**1) Trình bày rõ mô hình và tham số mô hình, cách inference trong mô hình**

Naïve Bayes classcification is a [classification technique](https://courses.analyticsvidhya.com/courses/introduction-to-data-science-2/?utm_source=blog&utm_medium=6stepsnaivebayesarticle) based on Bayes’ Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature.

For example, a fruit may be considered to be an apple if it is red, round, and about 3 inches in diameter. Even if these features depend on each other or upon the existence of the other features, all of these properties independently contribute to the probability that this fruit is an apple and that is why it is known as ‘Naive’.

Naive Bayes model is easy to build and particularly useful for very large data sets. Along with simplicity, Naive Bayes is known to outperform even highly sophisticated classification methods.

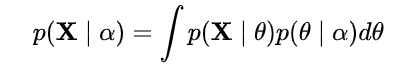
Bayes theorem provides a way of calculating posterior probability P(c|x) from P(c), P(x) and P(x|c). Look at the equation below:



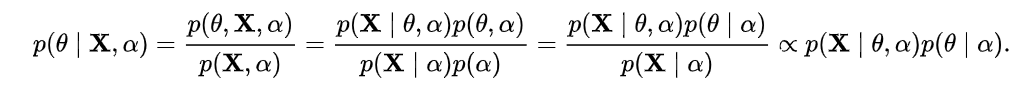
* *P*(*c|x*) is the posterior probability of *class* (c, *target*) given *predictor* (x, *attributes*).
* *P*(*c*) is the prior probability of *class*.
* *P*(*x|c*) is the likelihood which is the probability of *predictor* given *class*.
* *P*(*x*) is the prior probability of *predictor*.

**Bayesian inference** is a method of [statistical inference](https://en.wikipedia.org/wiki/Statistical_inference) in which [Bayes' theorem](https://en.wikipedia.org/wiki/Bayes%27_theorem) is used to update the probability for a hypothesis as more [evidence](https://en.wikipedia.org/wiki/Evidence) or [information](https://en.wikipedia.org/wiki/Information) becomes available. Bayesian inference is an important technique in [statistics](https://en.wikipedia.org/wiki/Statistics), and especially in [mathematical statistics](https://en.wikipedia.org/wiki/Mathematical_statistics). Bayesian updating is particularly important in the [dynamic analysis of a sequence of data](https://en.wikipedia.org/wiki/Sequential_analysis). Bayesian inference has found application in a wide range of activities, including [science](https://en.wikipedia.org/wiki/Science), [engineering](https://en.wikipedia.org/wiki/Engineering), [philosophy](https://en.wikipedia.org/wiki/Philosophy), [medicine](https://en.wikipedia.org/wiki/Medicine), [sport](https://en.wikipedia.org/wiki/Sport), and [law](https://en.wikipedia.org/wiki/Law). In the philosophy of [decision theory](https://en.wikipedia.org/wiki/Decision_theory), Bayesian inference is closely related to subjective probability, often called "[Bayesian probability](https://en.wikipedia.org/wiki/Bayesian_probability)".

* The [prior distribution](https://en.wikipedia.org/wiki/Prior_distribution) is the distribution of the parameter(s) before any data is observed, i.e.p( ){\displaystyle p(\theta \mid \alpha )}. The prior distribution might not be easily determined; in such a case, one possibility may be to use the [Jeffreys prior](https://en.wikipedia.org/wiki/Jeffreys_prior) to obtain a prior distribution before updating it with newer observations
* The [sampling distribution](https://en.wikipedia.org/wiki/Sampling_distribution) is the distribution of the observed data conditional on its parameters, i.e. p( ) {\displaystyle p(\mathbf {X} \mid \theta )}. This is also termed the [likelihood](https://en.wikipedia.org/wiki/Likelihood_function), especially when viewed as a function of the parameter(s), sometimes written {\displaystyle \operatorname {L} (\theta \mid \mathbf {X} )=p(\mathbf {X} \mid \theta )}L( = p( )
* The [marginal likelihood](https://en.wikipedia.org/wiki/Marginal_likelihood) (sometimes also termed the *evidence*) is the distribution of the observed data [marginalized](https://en.wikipedia.org/wiki/Marginal_distribution) over the parameter(s), i.e.



