Bayesian vs. Frequentist Approach to Probability

PROF. NAVNEET GOYAL

CS & IS

BITS, PILANI

Importance of Beliefs

“If I hadn’t believed it,

I would never have seen it”

Anon – The Film (2018)

Mind’s Eye

The Ether

What we are interested in?

Medicine

p(+ve test | cancer) vs p(cancer | +ve test)

- **Justice**

p(DNA match | guilty) vs p(guilty | DNA match)

- **Perception**

p(retinal image | cube) vs p(cube | retinal image)

posterior = likelihood * prior / evidence

Bayes' Theorem Overview

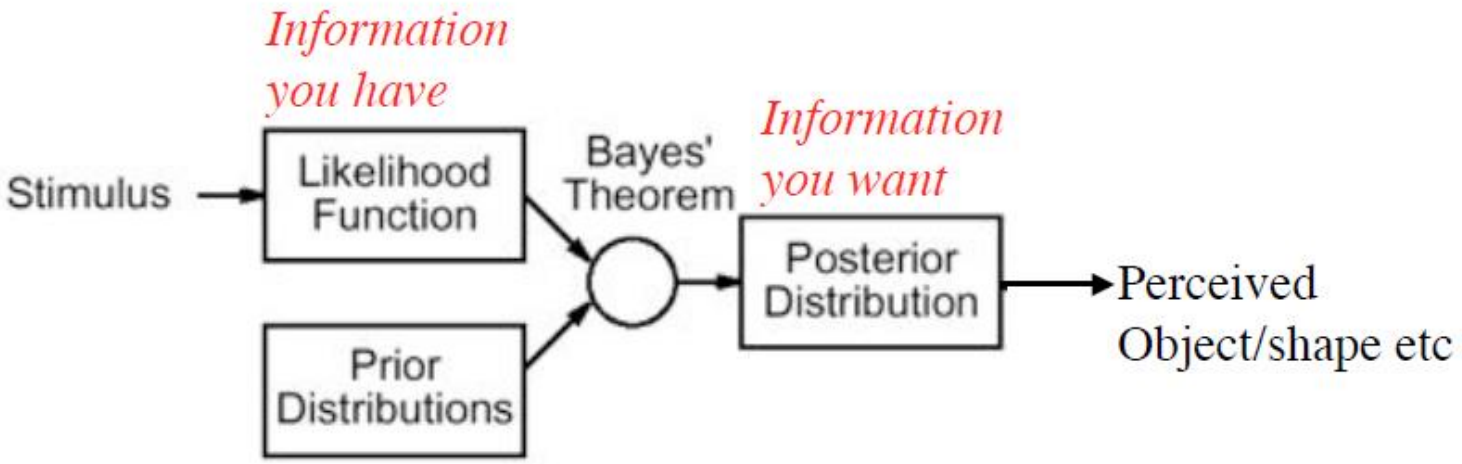
- There are two types of information:

information you **want** and information you **have**.

- Bayes rule *can be considered as* a method

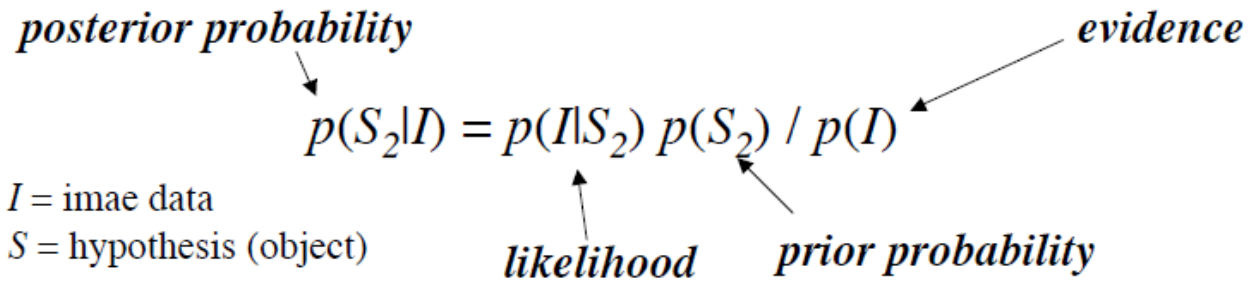
for obtaining the information you **want** from the information you **have**.

Bayes' Theorem Overview



Also known as ***Bayes' rule***.

