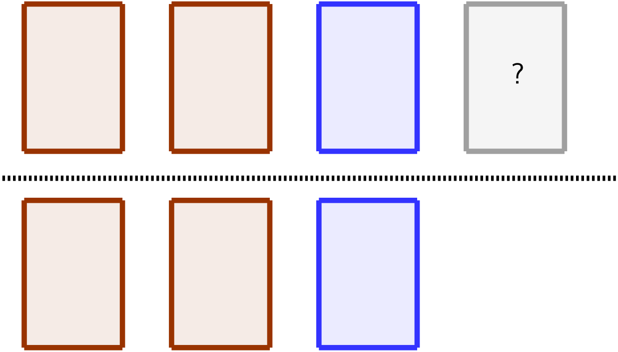
Naive bayes classifier

using python.

Iván Andrés Trujilllo Abella

ivantrujillo1229@gmail.com

Freund ilustration



Estimate Θ by maximun likelihood

The **Method of maximun likelihood** consist in estimate the parameter Θ the number of *red cards* that maximize the probability of see the **data** (three red cards and a blue car), in this case Θ could be 2 or 3.

$$\frac{\binom{3}{2}\binom{1}{1}}{\binom{4}{3}} > \frac{\binom{2}{2}\binom{2}{1}}{\binom{4}{3}} \tag{1}$$

Scipy

in Scipy the notation is:

$$p(k, M, n, N) = \frac{\binom{n}{k}\binom{M-n}{N-k}}{\binom{M}{N}} \tag{2}$$

```
from scipy.stats import hypergeom
k,M,n,N = 2,4,3,3
print(hypergeom.pmf(k,M,n,N))
k, M,n,N = 2,4,2,3
print(hypergeom.pmf(k,M,n,N))
```

Maximun Likelihood Estimation (MLE)

Suppose that you data it is generated by a theoretical distribution, the inverse problem is determine the most probable parameter that generate the data.

Laboratory

See lab (Click here)

Insights about MLE

we are going to say in a general term that $f(x_i)$ is PDF or PMF of a random variable.

Parameter estimation!

Problem

We are interested in known the population of turtles in a lake, the main restriction is that you can find all and you **capture** all and can't dry the lake.

How solve it?

Parameter estimation!

Problem

We are interested in known the population of turtles in a lake, the main restriction is that you can find all and you **capture** all and can't dry the lake.

How solve it?

Using two important tools:

- Sampling
- Maximun likelihood estimation in hypergeometric distribution for population

See Lincoln-Petersen(Click here)

MLE

...

Given a random sample $x_1, x_2, ..., x_n$ of independent and identically distributed (**iid**) random variables of the following *pmf* or *pdf* $f(x_i \mid \theta)$. The **likelihood** function is defined as:

$$L(\theta \mid data) = L(\theta \mid x_1, x_2, ..., x_n) \tag{3}$$

...

The likelihood function is the joint distribution of the data, therefore according to the **iid** assumption then:

$$L(\theta \mid x_1, x_2, ..., x_n) = \prod_{i=1}^{n} f(x_i \mid \theta) \tag{4}$$

MLE

the our estimator is

$$\hat{\theta}_{MLE} = \arg\max L(\theta \mid x_1, x_2 ... x_n) \tag{5}$$

Given that log is a **monotone** function sometimes is easier solve $\log L(\theta \mid x_1, x_2, ... x_n)$ it is important remark that some problems require numerical solutions.

Binomial example

Assume that you have:

MLE estimator

$$L(\theta \mid data) = \binom{n}{k}\theta^k(1-\theta)^{n-k} \quad (6)$$

Therefore:

$$\hat{\theta}_{MLE} = \frac{k}{n} \quad (7)$$

Parameter estimation!

All the problems consist in estimate a Unknown parameter with the information of a sample!

Parameter estimation!

All the problems consist in estimate a Unknown parameter with the information of a sample!

How describe a random variable?

- Mean
- Standard Deviation

The fundamental question is if we can get the population values from a sample?, Therefore we could try uses MLE to find the parameters.

$\hat{\sigma}^2_{MLE}$

Assume a sample of $x_1, x_2, ..., x_n \sim N(\mu, \sigma)$ therefore applying the MLE principle of maximizing the probability of get the data we need maximize

$$\mathbb{L} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{\sum (x_i - \mu)^2}{2\sigma^2}\right) \tag{8}$$

Maximizing the expression

First Order Condition (FOC)

$$\frac{\partial \mathbb{L}}{\partial \sigma^2} = 0 \tag{9}$$

$$\hat{\sigma}^2_{MLE} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{10}$$

Simulation

Benchmark of estimators

...

$$S_n^2 = \frac{\sum(x_i - \bar{x})^2}{n} \tag{11}$$

$$S^2 = \frac{\sum(x_i - \bar{x})^2}{n-1} \tag{12}$$

(Click here)

Probability and likelihood

Think in normal distribution $X \sim N(\sigma, \mu)$.