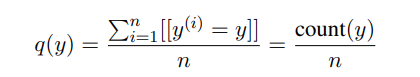
* The [posterior distribution](https://en.wikipedia.org/wiki/Posterior_distribution) is the distribution of the parameter(s) after taking into account the observed data. This is determined by [Bayes' rule](https://en.wikipedia.org/wiki/Bayes%27_rule), which forms the heart of Bayesian inference:



This is expressed in words as "posterior is proportional to likelihood times prior", or sometimes as "posterior = likelihood times prior, over evidence".

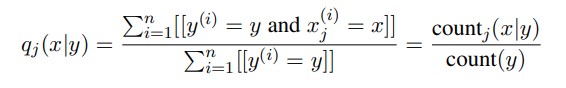
**2) Trình bày cách tính giá trị cho các tham số mô hình NB với giá trị rời rạc dựa vào Maximum Likelihood Estimation (mô hình mutinomial NB)**

We now consider how the parameters q(y) and qj (x|y) can be estimated from data. In particular, we will describe the maximum-likelihood estimates. We first state the form of the estimates, and then go into some detail about how the estimates are derived. Our training sample consists of examples (x (i) , y(i) ) for i = 1 . . . n. Recall that each x (i) is a d-dimensional vector. We write x (i) j for the value of the j’th component of x (i) ; x (i) j can take values −1 or +1. Given these definitions, the maximum-likelihood estimates for q(y) for y ∈ {1 . . . k} take the following form:

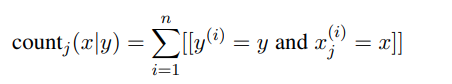


Here we define [[y (i) = y]] to be 1 if y (i) = y, 0 otherwise. Hence Pn i=1[[y (i) = y]] = count(y) is simply the number of times that the label y is seen in the training set.

Similarly, the ML estimates for the qj (x|y) parameters (for all y ∈ {1 . . . k}, for all x ∈ {−1, +1}, for all j ∈ {1 . . . d}) take the following form:



Where

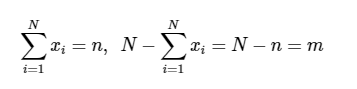


This is a very natural estimate: we simply count the number of times label y is seen in conjunction with xj taking value x; count the number of times the label y is seen in total; then take the ratio of these two terms.

Example 1:

Suppose toss a coin N times and get n faces. Find the probability that there is a head on the next coin toss.

Assume λ is the probability to get a head face. Let x 1 , x 2 , … , x N be the outputs received, where there are n values ​​of 1 corresponding to the head face and m = N − n zero values ​​corresponding to the tail face. We can immediately infer:



It can be seen whether getting a head or a tail when tossing a coin follows the Bernoulli distribution:



Then the model parameter λ can be evaluated by solving the optimization problem:

λ=argmaxλ[p(x1,x2,…,xN|λ)]

=argmaxλ[)]

=argmaxλ[N∏i=1λxi(1−λ)1−xi]

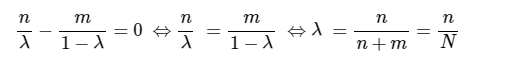
=argmaxλ[λ∑Ni=1xi(1−λ)N−∑Ni=1xi]

=argmaxλ[λn(1−λ)m] (\*)

=argmaxλ[nlog(λ)+mlog(1−λ)] (\*\*)

I assumed that the outcome of each coin toss was independent of each other. From ( \*) to ( \*\*) we have taken the log of the objective function.

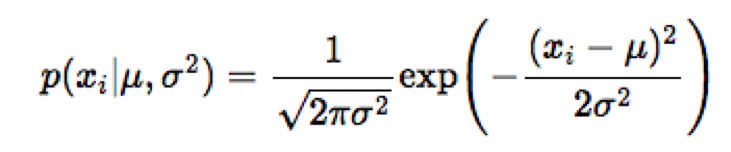
At this point, the optimization problem ( \*\*) can be solved by taking the derivative of the objective function to be 0. That is, λ is the solution of the equation:



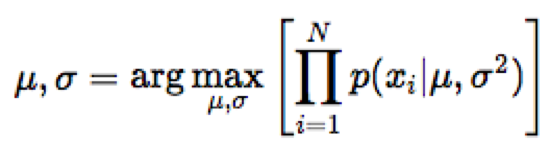
So the results we assert above are valid.