Bayes' Theorem Overview



posterior probability:	the thing we **want to know** (the probability that the object is S_2 given the image I)
likelihood:	the thing we **already know** (probability of of the image I given object S_2)
prior probability:	the thing we know from **prior experience** (the probability that any object could be S_2)
$p(I)$:	the thing we don't care about (probability of image I). We can assume it equals 1.0.

Bayes' Theorem Overview

All that you touch

All that you see

All that you taste

All that you hear

All you feel …

Bayes' Theorem Overview

… is delivered to your brain as a stream of spikes whizzing along neurons.

• These neurons are the only connection between you and the physical world, and the spikes they deliver are the only messages you can ever receive about that world.

• But the messages are corrupted, and the best guess at what they mean is given by ***Bayes' rule***.

Introduction

- Suppose you went to a doctor with some symptoms and your doctor says – 90% of the people who have a particular disease have these symptoms!!
- Should you be worried??
- Should the doctor have told you this?
- What should the doctor have told you?

The material for this presentation has been adapted from:

- "Data Mining Methods and Models", by D T Larose (chapter 5)
- "Foundations of Statistics – Frequentist and Bayesian" by Mary Parker, http://www.austincc.edu/mparker/stat/nov04/
- Bayes' Rule: A tutorial introduction to Bayesian Analysis by James Stone
- Bayesians and Frequentists: Models, Assumptions, and Inference by George Casella

Introduction: Probability

- Extension of logic to deal with uncertainty
- Logic – set of formal rules for determining what propositions are implied to be T/F given the assumption that other propositions are T/F
- Probability – set of formal rules for determining the likelihood of a proposition being T given the likelihood of other propositions.

Introduction: Probability

- Data Science/Machine Learning Context
- DS/ML needs to deal with uncertain and stochastic (non-deterministic) quantities
  - uncertainty and stochasticity can arise due to many reasons

Stochstic -

having a random probability distribution or pattern that may be analysed statistically but may not be predicted precisely

the word **stochastic** is an adjective in English that describes something that was randomly determined. The word first appeared in English to describe a mathematical object called a **stochastic** process, but now in mathematics the terms **stochastic** process and random process are considered interchangeable.

Introduction: Probability

Possible sources of uncertainty:

- Inherent stochasticity in the system being modelled
- Incomplete observability
- Incomplete Modeling

Introduction

- Frequentist – repeatable events (tossing of a coin)
  - Related to rates at which events occur
- Bayesian – non-repeatable events (a patient having 40% chance of having a disease)
  - Qualitative levels of certainty
- Degree of Belief – Doctor diagnosing a patient
  - 1 – absolute certainty about patient having disease
  - 0 - absolute certainty about patient does not have disease

Introduction

- The doctor should have told you about the probability of the disease if symptoms are present!!

- The doctor told you just the opposite – the probability of symptoms, given that the disease is present!!

- P(Symptoms|Disease)
- P(Disease|Symtopms)

Introduction

- The doctor should have told you about the probability of the disease if symptoms are present!! (useful information)
- The doctor told you just the opposite – the probability of symptoms, given that the disease is present!! (not useful information)
- P(Symptoms|Disease) (not useful information)
- P(Disease|Symtopms) (useful information)
- How to convert not useful information (probability) into useful information (probability)?
- How to convert prior probabilities into posterior probabilities?

Introduction

- Bayes' Rule
- Let's try to understand Bayes' rule through an example

Introduction

- Poxy Diseases – Chicken pox (CP) and Small pox (SP)
- CP is common whereas, SP is rare
- Spots on face and body are symptoms for both
- If a person get spots, he/she is interested in knowing the probability of getting the disease
- It is given that:

$P(spots|SP) = 0.9$ (likelihood of SP)

$P(spots|CP) = 0.8$ (likelihood of CP)

- Maximum Likelihood Estimation (MLE) of the disease is SP

Introduction

- CP is common whereas, SP is rare (priors) :

P(CP) = 0.1

P(SP) = 0.001

- Bayes' Rule:

P(SP|spots) = P(spots|SP)*P(SP)/P(spots) – posterior probabilities

P(CP|spots) = P(spots|CP)*P(CP)/P(spots) – posterior probabilities

Probability of spots, P(spots) = 0.081

P(SP|spots) = 0.9*0.001/0.081 = 0.011 – posterior probabilities

P(CP|spots) = 0.8*0.100/0.081 = 0.988 – posterior probabilities

Introduction

Bayes' Inference Engine

P(SP|spots) = 0.9*0.001/0.081 = 0.011 – posterior probabilities

P(CP|spots) = 0.8*0.100/0.081 = 0.988 – posterior probabilities

Inference based on Posterior probabilities:

Both likelihood and posterior probabilities depend on observed data, but posterior probabilities also depend on prior knowledge.

