Naïve Bayes Classification

(Course: Introduction to Data Science)

Tirtharaj Dash

BITS Pilani, K.K. Birla Goa Campus

tirtharaj@goa.bits-pilani.ac.in

November 14, 2020

Introduction I

Let's look at some prediction tasks:

- Spam Classification: Given an email, predict whether it is a spam or not.
- Medical diagnosis: Given a list of symptoms, predict whether a patient has fever or not.
- Weather prediction: Based on temperature, humidity, etc. predict if it will rain tomorrow.

Introduction II

Naïve Bayes is an approach to model probabilistic relationship between attribute set ($\boldsymbol{x}$) and the class variable (y).

Introduction III

Let's define the problem:

- Given an instance defined by features $X_1, X_2, \ldots, X_d$
- Predict a label Y

(Notice that I have written the variables in CAPITAL letters. In a probabilistic setting, we are treating them to be random variables.)

Revisiting Probability I

- Conditional probability:

$$P(C|A) = \frac{P(A, C)}{P(A)}$$

- Bayes Theorem:

$$P(C|A) = \frac{P(A|C)P(C)}{P(A)}; P(A) \neq 0$$

The L.H.S is called the posterior, $P(A|C)$ is called the likelihood, $P(C)$ is called prior. The denominator is $P(A)$ is the marginal probability; and in most of the cases calculating this is fairly straightforward using the numerator.

Revisiting Probability II

Example:

1. - Let say: a doctor knows that common cold causes fever 50% of the time.
   - Let the prior probability of any patient having cold is $\frac{1}{10000}$
   - Let prior probability of any patient having fever is $\frac{1}{50}$

   If a patient has fever, what is the probability that the patient has cold?

Revisiting Probability III

Answer. Let C represents the random variable (r.v.) for cold, and F represents the r.v. for fever. We want to find $P(C|F)$ (this is read as: Probability of cold given fever)

$$P(C|F) = \frac{P(F|C)P(C)}{P(F)} = \frac{0.5 \times 1/10000}{1/50}$$

Bayesian Classification I

Given a data instance with attributes $X_1, \ldots, X_d$

- Goal is to predict a lass Y
- Specifically, we want to find out the value of Y that maximises $P(Y|X_1, \ldots, X_d)$

Can we estimate $P(Y|X_1, \ldots, X_d)$ directly from given data?

Bayesian Classification II

Let's use Bayes theorem:

$$P(Y|X_1, \ldots, X_d) = \frac{P(X_1, \ldots, X_d|Y)P(Y)}{P(X_1, \ldots, X_d)}$$

Now the difficulty is calculating the likelihood: $P(X_1, \ldots, X_d|Y)$.

Why? (Homework!)

Answer.
- We will run out of space and time.
- We will require very large amount of data.

Naïve Bayes Classification I

Naïve Bayes assumes conditional independence among attributes to calculate the likelihood.

This is the very reason 'Naïve Bayes' is called "naive" because it assumes that each input variable is conditionally independent.

However, this is a strong assumption and unrealistic for real data; however, the technique is very effective on a large range of complex problems.

But, conditional on what? (On Y)

Naïve Bayes Classification II

Therefore, the likelihood is calculated as:

$$P(X_1, \ldots, X_d|Y) = P(X_1|Y)P(X_2|Y)\ldots P(X_d|Y)$$

It is now easier to calculate the individual conditional probabilities $P(X_i|Y)$ from the given data.

For a given class the probability that a data point belongs to class $Y = y_k$ is given as:

$$P(Y = y_k|X_1, \ldots, X_d) \propto P(Y = y_k)P(X_1|Y = y_k)\ldots P(X_d|Y = y_k)$$

or,

$$P(Y = y_k|X_1, \ldots, X_d) \propto P(Y = y_k)\prod_{i=1}^{d} P(X_i|Y = y_k)$$

Naïve Bayes Classification III

Where did that big denominator vanish?

$$P(X_1, \ldots, X_d) = \sum_{k=1}^{K} P(X_1, \ldots, X_d | y_k)$$

As you can notice, this is a normalising constant term and it has no role in determining the predictive decision. Therefore, we can safely ignore it.

