yes','No','Yes','No','Yes','Yes','Yes','Yes', 'Yes','No']

 le = preprocessing.LabelEncoder() outlook\_encoded=le.fit\_transform(outlook) humidity\_encoded=le.fit\_transform(humidity) wind\_encoded=le.fit\_transform(wind) label=le.fit\_transform(play) print ("Outlook:",outlook\_encoded)

print ("Humidity:",humidity\_encoded) print ("Wind:",wind\_encoded) print ("Play:",label)

features=zip(outlook\_encoded,humidity\_encoded, wind\_encoded) features\_list = list(features)

 model = GaussianNB()

 model.fit(features\_list,label)

 feature\_combination = [[0, 0, 1]] Play\_predicted= model.predict(feature\_ combination)

 print ("For the feature combination ", feature\_combination, " Predicted Play Label : ", Play\_predicted)

## Контрольні запитання

1. Де застосовується наївний Байес?
2. Поясніть теорему Байеса?
3. Які типи наївного байєсівського класифікатора існують?