Bayes' rule is used to combine prior experience/knowledge (in the form of prior probabilities) with observed data (in the form of likelihood) for making inference

Bayesian Inference!!

Introduction

The Perfect Inference Engine

- Bayesian Inference is not guaranteed to provide the correct answer!!
- It provides probabilities for each alternatives (SP or CP?)
- Probabilities can be used to find the answer that is most probably true
- Bayesian Inference rule provides an informed guess
- No other procedure can provide a better guess, so Bayesian Inference is considered as the perfect inference engine
- It is fallible, but it is provably less fallible than any other!!

Introduction

General Case

P(hypothesis|data) = P(data|hypothesis)*P(hypothesis)/P(data)

Hypothesis should be interpreted as “Hypothesis is True”

Marginal Probability P(symptoms) or P(spots) or P(data) has no effect on the decision regarding which disease the patient has.

P(data) is a scale factor which guarantees that the posterior probabilities add up to 1!!

MAP – Maximum A Posteriori

MAP can be considered as a regularization of MLE!!

MLE, MAP, Bayes’ are considered to be the holy trinity of parameter estimation!! (see article from Avinash Kak, Purdue University)

Bayesian vs. Frequentist

Two main approaches to probability in Statistics:

- Frequentist or Classical approach
- Bayesian approach

Frequentist Approach

The way we are taught probability!!

Population parameters are fixed constants whose values are unknown

Probability is defined as the relative frequencies of the various categories, where the experiment is repeated indefinitely large number of times

If we toss a fair coin 10 times, it may not be unusual to observer 7 H & 3 T

If we toss a fair coin 10 billion times, we can be fairly certain that the proportion of heads will be 50%

Long run behavior defines probability in the frequentist approach

Frequentist Approach

In lot of situations, for which the classical definition of probability is unclear

- What is the probability that there will be a terrorist attack in Pilani?
- Since such an event has never occurred, it is difficult to conceive the long term behavior this experiment might be

In frequentist approach, the parameters are fixed, and the randomness lies in the data

- Data is viewed as a random sample from a given distribution with unknown but fixed parameters

Bayesian Approach

Bayesian approach to probability, turns these assumptions around

Parameters are considered to be the random variables, and the data are considered to be known

Parameters are coming from a distribution of values

Bayesians look at the observed data to provide information on likely parameter values

Let θ represent the parameters of an unknown distribution

Bayesian analysis requires an elicitation of a prior distribution of θ, called the prior distribution, $p(\theta)$

Bayesian Approach

Prior distribution can come from expert or domain knowledge, if any, regarding the distribution of θ

For example, churn modeling experts may be aware that a customer exceeding a certain threshold number of calls to customer service may indicate a likelihood to churn

This knowledge can be distilled into prior assumption about the distribution of customer service calls, including its mean and std. dev.

Non-informative prior – assign equal probabilities to all the values of the prarameter

p(churners) = p(non churners) = 0.5 😟

Posterior distribution p($\theta|\boldsymbol{X}$)**,** where $\boldsymbol{X}$ represents the observed data

Bayesian Approach

$p(\theta|\boldsymbol{X}) = \frac{\boldsymbol{p(X|\theta)p(\theta)}}{\boldsymbol{p(X)}}$, $\boldsymbol{Bayes\ Theorem}$

$\boldsymbol{p(X|\theta)}$ ***likelihood function***

$\boldsymbol{p(\theta)}$ ***prior distribution***

$\boldsymbol{p(X)}$ ***marginal distribution of data***

Maximum a posteriori (MAP) approach: choose the value of θ ***that maximizes*** $p(\theta|\boldsymbol{X})$***.***

$p(\theta|\boldsymbol{X})$ ***is a distribution rather than a single value, we can conceivably examine any possible statistic of this distribution that we are interested in***

Bayesian Approach

$$p(\theta|\boldsymbol{X}) = \frac{\boldsymbol{p}(\boldsymbol{X}|\theta)\boldsymbol{p}(\theta)}{\boldsymbol{p}(\boldsymbol{X})}$$

