e-PGPathshala

Subject : Computer Science

Paper: Machine Learning

Module: Probability

Module No: CS/ML/6

Quadrant I – e-text

Welcome to the e-PG Pathshala Lecture Series on Machine Learning.

Learning Objectives:

The learning objectives of this module are as follows:

- To understand the definitions of Basics of Probability for Machine Learning
- To understand the mathematics for Machine Learning
- To know the concept of Bayes Theorem and supervised Machine learning
- To design the Machine Learning problem using Bayes Theorem

6.1 "7 Giants" of Data

Data and analysis of data play a very important role in machine learning. These are the seven so called giants of data

1. Basic Statistics – This includes basic statistical measures such as counts, contingency table, mean, median, variance, range queries (SQL queries) etc..
2. Generalized N-body Problems which include kernel summations, clustering, spatial correlations, etc.
3. Graph Theoretic Problems which include betweeness, centrality, commute distance, graphical model inference
4. Optimization- general methods of Optimization
5. Linear Algebraic Problems which include Linear algebra, Principal component analysis, Gaussian Process Regression, Mainfold Learning
6. Integration using Bayesian Inference
7. Alignment problem which include BLAST in genomics, string matching, phylogenies, SLAM, cross-match and some other operations specific to bioinformatics

However in this module we will be discussing only the basic aspects of probability and how these concepts are connected to machine learning.

6.2 Two views of Probability

When we talk about probability there are basically two aspects that we consider. One is the classical interpretation where we describe the frequency of outcomes in random experiments. The other is Bayesian viewpoint or subjective interpretation of probability where we describe the degree of belief about a particular event.

Before we go further let us first define probability. **Probability** is defined as the chance that an uncertain event will occur (always between 0 and 1). In this context **Sample Space** is defined as the collection of all possible events. Example of a sample space for all the 6 faces of a die is shown in Figure 6.1

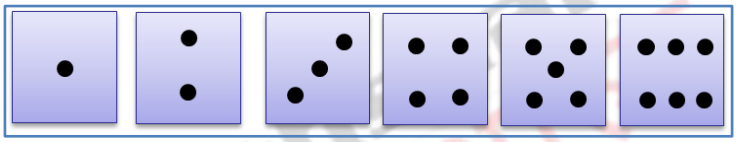

Figure 6.1 Six Faces of a Die

6.3 Types of Probability

There are three approaches to assessing the probability of an uncertain event. They are defined below.

6.3.1 Apriori classical probability: the probability of an event is based on **prior knowledge of the process involved**.

a priori classical probability

$$\text{Probability of Occurrence} = \frac{X}{T} = \frac{\text{number of ways the event can occur}}{\text{total number of possible outcomes}}$$

Priori classical probability – Example 1
Find the probability of selecting a face card (Jack, Queen, or King) from a standard deck of 52 cards**.**

$$\text{Probability of Face Card} = \frac{X}{T} = \frac{\text{number of face cards}}{\text{total number of cards}}$$

$$\frac{X}{T} = \frac{12 \text{ face cards}}{52 \text{ total cards}} = \frac{3}{13}$$

6.3.2 Empirical classical probabilityisthe probability of an event is based on **observed data**.
empirical classical probability

$$\text{Probability of Occurrence} = \frac{\text{number of favorable outcomes observed}}{\text{total number of outcomes observed}}$$

Equationsassume all outcomes are equally likely.

Empirical classical probability – Example 2
Find the probability of selecting a male taking statistics from the population described in the following table (Figure 6.2):

	Taking Stats	Not Taking Stats	Total
Male	84	145	229
Female	76	134	210
Total	160	279	439

Figure 6.2 Example2 for Classical Probability

$$\text{Probability of Male Taking Stats} = \frac{\text{number of males taking stats}}{\text{total number of people}} = \frac{84}{439} = 0.191$$

6.3.3 Subjective probability: the probability of an event is **determined by an individual**

6.4 Events – Sample Space

Associated with probability is the concept of an event. Event is defined as each possible type of occurrence or outcome. There are basically different concepts associated with an event.

- **Simple event** is an outcome from a sample space with one characteristic. ex. A red card from a deck of cards
- **Complement of an event A** (denoted $A^{/}$) is defined as all outcomes that are not part of event A
  ex. All cards that are not diamonds

- **Joint event**Involves two or more characteristics simultaneously.
  ex. An ace that is also red from a deck of cards
- **Mutually exclusive event**are eventsthat cannot occur together (simultaneously).
  ex. B = having a boy; G = having a girl
- **Collectively exhaustive events**–here **o**ne of the events must occur from the set of events covers the entire sample space

Example 3

A = aces; B = black cards; C = diamonds; D = hearts
Events A, B, C and D are **collectively exhaustive** (but not mutually exclusive)
Events B, C and D are **collectively exhaustive** and also mutually exclusive

6.5 Visualizing Events in Sample Space

There are basically two ways in which events can be visualized in the sample space namely contingency table and tree diagrams.

