The NAIVE Bayes

Tim Analisis
Oct 27 2016

KMKLabs - Senayan City

REFRESH .....

Unsupervised / Clustering	Supervised / Classification
1. K-Means 2. HCS clustering 3. Canopy 4. DBSCAN 5. Fuzzy Clustering 6. K-SVD 7. Pitman More than 100 approaches	1. Naive Bayes 2. Linear Regression 3. Decision Tree 4. Random Forest 5. Support Vector Machine 6. Neural Network 7. Deep Neural Network

Outline

1. Introduction
2. Bayes Theorem
3. Naive Bayes Classifier
4. Example of Calculation
5. Naive Bayes in Scikit Learn
6. Advantage and Disadvantage

1. Introduction

→ Is A family of simple probabilistic classifiers

→ Applying Bayes' theorem with strong (naive) independence assumptions

→ Has been studied extensively since the 1950's.

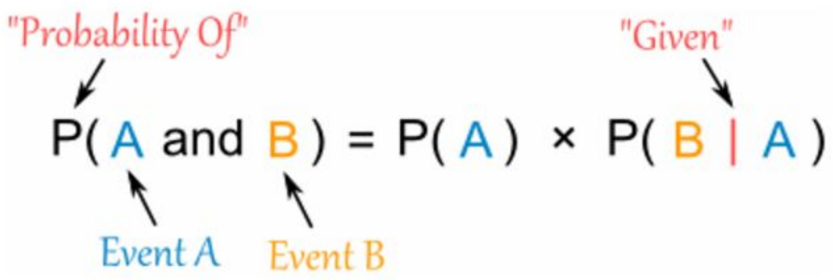
→ Remains a popular (baseline) method for text categorization,

Thomas Bayes
1702 - 1761

2. Bayes Theorem

→ Basically, bayes theorem defines the probability below:

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



Where:

- P(A) → probability to find A in population
- P(B) → probability to find B in population
- P(A|B) → peluang kejadian *A* bila *B* terjadi (Probability of Event A **given** Event B)
- P(B|A) → peluang kejadian *B* bila *A* terjadi (Probability of Event B **given** Event A)

2. Bayes Theorem

Example: Ice Cream

70% of your friends like Chocolate, and 35% like Chocolate AND like Strawberry.

What percent of those who like Chocolate also like Strawberry?

P(Strawberry|Chocolate) = P(Chocolate and Strawberry) / P(Chocolate)

$\rightarrow$ 0.35 / 0.7 = 50%

50% of your friends who like Chocolate also like Strawberry

3. Naive Bayes Classifier

→ Naive Bayes works as follow:

1. Let's say we have:
- Training set T
- Class set C where C = $\{C_1, C_2, .. C_k\}$
- Data X where X = $\{X_1, X_2, X_3, X_4, ... X_n\}$. and Xi = $[a_1, a_2, a_3, ... a_m]$

2. Given a new sample X, **X is predicted to belong to the class Ci if and only if**

$$P(C_i|\mathbf{X}) > P(C_j|\mathbf{X}) \qquad \text{for } 1 \le j \le m,\ j \ne i.$$

$$P(C_i|\mathbf{X}) = \frac{P(\mathbf{X}|C_i)\ P(C_i)}{P(\mathbf{X})}.$$

P(X|Ci) gimana kalo
X = $[a_1,a_2,,a_3,..a_n]$
???

$$P(\mathbf{X}|C_i) \approx \prod_{k=1}^{n} P(x_k|C_i).$$

4. Example of Calculation

Assume that we have two classes

c_1 = male, and c_2 = female.

We have a person whose sex we do not know, say "*drew*" or *d*.

Classifying *drew* as male or female is equivalent to asking is it more probable that *drew* is male or female, I.e which is greater $p(\text{male} \mid drew)$ or $p(\text{female} \mid drew)$

(Note: "Drew can be a male or female name")

Drew Barrymore

Drew Carey

4. Example of Calculation

What is the probability of being called "*drew*" given that you are a male?

What is the probability of being a male?

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

What is the probability of being named "*drew*"?
(actually irrelevant, since it is that same for all classes)

Officer Drew

This is Officer Drew (who arrested me in 1997). Is Officer Drew a Male or Female?

Luckily, we have a small database with names and sex.

We can use it to apply Bayes rule…

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Officer Drew

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(\text{male} \mid drew) = \frac{\mathbf{1/3 * 3/8}}{\mathbf{3/8}} = \frac{\mathbf{0.125}}{3/8}$$

$$p(\text{female} \mid drew) = \frac{\mathbf{2/5 * 5/8}}{\mathbf{3/8}} = \frac{\mathbf{0.250}}{3/8}$$

Officer Drew is more likely to be a Female.

Officer Drew

Officer Drew IS a female!

$$p(\text{male} \mid drew) = \frac{1/3 * 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 * 5/8}{3/8} = \frac{0.250}{3/8}$$

So far we have only considered Bayes Classification when we have one attribute (the "*antennae length*", or the "*name*"). But we may have many features.
How do we use all the features?

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	**Over 170CM**	**Eye**	**Hair length**	**Sex**
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

- To simplify the task, **naïve Bayesian classifiers** assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * \ldots * p(d_n|c_j)$$

The probability of class c_j generating instance d, equals....

The probability of class c_j generating the observed value for feature 1, multiplied by..

The probability of class c_j generating the observed value for feature 2, multiplied by..

- To simplify the task, **naïve Bayesian classifiers** assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * \ldots. * p(d_n|c_j)$$

$$p(\text{officer drew}|c_j) = p(\text{over\_170}_{\text{cm}} = \text{yes}|c_j) * p(\text{eye} = blue|c_j) * \ldots.$$

Officer Drew is blue-eyed, over 170_{cm} tall, and has long hair

$$p(\text{officer drew}|\ \text{Female}) = 2/5 \ * \ 3/5 \ * \ \ldots.$$

$$p(\text{officer drew}|\ \text{Male}) = 2/3 \ * \ 2/3 \ * \ \ldots.$$

DAN SETERUSNYA …………..

5. Naive Bayes in Scikit learn → Gaussian

```
>>> import numpy as np
>>> X = np.array([[-1, -1], [-2, -1], [-3, -2], [1, 1], [2, 1], [3,
2]])
>>> Y = np.array([1, 1, 1, 2, 2, 2])
>>> from sklearn.naive_bayes import GaussianNB
>>> clf = GaussianNB()
>>> clf.fit(X, Y)
GaussianNB(priors=None)
>>> print(clf.predict([[-0.8, -1]]))
[1]
>>> clf_pf = GaussianNB()
>>> clf_pf.partial_fit(X, Y, np.unique(Y))
GaussianNB(priors=None)
>>> print(clf_pf.predict([[-0.8, -1]]))
[1]
```

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

6. Advantage and Disadvantage

1. Advantage
   - Fast to train (just one scan)→ Spam Email Google use Naive Bayes
   - Fast to classify
   - Not sensitive to Irrelevant feature
   - Handle streaming data well

2. Disadvantage
   - Assume independence feature
   - problem happens when we are drawing samples from a population and the drawn vectors are not fully representative of the population.
   - need a big data set in order to make reliable estimations of the probability of each class

HOMEWORK :D

Given data:

Name	Over 170CM	Eye	Hair length	Sex
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

Please predict this data:

1. Drew, Yes, Brown, Short

Male or Female? Describe the calculation **OR** the python code used.