University of Diyala - College of Science

Computer Science Department

Advertisement Gaussian Naive Bayes Classifier

This research was presented to the Council of the College of Science - University of Diyala - Department of Computing as part of the requirements to get a bachelor's degree in Computer science

BY:

Ibrahim Talaat Kanaan

Alaa Saddam Musa

Omar Adnan Abdullah

Under the Supervision Of:

Dr. Dahir Abdulhadi Abdullah

1442 A.H. 2021 A.D.

ABSTRUCT

We know the importance of social media at the present time for marketing and selling products through advertising, as most people prefer buying and selling through social media because of the ease of presentation and speed of sale and without any fatigue. Therefore, our project is to build classifier based on the Gaussian Naive Bayes to use targeted advertising to select users who are likely to purchase a particular product. And it was applied to dataset is a set of users of a fictitious social network and some of their attributes. Attributes of the users are key in determining the outcome of their purchases, and hence a Gaussian Naive Bayes classifier is developed to improve the targeted advertising.

SUPERVISOR CERTIFICATION

I certify that the preparation of this project entitled

Advertisement Gaussian Naive Bayes Classifier

Prepared By

Ibrahim Talaat Kanaan

Alaa Saddam Musa

Omar Adnan Abdullah

was made under my supervision in the Department of Computer Science/College of Science/University of Diyala and it is part of the requirements for obtaining a Bachelor's degree in Computer Science

Signature :

Name :

Date :

الاهداء

إلى النور الذي ينير لي درب النجاح (أبي)

و يا من علمتني الصمود مهما تبدلت الظروف (أمي)

إلى من يضيئون لي الطريق ويساندوني ويتنازلون عن حقوقهم لإرضائي والعيش في هناء (أخوتي)

إلى من أنار لي الطريق وأمسك لي مشعل النور أستاذي الفاضل

د. ظاهر وجميع اساتذة قسم علوم الحاسبات،

إلى كل من أضاء بعلمه عقل غيره،

أو هدى بالجواب الصحيح حيره سائليه

فأظهر بسماحته تواضع العلماء

وبرحابته سماحه العارفين.

شكر وتقدير

اشكر الله العلي القدير الذي أنعم عليَّ بنعمة العقل والدين. القائل في محكم التنزيل "وَفَوْقَ كُلِّ ذِي عِلْمٍ عَلِيمٌ" سورة يوسف آية 76.... صدق الله العظيم .

وقال رسول الله (صلي الله عليه وسلم):"من صنع إليكم معروفاً فكافئوه, فإن لم تجدوا ما تكافئونه به فادعوا له حتى تروا أنكم كافأتموه" ..... (رواه أبو داوود) .

فبعد شكر المولى عز وجل ، المتفضل بجليل النعم ، وعظيم الجزاء.. يجدر بي أن أتقدم ببالغ الامتنان ، وجزيل العرفان إلى كل من وجهني ، وعلمني ، وأخذ بيدي في سبيل إنجاز هذا البحث .. وأخص بذلك مشرفي الذي قوم ، وتابع ، وصوب ، بحسن إرشاده لي في كل مراحل البحث ، والذي وجدت في توجيهاته حرص المعلم ، التي تؤتي ثمارها الطيبة بإذن الله ...

كما أحمل الشكر والعرفان لكل من أمدني بالعلم ، والمعرفة ، وأسدى ليَّ النصح ، والتوجيه ، وإلى ذلك الصرح العلمي الشامخ متمثلاً في جامعة ديالى ، وأخص بالذكر كلية العلوم ، قسم الحاسبات، والقائمين عليها ,كما أتوجه بالشكر إلى كل من ساندني بدعواته الصادقة ، أو تمنياته المخلصة

أشكرهم جميعاً وأتمنى من الله عز وجل أن يجعل ذلك في موازين حسناتهم.

Table of Contents

List of Figures

CHPTER 1

INTRODUCTION

1.1 Introduction

The effectiveness of targeting a small portion of customers for advertising has long been recognized by businesses [1]. Research in this area has attracted considerable attention, for two main reasons. First, the amount of product/service information available to customers is ever-increasing, and hence it is desirable to help customers wade through the information to find the product/service they want. Second, understanding the needs of current and potential customers is an essential part of customer relationship management. The ability to accurately and efficiently identify the needs of customers and subsequently advertise products/services that they will find desirable opens up new possibilities to increase the customer retention, growth, and profitability of a business [1].