Naïve Bayes Classification IV

As you can notice: $P(X_i|Y)$ is nothing but a probability value. That is: it is $\frac{count}{total}$. That is:

$$P(X_i|Y) = \frac{N_{ik}}{N_k}$$

There could be situations where the count could be 0. In that case, the whole probability term will be 0. This can severely affect predictive decision.

To overcome this issue, we can use a Laplace estimate of probability:

$$P(X_i|Y) = \frac{N_{ik} + \alpha}{N_k + \alpha c}$$

where, c is the number of class labels; the parameter is an additive pseudo-count $\alpha > 0$. By default, we can just put $\alpha = 1$.

Problem solving I

2. Consider that you are given a task to filter incoming mails as spam or non-spam. You have a database of a set of mails with their class (i.e. spam or non-spam) where a set of words are used as feature to classify a mail to be one of these types. Let say the words be A, B, C and D; and the class is represented as S or NS.

A	B	C	D	Type
3	1	0	2	NS
2	0	1	1	NS
1	1	1	1	NS
4	1	1	0	NS
0	1	0	0	NS
0	2	5	0	S
1	3	4	4	S
2	0	4	5	S
1	0	0	8	S
4	1	0	7	S

Problem solving II

(a) For the given problem above, construct a Naïve-Bayes classification model by treat the occurrence of the words (A, B, C and D) as Bernoulli's trial.

(b) Using the classifier which you just modeled in question (a), classify a new email with word occurrences (0,2,6,0).

Problem solving III

Answer. (a) Bernoulli's trial as occurrence of the words means the following: if a word occurs (i.e. if number of occurrence is $>$ 0), then it is considered 1; otherwise 0. (1 means occurs, 0 means does not occur). With this transformation, our new dataset looks like this:

A	B	C	D	Type
1	1	0	1	NS
1	0	1	1	NS
1	1	1	1	NS
1	1	1	0	NS
0	1	0	0	NS
0	1	1	0	S
1	1	1	1	S
2	0	1	1	S
1	0	0	1	S
1	1	0	1	S

Problem solving IV

The probability estimates for NB are:

P(S)=5/10

P(A=1|NS)=4/5 P(A=0|NS)=1-P(A=1|NS)=1-4/5 (Optional)

P(B=1|NS)=4/5

P(C=1|NS)=3/5

P(D=1|NS)=3/5

Similarly,

P(A=1|S)=4/5 P(A=0|S)=1-P(A=1|S)=1-4/5 (Optional)

P(B=1|S)=3/5

P(C=1|S)=3/5

P(D=1|S)=4/5

Problem solving V

(b) Test the new email:

E-mail2(0,2,6,0)=E-mail2(0,1,1,0)

P(mail=NS|E-mail2)

= P(mail=NS|A=0,B=1,C=1,D=0)

= P(NS)×P(A=0|NS)×P(B=1|NS)×P(C=1|NS)×P(D=0|NS)

= P(NS)×[1-P(A=1|NS)]×P(B=1|NS)×P(C=1|NS)×[1-P(D=1|NS)]

= 0.0192

P(mail=S|Email2)

= P(mail=S|A=0,B=1,C=1,D=0)

= P(S)×P(A=0|S)×P(B=1|S)×P(C=1|S)×P(D=0|S)

= P(S)×[1-P(A=1|S)]×P(B=1|S)×P(C=1|S)×[1-P(D=1|S)]

= 0.0072

Problem solving VI

Since, $P(\text{NS}|\text{Email}) > P(\text{S}|\text{Email})$, it will be classified as non-spam.

Homework:

- Practice the above spam-classification problem using Laplace estimate of probabilities with $\alpha = 1$.
- Use Scikit-learn to model any real-world problem using Naïve Bayes.

Summary

- Very simple-yet-powerful probabilistic approach
- Has been shown to be remarkably effective for predictive decisions
- One of the most common approach to try given any new problem (to get a baseline predictive performance)