Probability

$$P(a < x < b \mid \mu, \sigma) \tag{13}$$

You calculate the probability given that the location and the distribution is defined.

Likelihood

$$L(\mu, \sigma \mid x) \tag{14}$$

A measure that quantify how much the parameters μ and σ describe the observed data x.

Disease diagnose and bayes

The sensitivity (The probability that test is positive given that the person really have the disease $P(+test \mid disease)$ for instance is 90% therefore if a person is positive test could be seen the life pass for eyes?, but we need is:

$$P(disase \mid +test) \tag{15}$$

Here appear Bayes theorem.

Bayes in diagnose

Causality

In a causal sense the proper order is **Cause produce effect**, note here the cronological order, notice that an effect is associated with multiple causes.

Bayes in diagnose

In a medical sense we observe different sympstoms and we need diagnose could have:

$$P(cause \mid effect) = \frac{P(effect \mid cause)P(cause)}{P(effect)} \tag{16}$$

Bayes in diagnose

we can calculate $P(disease \mid +test)$ with the following formula:

$$\frac{P(+test \mid disease)}{P(+test \mid disease)P(disease) + P(positive \mid disease^c)P(disease^c)} \tag{17}$$

The denominator is calculated with total law of probability.

Prosecutor fallacy

Innocent person have a probability of $\frac{1}{10.000}$ that the evidence be damning, namely:

$$P(evidence \mid inoccent) = \frac{1}{10.0000} \tag{18}$$

Bayes theorem in prosecutor fallacy

$$P(innocent \mid evidence) = \frac{P(evidence \mid inoccent)P(inoccent)}{P(evidence)} \tag{19}$$

Sally clark's history

- in 1996 born her first son and die few months later
- in 1997 born her second son and also die few months laters
- She was the last person with stay with both childrens.

Meadow (He was a pediatrician) argument was that the probability of a child die by sudden death is:

$$P(S) = \frac{1}{8573} \tag{20}$$

Sally clarks history

- Clark was convicted in 1999
- Clark was released in 2003.
- Clak died four years later by alcohol intoxication.

Sally clark's history

Remember that

$$P(A \cap B) = P(A)P(B) \tag{21}$$

and therefore:

$$\begin{aligned} P(A \mid B) &= \frac{P(A \cap B)}{P(B)} \\ P(A \mid B) &= \frac{P(A)P(B)}{P(B)} = P(A) \end{aligned} \tag{22}$$

This is very important because the events are not independent.
Given that, the second child die is more probable if the first die suddently (could be appear genetic predisposition), namely: $P(S_2 \mid S_1) > P(S_1)$.

Sally clark's history

Fallacy?

Asumming that sally is innoncent then the trial uses:

$$P(evidence \mid innocent) = \frac{1}{8573} \cdot \frac{1}{8573} \tag{23}$$

what is wrong? what formula you need?.

Bayes in ML lingo

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)} \tag{24}$$

we can descompose (24) as :

- $P(Y \mid X)$ posterior
- $P(Y)$ prior
- $P(X)$ evidence
- $P(X \mid Y)$ likelihood

Excersie

rolling dice

what is the probability of have a odd in a rolling dice if the number given that we had is equal or greater than four.

Spam detector

$$P(broma \mid spam) = b \tag{25}$$

b is the conditional probability, the probability that word **broma** is contained given belong to spam.

Naive Bayes classifier

Messages	**Category**
This message contain the words; a, b, c	No spam
This message contain the words; a, b	No spam
This message contain the words; a, c	No spam
This message contain the words; a, e	Spam
This message contain the words; b, d	Spam
This message contain e	Spam
This message contain d, f	Spam

Split data by category

Messages	Category
This message contain the words; a, e	Spam
This message contain the words; b, d	Spam
This message contain e	Spam
This message contain d, f	Spam

Table: Spam

Messages	Category
This message contain the words; a, b, c	No spam
This message contain the words; a, b	No spam
This message contain the words; a, c	No spam

Table: No spam

Conditional probability of the words according to the category

The **prior probability** of $word_a$ is:

$$P(word_a) = \frac{4}{7} \tag{26}$$

$$P(word_a \mid Spam) = \frac{1}{4} \tag{27}$$

$$P(word_a \mid Spam^c) = \frac{3}{3} \tag{28}$$

Notice, that the probability that the $word_a$ appear in a spam message is greater than in a not spam message. What are the probabilities for another words?.