Traditional approach to targeted advertising is to (manually) analyze a historical database of previous transactions and the features associated with the (potential) customers, possibly with the help of some statistical tools and identify a list of customers most likely to respond to the advertisement of the product. The first type of recommendation technique was called the content-based approach [2]. A content-based approach characterizes recommendable products by a set of content features, and represents users' interests by a similar feature set. Content-based approaches select target customers whose interests have a high degree of

similarity to the product's content profile. But The content-based approach is inappropriate for products whose content is not electronically available. Another type of recommendation technique was called the collaborative approach (or sometimes called the social-based approach) [2]. The collaborative approach looks for relevance among users by observing their ratings assigned to products in a training set of limited size. The "nearest-neighbor" users are those that exhibit the strongest relevance to the target user. These users then act as "recommendation partners" for the target user. But it still has limitations, including: failing to advertise newly introduced products that have yet to be rated by users.

In this research, we have built a Gaussian Naive Bayes classifier that, based on the information of the person on the access social media, can categorize that the advertisement is appropriate for this person or not.

1.2 Problem Statement

Today, with the advent of information techniques, especially communication systems (i.e., email, bulletin board, and messaging etc.) and contacting systems (i.e., MSN, ICQ, Orkut, LinkedIn, and Friendster… etc.), we are seeing the growth of computer mediated social networks. These computers mediated human activities leave a huge amount of records in behind. These datasets can be used to analyze the social networks and facilitate the automatic construction of a targeted

advertising system. Therefore, in this research, we investigate the problem of targeted advertisement.

1.3 Research Objective

The objective of this study is to is to build classifier based on the Gaussian Naive Bayes to use targeted advertising to select users who are likely to purchase a particular product.

1.4 Research Scope

This research focuses on improving advertisements on social networking sites by classifying people and choosing the right people to show them the advertisement.

1.5 Expected Contribution / Benefit

The contributions of this study are to provides a classification method to improve and facilitate the access of advertising for a product or something else to the desired user easily.

CHPTER 2

LITERATURE REVIEW

2.1 Review of Related Work

Modern recommendation techniques have their roots in information filtering [2], and they aim to filter out information that is not relevant or uninteresting to a given user. With the advent of the World Wide Web and the rapid growth of e-commerce, recommender systems have been applied to a wide spectrum of domains not limited to digital information products, such as news, web pages, and documents. Recommender systems have been shown to be suitable for suggesting a wide range of products and services, such as books, restaurants, dry cleaners, plumbers, physicians, lawyers, financial institutions, and real estate brokers [3]. A more comprehensive survey of various recommendation techniques is listed in the work. [4]

Two broad classes of recommendation approaches that are commonly used by current recommender systems are content-based filtering and collaborative filtering. Content-based filtering is typically applied to recommend items that have par sable content or description. Unlike content-based filtering, which considers the preferences of only a given user when making recommendations, collaborative filtering recommends items to a user by considering the preferences of other users. The preferences of a user for unrated items are predicted based on a combination of known ratings from other users. Due to the simplicity and effectiveness of collaborative filtering, as indicated in empirical studies [6] [5], it is by far the most popular approach used in current recommender systems

2.2 Bayes Theorem

Bayes Theorem can be used to calculate conditional probability. Being a powerful tool in the study of probability, it is also applied in Machine Learning.

The Formula For Bayes' Theorem Is

$$P(A|B) = \frac{P(A \bigcap B)}{P(B)} = \frac{P(A) \cdot P(B|A)}{P(B)}$$

where:

$P(A) =$ The probability of A occurring
$P(B) =$ The probability of B occurring
$P(A|B) =$ The probability of A given B
$P(B|A) =$ The probability of B given A
$P\left(A \bigcap B\right)) =$ The probability of both A and B occurring

Bayes Theorem has widespread usage in variety of domains.