For Non-informative priors – MAP estimate and frequentist MLE often coincide, since the data dominate the prior

Criticism of Bayesian Framework:

1. Elicitation of prior very subjective
   a. select non-informative prior if the choice of prior is controversial
   b. Apply lots of data so that the relative importance of prior is diminished
      (that's exactly what we do in ML & DM)
   a. You get two or more models (depending upon no. of priors/beliefs you have) and then do a model comparison/selection

Bayesian Approach

$$p(\theta|\boldsymbol{X}) = \frac{\boldsymbol{p}(\boldsymbol{X}|\theta)\boldsymbol{p}(\theta)}{\boldsymbol{p}(\boldsymbol{X})}$$

Criticism of Bayesian Framework:

2. Scalability Issues (CoD)

   Normalizing factor $\boldsymbol{p}(\boldsymbol{X})$ requires integration or summation over all possible values parameter values which is computationally infeasible when applied directly. MCMC methods like Gibbs sampling and Metropolis algorithm has greatly helped in expanding the range of of problems and dimensionality that Bayesian analysis can handle

Typical Conclusions

	Frequentist	Bayesian
Estimation	I have 95% confidence that the population mean is between 12.7 and 14.5 mcg/liter.	There is a 95% probability that the population mean is in the interval 136.2 g to 139.6 g.
Hypothesis Testing	If H0 is true, we would get a result as extreme as the data we saw only 3.2% of the time. Since that is smaller than 5%, we would reject H0 at the 5% level. These data provide significant evidence for the alternative hypothesis.	The odds in favor of H0 against HA are 1 to 3.

second set of conclusions are easier for most readers to understand

Bayesian

So why don’t we all do Bayesian statistics?

1. The calculations needed for Bayesian statistics can be overwhelming.

2. The structure requires a “prior distribution” on the parameter of interest. If you use a different prior you will obtain different results and this “lack of objectivity” makes some people uncomfortable.

Bayesian

In the last twenty years or so, advances in computing have made it possible to do calculations that earlier would have been completely impractical. This has generated a resurgence of interest in Bayesian methods and it is a rapidly growing field

Bayesian Estimation

- I take a coin out of my pocket and I want to estimate the probability of heads when it is tossed.
- I am only able to toss it 10 times. When I do that, I get seven heads.
- I ask three statisticians to help me decide on an estimator of p, the probability of heads for that coin.

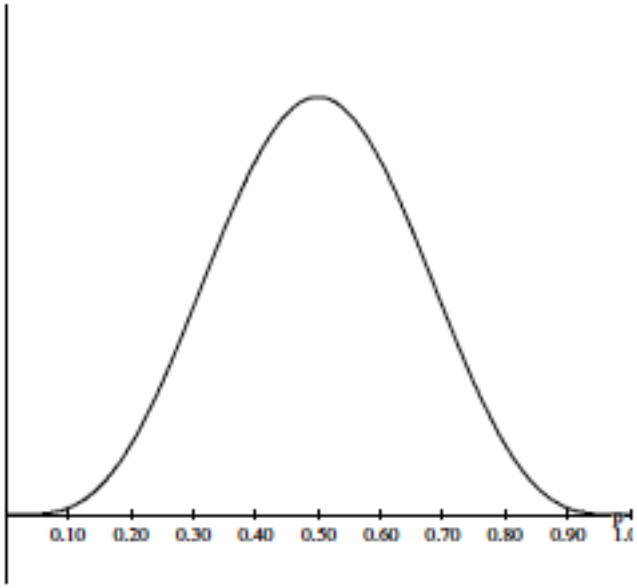
Case 1. Sue, a frequentist statistician, used $\hat{p} = \frac{X}{10}$ =0.7 (MLE??)

Case 2. Jose, who doesn't feel comfortable with this estimator, says that he already has an idea that p is close to 0.5, but he also wants to use the data to help estimate it.

How can he blend his prior ideas and the data to get an estimate? Answer: Bayesian statistics.

Bayesian Estimation

Jose makes a sketch of his prior belief about p. He thinks it is very unlikely that p is 0 or 1, and quite likely that it is somewhere pretty close to 0.5. He graphs his belief.



