Naive Bayes

Author: Rodriguez Noh Santiago Miguel
Professor: Ph.D. Anabel Martin Gonzalez
Link to code: https://github.com/Santiagomrn/Naive_Bayes

I. Introduction

A naive Bayesian classifier is a probabilistic method that has its bases in Bayes' theorem and is called naive given some additional simplifications that determine the hypothesis of independence of the predictor variables.

A. Theoretical framework

1) Bayas Teorem: In statistics and probability theory, Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. It serves as a way to figure out conditional probability.

Bayes' theorem is expressed by the following equation:

$$P(H|E) = \frac{P(E|H)P(H)}{P(E)} \tag{1}$$

P(H) is called the prior probability. $P(H|E)$ is called the posterior probability. $P(H|E)$ is called the likelihood.

2) Gaussian distribution: In statistics and probability it is called normal distribution, Gaussian distribution, Gaussian distribution or Laplace-Gauss distribution, to one of the probability distributions of continuous variable that more frequently appears in statistics and in probability theory.

Gaussian distribution is expressed by the following equation:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma_a} e^{-\frac{(x-\mu_a)^2}{2\sigma_a^2}}$$

where μ_a is the *sample mean*: $\mu_a = \frac{1}{|D_a|}\sum_{x\in D_a} x.a$
σ_a is the *sample standard deviation*, and
σ_a^2 the *sample variance*: $\sigma_a^2 = \frac{1}{|D_a|-1}\sum_{x\in D_a}(x.a-\mu_a)^2$

Fig. 1. Gaussian distribution equation.

II. Development of practice

Implement the Naive Bayes algorithm in Python or Matlab as follows data set:

	PRONOSTICO	TEMPERATURA	HUMEDAD	VIENTO	ASADO
0	soleado	36	alta	leve	no
1	soleado	28	alta	fuerte	no
2	nubaldo	30	alta	leve	si
3	lluvioso	20	alta	leve	si
4	lluvioso	2	normal	leve	si
5	lluvioso	5	normal	fuerte	no
6	nublado	11	normal	fuerte	si
7	soleado	22	alta	leve	no
8	soleado	9	normal	leve	si
9	lluvioso	17	normal	leve	si
10	soleado	19	normal	fuerte	si
11	nublado	22	alta	fuerte	si
12	nublado	27	normal	leve	si
13	lluvioso	21	alta	fuerte	no

Fig. 2. Data

to identify whether, according to the observed characteristics, a meat roast will be carried out that day or not. Asado (No = class 0, Yes = class 1) is the class to find. The other fields are the observed characteristics.

A. train

The first step was to obtain the likelihood tables of all the fields.

	no	si	total
soleado	0.6	0.222222	0.357143
nubaldo	0.0	0.111111	0.071429
lluvioso	0.4	0.333333	0.357143
nublado	0.0	0.333333	0.214286

Fig. 3. Pronostico likelihood table.

This is applicable to all fields except the temperature field. This is because temperature is a continuous variable unlike the other fields that are categorical variables.

For this reason, use the Gaussian distribution to obtain the probability of this variable, for this the following is obtained:

```
no   [36, 28, 5, 22, 21]
no   {'variance': 130.29999999999998, 'mean': 22.4}
si   [30, 20, 2, 11, 9, 17, 19, 22, 27]
si   {'variance': 78.77777777777779, 'mean': 17.444444444444443}
```

Fig. 4. mean and variance of both sets.

from this the following graphs were obtained.

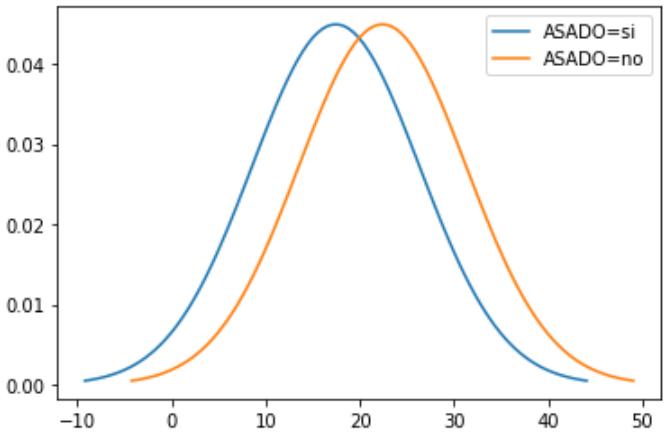

Fig. 5. Normal distributions.

B. Predictions

Once the classifier has been generated, perform the classification of the following instances:

PRONÓSTICO	TEMPERATURA	HUMEDAD	VIENTO
soleado	19	normal	leve
lluvioso	34	alta	leve
nublado	14	normal	fuerte

Fig. 6. Instances to prediction.

```
['soleado', 19, 'normal', 'leve']
ASADO = no   0.1693925266189205
ASADO = si   0.8306074733810795

['lluvioso', 34, 'alta', 'leve']
ASADO = no   0.7172465769160923
ASADO = si   0.2827534230839078

['nublado', 14, 'normal', 'fuerte']
ASADO = no   0.0
ASADO = si   1.0
```

Fig. 7. Results of prediction.

III. FINAL COMMENTS

Unlike other algorithms, this one had a very fast training. Something that I learned from this practice is about handling continuous variables and how the Gaussian distribution helps with that problem.