2.3 Naive Bayes Classifier

Naive Bayes Classifiers are based on the Bayes Theorem. One assumption taken is the strong independence assumptions between the features. These classifiers assume that the value of a particular feature is independent of the value of any other feature. In a supervised learning situation, Naive Bayes Classifiers are trained very efficiently. Naive Bayed classifiers need a small training data to estimate the parameters needed for classification. Naive Bayes Classifiers have simple design and implementation and they can applied to many real life situations.

2.4 Gaussian Naive Bayes

A Gaussian Naive Bayes algorithm is a special type of NB algorithm. It's specifically used when the features have continuous values. It's also assumed that all the features are following a gaussian distribution i.e., normal distribution. When we work with statistics and specifically probabilities, Gaussian Naive Bayes law is one of the most popular and fascinating theorems to dive into. The Bayesian theorem when working with statistics will allow you to calculate the probability that an event will occur provided that you have prior knowledge and information related to the specific event.

When working with continuous data, an assumption often taken is that the continuous values associated with each class are distributed according to a normal (or Gaussian) distribution. The likelihood of the features is assumed to be:

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

Sometimes assume variance

- is independent of Y (i.e., σi),
- or independent of Xi (i.e., σk)
- or both (i.e., σ)

Gaussian Naive Bayes supports continuous valued features and models each as conforming to a Gaussian (normal) distribution.

An approach to create a simple model is to assume that the data is described by a Gaussian distribution with no co-variance (independent dimensions) between dimensions. This model can be fit by simply finding the mean and standard deviation of the points within each label, which is all what is needed to define such a distribution.

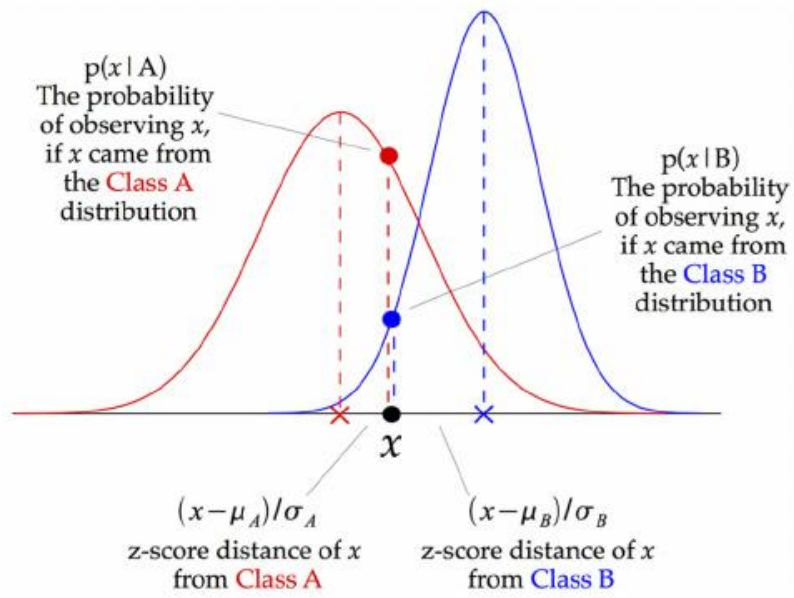

Figure 2.1. illustration of Gaussian Naive Bayes (GNB) classifier

The above illustration indicates how a Gaussian Naive Bayes (GNB) classifier works. At every data point, the z-score distance between that point and each class-mean is calculated, namely the distance from the class mean divided by the standard deviation of that class.

Thus, we see that the Gaussian Naive Bayes has a slightly different approach and can be used efficiently.

2.5 Python (programming language)

We will use Python to build the project. Python is an interpreted high-level general-purpose programming language. Python's design philosophy emphasizes code readability with its notable use of significant indentation. Its language constructs as well as its object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects[7] .

Python is dynamically-typed and garbage-collected. It supports multiple programming paradigms, including structured (particularly, procedural), object-oriented and functional programming. Python is often described as a "batteries included" language due to its comprehensive standard library[8] .

2.5.1 Scikit-learn

Scikit-learn (formerly scikits.learn and also known as sklearn) is a free software machine learning library for the Python programming language.[3] It features various classification, regression and clustering algorithms including support vector machines, random forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy.