Beta (5,5)

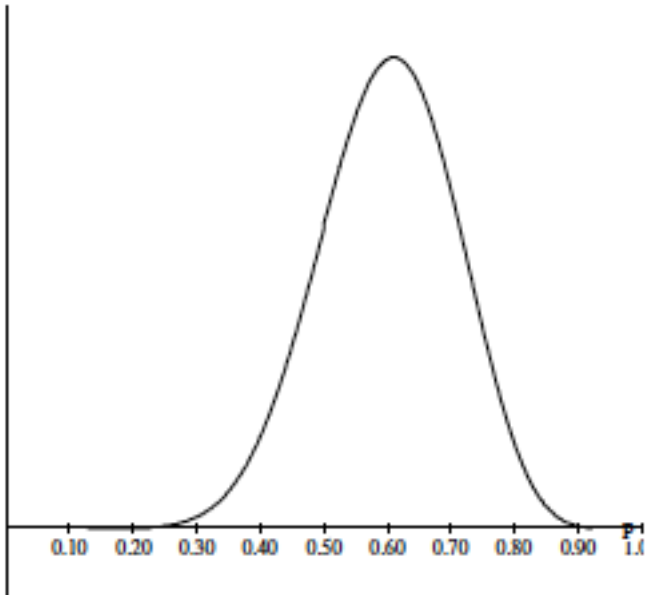
$$\text{Mean} = \frac{a}{a+b} \qquad \text{Variance} = \frac{ab}{(a+b)^2(a+b+1)}$$

Mean = 0.5, var. = 0.022727 std. dev. = 0.15

Bayesian Estimation

Jose makes a sketch of his prior belief about p. He thinks it is very unlikely that p is 0 or 1, and quite likely that it is somewhere pretty close to 0.5. He graphs his belief.

Beta(12,8)

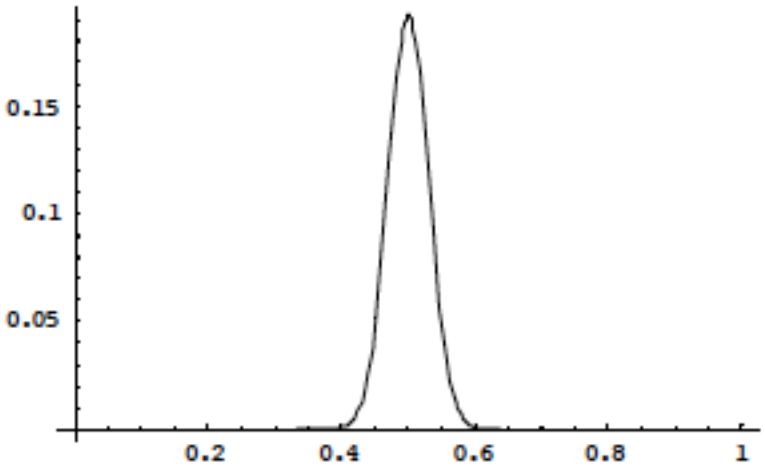


Then, by some magic! (i.e. Bayes Theorem), he combines the data and his prior distribution and gets that the distribution of p, given the data, is a Beta(12,8). This is called the posterior distribution of p.

mean = 0.6 , var. = 0.01143, and std. dev. =0.107

Bayesian Estimation

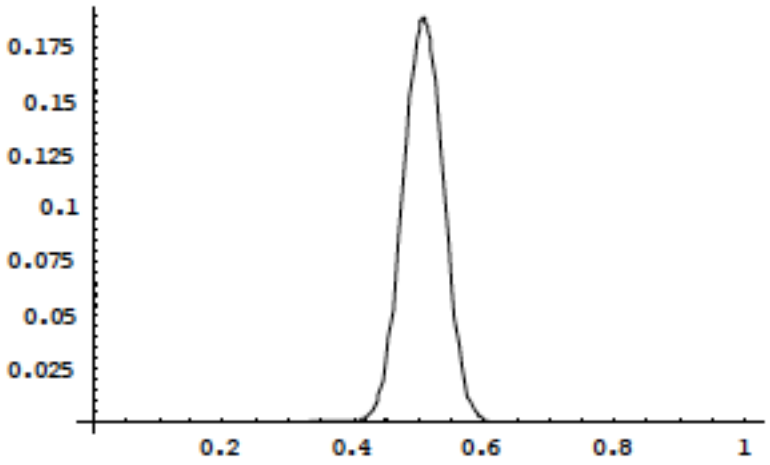
- Case 3: Vicki, who is very sure that coins are unbiased, has a prior distribution like Jose's, but much narrower. There's a much higher probability on values very close to 0.5. She graphs her belief



She notices that this corresponds to a particular probability distribution, which is Beta(138,138), so that is
her prior distribution of p.
mean = 0.5, var.= 0.0009, std. dev. = 0.03.
Notice that her standard deviation is much smaller than Jose's.