Figure 6.3 and Figure 6.4 shows an example of black and red aces as events in a sample space which is a pack of 52 cards represented as contingency table and tree diagrams repectively.

	Ace	Not Ace	Total
Black	2	24	26
Red	2	24	26
Total	4	48	52

Figure 6.3 Contingency Table

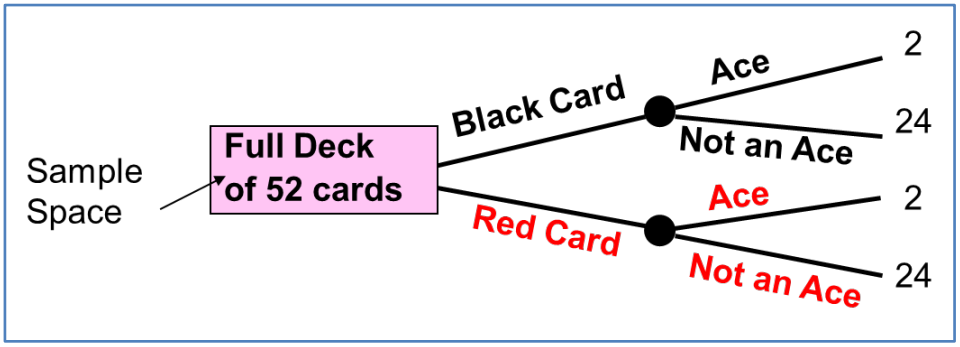

Figure 6.4Tree Diagram

6.6 Simple vs. Joint Probability

Simple (Marginal) Probability refers to the probability of a simple event.ex. P(King). The probability of an event A given as P(A) is given below:

$$P(A)=P(A \text{ and } B_1)+P(A \text{ and } B_2)+\cdots+P(A \text{ and } B_k)$$

Joint Probability refers to the probability of an occurrence of two or more events.ex. P(King and Spade)

$$P(A \text{ and } B) = \frac{number\ of\ outcomes\ satisfying\ A\ and\ B}{total\ number\ of\ elementary\ outcomes}$$

The contingency table for joint probability is gen in Figure 6.5. In this figure we have taken the example of two sets of events A_1 ,$A_2$and B_1, B_2. When events A and B occur together we have joint probability but the total is the marginal probability of events A and B.

Event	Event		Total
	B_1	B_2	
A_1	$P(A_1 \text{ and } B_1)$	$P(A_1 \text{ and } B_2)$	$P(A_1)$
A_2	$P(A_2 \text{ and } B_1)$	$P(A_2 \text{ and } B_2)$	$P(A_2)$
Total	$P(B_1)$	$P(B_2)$	1

Joint Probabilities

Marginal (Simple) Probabilities

Figure 6.5 Contingency Table for Joint Probability

Marginalization

Let us now explain the concept of marginalization. Consider the probability of X irrespective of Y.

$$p(X = x_j) = \frac{c_j}{N}$$

The number of instances in column j is the sum of instances in each cell of that column and is as given below:

$$c_j = \sum_{i=1}^{L} n_{ij}$$

Therefore, we can **marginalize** or "sum over" Y:

$$p(X = x_j) = \sum_{j=1}^{L} p(X = x_j, Y = y_i)$$

Sum and Product Rules

In general, we'll refer to a distribution over a random variable as p(X) and a distribution evaluated at a particular value as p(x). Two important rukes of probability include the sum and product rules.

Sum Rule $p(X) = \sum_{y} p(x, y)$

Product Rule $p(X,Y) = p(Y \mid X)p(X)$

6.7 Conditional Probability and Bayes Theorem

Another very important concept that we need to understand is the concept of conditional probability. Consider only instances where X = x_j. The fraction of these instances where Y = y_i is the conditional probability. In other words the The probability of y given x is as given below

$$p(X,Y) = p(Y \mid X)p(X)$$

A conditional probability is the probability of one event, given that another event has occurred is given below:

$$\mathrm{P(A \mid B)} = \frac{\mathrm{P(A \text{ and } B)}}{\mathrm{P(B)}}$$

The conditional probability of A given that B has occurred

The conditional probability can also be with respect to B given A. and is as given below:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

The conditional probability of B given that A has occurred

In the above definitions

P(A and B) = joint probability of A and B

P(A) = marginal probability of A

P(B) = marginal probability of B

Based on the concept of conditional probability we go on to discuss a very important probability theorem which is used extensively in machine learning.