CHPTER 3

SYSTEM IMPLEMENTATION

3.1 Project Data Set

We will implement he Gaussian Naive Bayes using Python and Scikit Learn. The data we will be working on for the exercise looks like this: -

	User ID	Gender	Age	EstimatedSalary	Purchased
0	15624510	Male	19	19000	0
1	15810944	Male	35	20000	0
2	15668575	Female	26	43000	0
3	15603246	Female	27	57000	0
4	15804002	Male	19	76000	0
5	15728773	Male	27	58000	0
6	15598044	Female	27	84000	0
7	15694829	Female	32	150000	1
8	15600575	Male	25	33000	0
9	15727311	Female	35	65000	0

Figure 3.1 project data set.

This is a database of various people based on their gender, age and estimated salary. The target is whether the person buys a product or not.

If the "Purchased" column has value "1", then it means that the person has bought the product, and if the value is "0", then it means that the person has not bought the product.

3.2 Project Implementation

In our project, using the Python language, we will import the necessary library in the first step of the project.

```
#Importing the libraries

import numpy as np

import matplotlib.pyplot as mtp

import pandas as pd
```

Now we import the data.

```
data= pd.read_csv('User.csv')

x = dataset.iloc[:, [2, 3]].values

y = dataset.iloc[:, 4].values
```

Now we do the step of splitting data into test and train sets.

```
#splitting the data into training and testing sets

from sklearn.model_selection import train_test_split

x_train, x_test, y_train, y_test = train_test_split(x, y, test_size =
0.25, random_state = 5)
```

Now, heading into Feature Scaling.

```
from sklearn.preprocessing import StandardScaler

sc = StandardScaler ()

x_train = sc.fit_transform(x_train)

x_test = sc.transform(x_test)
```

Now we applying the classifier model.

```
#Fitting Naive Bayes to the Training set

from sklearn.naive_bayes import GaussianNB

classifier = GaussianNB ()

classifier.fit(x_train, y_train)
```

```
#Predicting the Test set results
y_pred = classifier.predict(x_test)
```

```
#Accuracy score
from sklearn.metrics import accuracy_score
```

It gives a value of 0.89, hence an 89% accuracy. Next, we try to find the confusion matrix.

3.2.1 Confusion Matrix

is a performance measurement method for Machine learning classification. It helps you to the know the performance of the classification model on a set of test data for that the true values and false are known. It helps us find out, how many times our model has given correct or wrong output and of what type. Hence, it is a very important tool for evaluating classification models. It is illustrated in the following figure

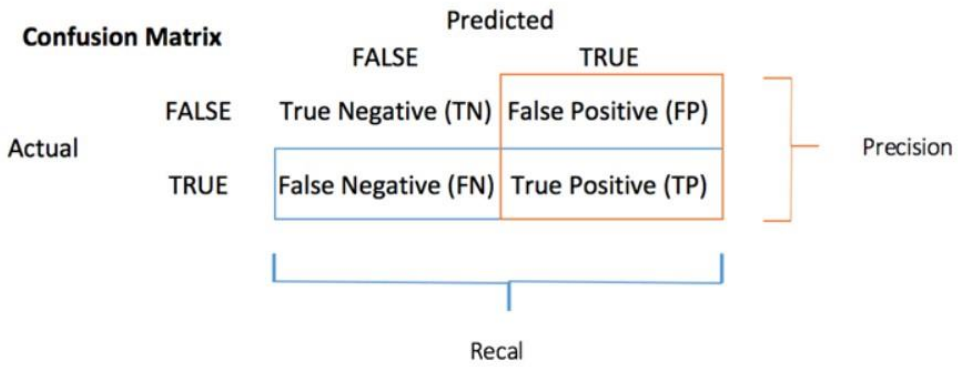

Figure 3.2 confusion matrix.

- **TP: True Positive**: Predicted values correctly predicted as actual positive.
- **FP: False Positive**: Predicted values incorrectly predicted an actual positive.
- **FN: False Negative**: Positive values predicted as negative.
- **TN: True Negative**: Predicted values correctly predicted as an actual negative.