Bayesian Estimation

- Vicki's posterior distribution



Now, she also uses Bayes Theorem to combine the data and her prior distribution and finds that her posterior distribution of p is a Beta(145,141). So her posterior mean = 0.507, var.=0.0008709, and std. dev. = 0.0295.

Bayesian Estimation

- Jose and Vicki are both doing Bayesian estimation.
- Both of them decide to use the mean of the posterior distribution of the parameter as their estimator.

Summary:

Sue's estimate of the probability of heads: 0.700

Jose's estimate of the probability of heads: 0.600

Vicki's estimate of the probability of heads: 0.507

Bayesian Estimation

Now, Jennifer offers you a bet. You pick one of these values. She chooses one of the other two.

An impartial person tosses the coin 1000 times, and get a sample proportion of heads.

If that sample proportion is closer to Jennifer's value, you pay her $25. If it is closer to yours, she pays you $25.

Which value would you choose?

Bayesian Estimation

- Jose and Vicki are both doing Bayesian estimation. But they get different answers.
- This is one of the criticism of Bayesian methods (quite subjective)
- Different people can get different estimates based on the same data
- And of course, Jose and Vicki did get different estimators
- But, were they really using the same data?
- Well, not if we consider data in the broad sense.
- If their prior beliefs are considered data, then these are not based on the same data.
- If prior beliefs are not considered data, then we are back to frequentist statistics.
- Would any of you be willing to take up Jennifer's bet and choose Sue's frequentist statistics estimator of 0.7?

Bayesian Estimation

What are some of the difficulties in the Bayesian approach?

1. Quantifying prior beliefs into probability distributions is not simple. First, we haven't all thought much about our prior beliefs about most things, and, even if we have some beliefs, those aren't usually condensed into a probability distribution on a parameter.

2. We might not agree with colleagues on the prior distribution.

3. Even if we can find a formula for the distribution describing our prior beliefs about the parameter, actually doing the probability calculations to find the posterior distribution using Bayes Theorem may be more complex than we can do in closed form. Until people had powerful computing easily available, this was a major obstacle to using Bayesian analysis.

Posterior Distribution

- Everyone is a Bayesian in some situations. I don't think I could find anyone to actually place a bet on the estimate of 0.7 in the example.
- So does that mean that we are all really Bayesians, and there is no point to learning frequentist statistics?
- The basic issue in this example is that we had a quite small set of additional data and most of us feel that we have quite a lot of prior information about whether a coin is likely to be fair.
- It is an example "made to order" to make frequentist analysis look bad.
- In other situations where prior information is not agreed upon, then Bayesian analysis looks less attractive.

Conclusions

The major conceptual difference between Bayesian statistics and frequentist statistics is that, in Bayesian statistics, we consider the parameters to be random, so one consequence of that is that we can write probability intervals for parameters.

Typical Conclusions

	Frequentist	Bayesian
Estimation	I have 95% confidence that the population mean is between 12.7 and 14.5 mcg/liter.	There is a 95% probability that the population mean is in the interval 136.2 g to 139.6 g.
Hypothesis Testing	If H0 is true, we would get a result as extreme as the data we saw only 3.2% of the time. Since that is smaller than 5%, we would reject H0 at the 5% level. These data provide significant evidence for the alternative hypothesis.	The odds in favor of H0 against HA are 1 to 3.

second set of conclusions are easier for most readers to understand

Bayesian

So why don’t we all do Bayesian statistics?

1. The calculations needed for Bayesian statistics can be overwhelming.

2. The structure requires a “prior distribution” on the parameter of interest. If you use a different prior you will obtain different results and this “lack of objectivity” makes some people uncomfortable.

Naïve Bayesian Classification

Example: Who buys computer?

Bayesian Classifiers

- Consider each attribute and class label as random variables

- Given a record with attributes (A_1, A_2,…,A_n)
  - Goal is to predict class C
  - Specifically, we want to find the value of C that maximizes P(C| A_1, A_2,…,A_n)

- Can we estimate P(C| A_1, A_2,…,A_n) directly from data?