6.7.1 Bayes' Theorem

Bayes' Theorem is used to revise previously calculated probabilities based on new information. This theorem was developed by Thomas Bayes in the 18th Century. It is an extension of conditional probability and is given by

$$P(B_i \mid A) = \frac{P(A \mid B_i)P(B_i)}{P(A \mid B_1)P(B_1) + P(A \mid B_2)P(B_2) + \ldots + P(A \mid B_k)P(B_k)}$$

Where: B_i = i^{th} event of k mutually exclusive and collectively exhaustive events

A = new event that might impact $P(B_i)$.

6.7.2 Bayes' Theorem – Example 4

A drilling company has estimated a 40% chance of striking oil for their new well. A detailed test has been scheduled for more information. Historically, 60% of successful wells have had detailed tests, and 20% of unsuccessful wells have had detailed tests. Given that this well has been scheduled for a detailed test, what is the probability that the well will be successful?

- Let S = successful well U = unsuccessful well
- P(S) = .4 , P(U) = .6 (prior probabilities)
- Define the detailed test event as D
- Conditional probabilities: P(D|S) =0.6 P(D|U) =0.2

Goal: To find P(S|D) using bayes Theorem

$$P(S \mid D) = \frac{P(D \mid S)P(S)}{P(D \mid S)P(S) + P(D \mid U)P(U)}$$

$$= \frac{(.6)(.4)}{(.6)(.4) + (.2)(.6)}$$

$$= \frac{.24}{.24 + .12} = .667$$

Given the detailed test, the revised probability of a successful well has risen to .667 from the original estimate of 0.4. The given probabilities can be represented using a contingency table (Figure 6.6)

Event	Prior Prob.	Conditional Prob.	Joint Prob.	Revised Prob.
S (successful)	.4	.6	.4*.6 = .24	.24/.36 = .667
U (unsuccessful)	.6	.2	.6*.2 = .12	.12/.36 = .333

Figure 6.6 Contingency Table for Example 4.

6.7.3 Interpretation of Bayes Rule

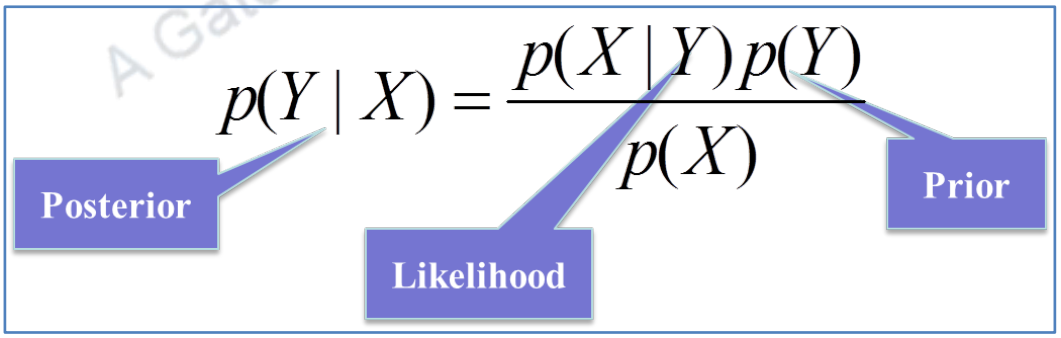

Figure 6.7 The Bayes Rule

The three terms (Figure 6.7) of the Bayes Theorem are as follows:

- **Prior**: Information we have before observation.
- **Posterior**: The distribution of Y after observing X
- **Likelihood:** The likelihood of observing X given Y

6.9 Naïve Bayesian Classification

Naïve Bayes Classifier is based on the Bayes Theorem. Given an instance of an entity we want to find its Class. In other words we want to find the posterior probability of the Class C given the sample data d_i. In order to find this class we need prior probability that is the frequency of items of Class C in the complete dataset and Likelihood which the probability that given the class C the data item d_i is likely to occur. If i-th attribute is **categorical** $P(d_i|C)$ is estimated as the relative freq of samples having value d_i as i-th attribute in class C. If i-th attribute is **continuous** $P(d_i|C)$ is estimated through a Gaussian density function. It is computationally easy in both cases to find this likelihood.