We can compute the accuracy test from the confusion matrix:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

```
#Confusion Matrix

from sklearn.metrics import confusion_matrix

cm = confusion_matrix(y_test, y_pred)

print(cm)
```

Our result:

```
[[62  4]
 [ 7 27]]
```

Our result comes as given above. We can see that,

- TP: 27
- FP: 4
- FN: 7
- TN: 62

So, Accuracy= (89/100) = 0.89, Hence an 89% accuracy. After visualizing the results, we see.

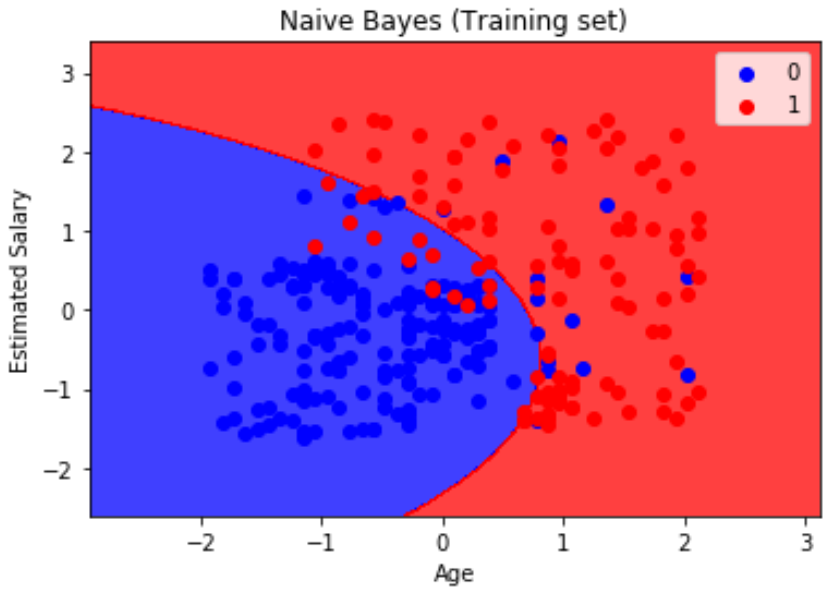

Figure 3.3 Gaussian Naive Bayes (Training set)

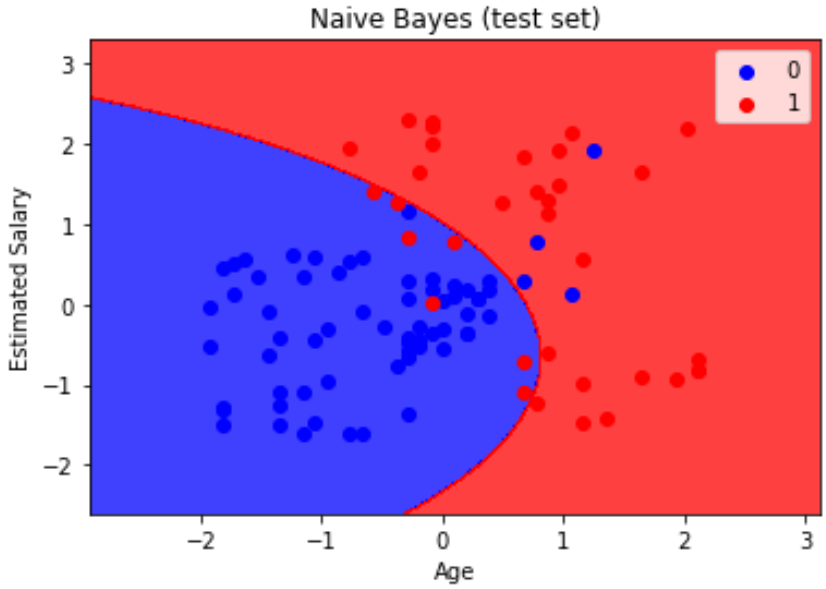

Figure 3.4 Gaussian Naive Bayes (Test set)

Hence, we can say that Gaussian Naive Bayes Classifier works well and can be used for a variety of Classification Problems.