Naïve Bayesian Classification

- Also called Simple BC
- Why Naïve/Simple??
- Class Conditional Independence

  Effect of an attribute values on a given class is independent of the values of other attributes

- This assumption simplifies computations

Bayesian Classifiers

- Approach:
  - compute the posterior probability $P(C \mid A_1, A_2, \ldots, A_n)$ for all values of C using the Bayes theorem

$$P(C \mid A_1 A_2 \ldots A_n) = \frac{P(A_1 A_2 \ldots A_n \mid C) P(C)}{P(A_1 A_2 \ldots A_n)}$$

  - Choose value of C that maximizes

    $P(C \mid A_1, A_2, \ldots, A_n)$

  - Equivalent to choosing value of C that maximizes

    $P(A_1, A_2, \ldots, A_n \mid C)\, P(C)$

- How to estimate $P(A_1, A_2, \ldots, A_n \mid C)$?

Naïve Bayes Classifier

- Assume independence among attributes A_i when class is given:
  - $P(A_1, A_2, \ldots, A_n \mid C) = P(A_1 \mid C_j)\, P(A_2 \mid C_j) \ldots P(A_n \mid C_j)$
  - Can estimate $P(A_i \mid C_j)$ for all A_i and C_j.
  - New point is classified to C_j if $P(C_j) \prod P(A_i \mid C_j)$ is maximal.

Naïve Bayesian Classification

Example

Age	Income	Student	Credit_rating	Class:Buys_comp
<=30	HIGH	N	FAIR	N
<=30	HIGH	N	EXCELLENT	N
31…..40	HIGH	N	FAIR	Y
>40	MEDIUM	N	FAIR	Y
>40	LOW	Y	FAIR	Y
>40	LOW	Y	EXCELLENT	N
31…..40	LOW	Y	EXCELLENT	Y
<=30	MEDIUM	N	FAIR	N
<=30	LOW	Y	FAIR	Y
>40	MEDIUM	Y	FAIR	Y
<=30	MEDIUM	Y	EXCELLENT	Y
31….40	MEDIUM	N	EXCELLENT	Y
31….40	HIGH	Y	FAIR	Y
>40	MEDIUM	N	EXCELLENT	N

Naïve Bayesian Classification

Example

A= (<=30,MEDIUM, Y,FAIR, ???)

We need to max.

$P(A|C_j)P(C_j)$ for j =1,2.

$P(C_j)$ is computed from training sample

P(buys_comp=Y) = 9/14 = 0.643

P(buys_comp=N) = 5/14 = 0.357

How to calculate $P(X|C_i)P(C_i)$ for i=1,2?

$P(A|C_j)$ $P(A_1, A_2, A_3, A_4|C) = \Pi P(A_k|C)$

Naïve Bayesian Classification

Example

P(age**<=30** | buys_comp=**Y**)=2/9=0.222

P(age**<=30** | buys_comp=**N**)=3/5=0.600

P(income=**medium** | buys_comp=**Y**)=4/9=0.444

P(income=**medium** | buys_comp=**N**)=2/5=0.400

P(student=**Y** | buys_comp=**Y**)=6/9=0.667

P(student=**Y** | buys_comp=**N**)=1/5=0.200

P(credit_rating=**FAIR** | buys_comp=**Y**)=6/9=0.667

P(credit_rating=**FAIR** | buys_comp=**N**)=2/5=0.400

Naïve Bayesian Classification

Example

P(X | buys_comp=**Y**)=0.222*0.444*0.667*0.667=0.044

P(X | buys_comp=**N**)=0.600*0.400*0.200*0.400=0.019

P(X | buys_comp=Y)P(buys_comp=Y) = 0.044*0.643=0.028

P(X | buys_comp=N)P(buys_comp=N) = 0.019*0.357=0.007

CONCLUSION: ***A buys computer***

Beta Distribution

Beta distribution is a family of continuous probability distributions defined on the interval [0, 1] parametrized by two positive shape parameters, denoted by α and β, that appear as exponents of the random variable and control the shape of the distribution.

Beta distribution can be used in Bayesian analysis to describe initial knowledge concerning probability of success. The beta distribution is a suitable model for the random behavior of percentages and proportions.

Source: Wikipedia