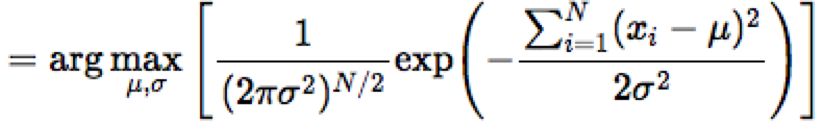
Example 2:

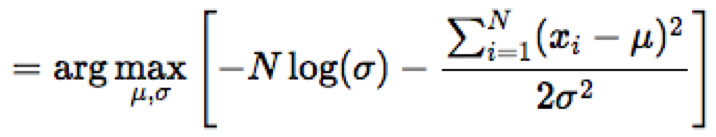
Suppose the problem is that there are 5 students taking the test with scores of 3, 6, 5, 9, 8 respectively. To model the scores of these students, we assume that the data points are segregated. distributed according to the Gaussian distribution:



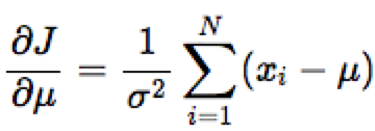
To predict the parameter set of the normal distribution, we use the MLE method:

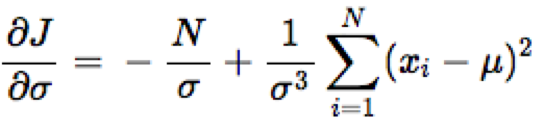




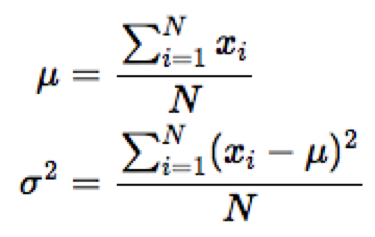


To find μ and σ such that the expression in square brackets reaches its maximum value, we derive the expression in terms of each variable and solve the equation when that value is zero.



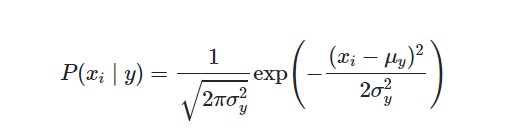


Then we have :



Substituting the data points into the above formula, we find μ = 6.2 and σ = 2.14.

**3) Trình bày các mô hình NB với biến liên tục: mô hình Gausian NB và mô hình Bernoulli NB**

Gausian NB  
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-

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.

Berboulli NB  
Bernouli Naive Bayes model is used for discrete data and it works on Bernoulli distribution. The main feature of Bernoulli Naive Bayes is that it accepts features only as binary values like true or false, yes or no, success or failure, 0 or 1 and so on. So when the feature values are binary we know that we have to use Bernoulli Naive Bayes classifier.  
Advantage of Bernouli NB

1. They are extremely fast as compared to other classification models
2. As in Bernoulli Naive Bayes each feature is treated independently with binary values only, it explicitly gives penalty to the model for non-occurrence of any of the features which are necessary for predicting the output y. And the other multinomial variant of Naive Bayes ignores this features instead of penalizing.
3. In case of small amount of data or small documents (for example in text classification), Bernoulli Naive Bayes gives more accurate and precise results as compared to other models.
4. It is fast and are able to make to make real-time predictions
5. It can handle irrelevant features nicely
6. Results are self explanatory

Disadvantage of Bernouli NB

1. Being a naïve (showing a lack of experience) classifier, it sometimes makes a strong assumption based on the shape of data
2. If at times the features are dependent on each other then Naive Bayes assumptions can affect the prediction and accuracy of the model and is sensitive to the given input data
3. If there is a categorial variable which is not present in training dataset, it results in zero frequency problem. This problem can be easily solved by Laplace estimation.