CHPTER 4

CONCLUSION AND SUGGESTIONS

4.1 Conclusions

Generating targeted advertisements for online social networks is a problem of growing interest. Monetizing activity in online social networks has been the topic of heated discussion lately. We presented a project that, through the information of individuals, can deliver advertising to people who have more ability to purchase products in the advertisement. Based on Naive Bayes algorithm.

4.2 Suggestions

Our suggestions for future work are:-

- We aspire to add new features.
- We aspire to have a greater ability to treat different cases.

References

[1] Armstrong, G., and Kotler, P., “Marketing: an introduction,” Prentice-Hall, 1999. (Upper Saddle River, NJ).

[2] Cohen, J., et al, Special issue on information filtering, Communication of the ACM, 35(12), 1992. 26-81.

[3] Ansari, A., Essegaier, S., and Kohli, R., “Internet recommendation systems,” Journal of Marketing Research, 37(3), 2000. 363–375.

[4] Wellman, B. 1996, “For a social network analysis of computer networks: A sociological perspective on collaborative work and virtual community,” Proceedings of The ACM SIGCPR/SIGMIS conference, Denver, USA, 1996. 11-13.

[5] Pazzani, M. J., “A framework for collaborative, content-based and demographic filtering,” Artificial Intelligence Review, 13(5–6), 1999. 393–408.

[6] Mooney, R. J., and Roy, L., “Content-based book recommending using learning for text categorization,” ACM Conference on Digital Libraries, San Antonio, USA, 2000. 195–240.

[7] Kuhlman, Dave. "A Python Book: Beginning Python, Advanced Python, and Python Exercises". Section 1.1. Archived from the original (PDF) on 23 June 2012.

[8] "About Python". Python Software Foundation. Retrieved 24 April 2012.,

second section "Fans of Python use the phrase "batteries included" to describe the standard library, which covers everything from asynchronous processing to zip files."

[9] Fabian Pedregosa; Gaël Varoquaux; Alexandre Gramfort; Vincent Michel; Bertrand Thirion; Olivier Grisel; Mathieu Blondel; Peter Prettenhofer; Ron Weiss; Vincent Dubourg; Jake Vanderplas; Alexandre Passos; David Cournapeau; Matthieu Perrot; Édouard Duchesnay (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research. 12: 2825–2830.

((إقرار المشرف))

اشهد بأن اعداد هذا المشروع الموسوم

Advertisement Gaussian Naive Bayes Classifier

والمعد من قبل الطلاب :-

ابراهيم طلعت كنعان

علاء صدام موسى

عمر عدنان عبدالله

قد تم تحت أشرافي في قسم علوم الحاسوب / كلية العلوم/جامعة ديالى وهي جزء من متطلبات نيـــــــل شهادة البكالوريوس في اختصاص علوم الحاسوب

التوقيع:

الاسم:

المرتبة العلمية :

التاريخ :

الخلاصة

ندرك أهمية وسائل التواصل الاجتماعي في الوقت الحاضر لتسويق وبيع المنتجات من خلال الإعلانات ، حيث يفضل معظم الناس البيع والشراء عبر وسائل التواصل الاجتماعي لسهولة العرض وسرعة البيع وبدون أي تعب. لذلك ، فإن مشروعنا هو بناء مصنف يعتمد على Gaussian Naive Bayes لتحديد المستخدمين الذين من المحتمل أن يشتروا منتجًا معينًا. وقد تم تطبيقه على مجموعة البيانات وهي مجموعة من مستخدمي شبكة اجتماعية وهمية وبعض سماتهم. تعتبر سمات المستخدمين أساسية في تحديد نتيجة مشترياتهم ، وبالتالي تم تطوير مصنف Gaussian Naive Bayes لتحسين ايجاد المستهدف من الإعلان.

جامعة ديالى – كلية العلوم
قسم علوم الحاسبات

Advertisement Gaussian Naive Bayes Classifier

بحث مقدم الى مجلس كلية العلوم – جامعة ديالى – قسم الحاسبات كجزء من متطلبات الحصول على شهادة البكالوريوس في علوم الحاسوب

إعداد الطلاب

ابراهيم طلعت كنعان
علاء صدام موسى
عمر عدنان عبدالله

اشرف على البحث

د. ظاهر عبدالهادي عبدالله

1442هـ 2021م