Scoring

$$P(spam) \prod_{i}^{n} (word_i \mid spam) \tag{29}$$

according to (29) then we can have a problem if in the training data sets there are not a set of words that appear in another dataset then the conditional probabilities are equal to zero, to tackle this we could add α(integer) count to each category.

There are something behind

There is a bias in the order

Naive ignore the language! Whats means that Naive have low variance?

Naive Bayes in Bankruptcy

Sector	Income	Bankruptcy
Financial	High	No
Financial	Small	Yes
Agricultural	Small	Yes
Agricultural	High	Yes
Financial	Small	No
Financial	Small	No
Agricultural	High	No
Agriulcural	Small	Yes
Agricultural	Small	No

Table: Dataset Bankruptcy

Notice, that we want to uses this dataset to answer the question $P(Bankruptcy \mid Sector \cap Income) = P(Bankruptcy \mid Sector, Income)$.

Example

Naive Bayes Classifier

let's consider that we want guess the probability that a occur a **default** given that te firm belong to the **agricultural** sector and its size or income is **small**:

$$P(Banruptcy = Yes \mid Sector = Agricultural, Income = Small) \tag{30}$$

Therefore we could use **Bayes theorem** using (3) to estimate:

$$P(Sector = Agricultural, Income = Small \mid Bankruptcy = Yes) \tag{31}$$

Split data

Naive Bayes classifier

Sector	Income	Bankruptcy
Financial	Small	Yes
Agricultural	Small	Yes
Agricultural	High	Yes
Agricultural	Small	Yes

Table: Banruptcy = Yes

With this data we can get:

$$P(Sector = Agricultural, Income = Small \mid Bankruptcy = Yes) = \frac{2}{4} = \frac{1}{2}.$$

a lot of if!

Naive bayes classifier

Could be impractical find concrete combinations of values($X_1 = x_1, X_2 = x_2, ..., X_n = x_n$), with a great number of features(zeros could be appear). if Assume that the features are indepedent, we dont need be corncened by combinatios! and therefore:

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n \mid Y = y_0) = \prod_{i=1}^{n} P(X_i = x_i \mid Y = y_0)$$

arg max

Assume a function $f : X \to Y$ then $\arg\max_x$ is

$$\arg\max_x f(x) = \{x \mid \forall x' : f(x') \leq f(x)\} \tag{32}$$

in other words are set of values in domain that give us the maximun value in the function, what is the $\arg\max_x (1 - |x|)$?.

$$\max_x f(x) = \{f(x) \mid \forall f(x') : f(x') \leq f(x)\} \tag{33}$$

$$f\left(\arg\max_x f(x)\right) = \max_x f(x) \tag{34}$$

$$\arg\max_{y} P(y|X) \tag{35}$$

Pseudocode

```
Count the number of classes,
get the prior probabilities,
determine likelihoods table;
    how many times appear the feature (i) in each class.
```

Nb is the baseline model...

NB

NB could be:

- **Multinomial NB:**
  Allow us model multiple occurrences of the feauture for instance the number of times that occur a especific word.
- Binomial NB (Allow us model if the word occur or not).
- Gaussian NB

Gaussian Bayes Continued

Assume f features and n samples, therefore we

Rule chain

$$
\begin{aligned}
P(A \cap B \cap C \cap D) &= P(D \mid A \cap B \cap C)P(A \cap B \cap C) \\
&= P(D \mid A \cap B \cap C)P(C \mid A \cap B)P(A \cap B) \\
&= P(D \mid A \cap B \cap C)P(C \mid A \cap B)P(B \mid A)P(A) \\
&= P(A)P(B \mid A)P(C \mid A \cap B)P(D \mid A \cap B \cap C)
\end{aligned}
$$

Rule chain

Generalization

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_n \mid A_1 \cap \ldots \cap A_n)P(A_1 \cap \ldots \cap A_{n-1}) =$$
$$P(A_n \mid A_1 \cap \ldots \cap A_n)P(A_{n-1} \mid A_1 \cap \ldots \cap A_{n-2})P(A_1 \cap \ldots \cap A_{n-2})$$

Recursively

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = \prod_{i=1}^{n} P\left(A_i \middle| \bigcap_{j=1}^{i-1} A_j\right)$$