6.9.1 Theoretical basis:

Now given the data items X_1, X_2, ...X_n, we want to find a class C* that maximizes the posterior probability. We replace the posterior probability by prior probability and likelihood using Bayes Theorem,

$$c^* = \underset{c}{\operatorname{argmax}} p(c \mid x_1, x_2, x_3, \ldots\ldots x_n)$$

$$c^* = \underset{c}{\operatorname{argmax}} \frac{p(x_1, x_2, x_3, \ldots\ldots x_n \mid c)p(c)}{p(c \mid x_1, x_2, x_3, \ldots\ldots x_n)}$$

$$p(c \mid x_1, x_2, x_3, \ldots\ldots x_n) = p(x_1 \mid c)p(x_2 \mid c), p(x_3 \mid c) \ldots\ldots p(x_n \mid c)$$

According to Naïve Bayes we make the simplified assumption that all the attributes are conditionally independent as given below::

$$P(C_j \mid D) \propto P(C_j) \prod_{i=1}^{n} P(d_i \mid C_j)$$

6.9.2 Maximum A Posteriori (MAP) Hypothesis and Maximum Likelihood

Our goal is to find the most probable hypothesis *h* from a set of candidate hypotheses *H* given the observed data *D*. In other words we want to find hypothesis h which gives the maximum posterior probability. In Figure 6.8 the

MAP Hypothesis, $h_{MAP} = \operatorname{argmax}_{h \in H} P(h|D)$

$= \operatorname{argmax}_{h \in H} P(D|h)P(h)/P(D)$

$= \operatorname{argmax}_{h \in H} P(D|h)P(h)$

Figure 6.8 MAP

Posterior probability of the hypothesis h given the Data D is replaced by likelihood (P(D/h)-probability of the hypothesis given the Data) and the prior of the hypothesis (P(h)) by applying the Bayes Theorem. If every hypothesis in *H* is equally probable a priori, we only need to consider the likelihood of the data *D* given *h*, *P(D|h).* In this case h_{MAP} becomes the Maximum Likelihood,

6.9.3 Bayes Classifiers

For describingthe Bayes classifier we assume that training set consists of instances of different classes described ***cj*** as conjunctions of attributes values. Now the task is to classify a new instance ***d***based on a tuple of attribute values into one of the classes **cj∈C.** Thekey idea **is to** assign the most probable class c_{MAP} using Bayes Theorem. The hypothesis we need here is the class C given the set of attribute values.

$$c_{MAP} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid x_1, x_2, \ldots, x_n)$$
$$= \underset{c_j \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c_j) P(c_j)$$

6.9.3.1 Example 5 – Text Classification

We consider a typical natural language processing application namely text classification to illustrate the method. In order to be able to perform all calculations, we will use an example with extremely small documents.

Type	Text	Class
Training	Good happy Good	positive
Training	Good good Service	Positive
Training	Good friendly	Positive
Training	Lousy good chat	Negative
Test	Good goodgood cheat lousy	??

Figure 6.9(a) Example 5 - Set of Documents

As given in Fig 6.9 (a), we are given 4training documents, and the class of each document is also given. Given a new document (called test) we want to find it's class.

We need nonzero probabilities for all words, even words that don't exist. For this we just count every word one time more than it actually occurs (Figure 6.9(b)).Since we are only concerned with relative probabilities, this inaccuracy should not be a problem.

The first step in the calculation (Figure 6.9(c)) is to find the prior probability of positive and negative in the whole corpus (here it is a toy size is4).One additional document is used as test).

$$P(word|C) = \frac{count(word|C) + 1}{count(C) + V}$$

(V is the total vocabulary, so that our probabilities sum to 1.)

Figure 6.9 (b) Probability Calculation

Now the words in the test document are *good, cheat and lousy*. We need to find the likelihood for each of these words when the class is positive and when it is negative.

Calculations for the Example:

First we need to find prior probability. Here we have 3 positive and 1 negative document.
Therefore Prior Probability of positive and negative are P(pos) =3/4 and P(neg) =1/4.
Now we have to find Likelihood:
Likelihood of good in positive documents & negative documents
P(good/pos) = 5 (number of goods in positive documents) +1 divided by total number of words in positive documents (8) + vocabulary (6)
= 6/14 =3/7
Similarly P(good/neg) = (1+1)/(3+6) =2/9
P(cheat/pos) = (0+1)/(8+6) = 1/14 and P(cheat/neg) = (1+1)/(3+6) = 2/9
P(lousy/pos) = (0+1)/(8+6) = 1/14 and p(lousy/neg) = (1+1)/(3+6) = 2/9

Now we can calculate the probability of the test document being positive
P(Pos/D5) = P(POS) *(P(good/POS) * P(cheat/POS) * P(lousy/POS)
= ¾ * 3/7 * 1/14 *1/14 = 0.0003

Similarly P(neg/D5) = ¼ * 2/9 *2/9*2/9 = 0.0001

Now we know that probability of positive is higher and so test document is classified aspositive.

Bayes Theorem in general and Naïve Bayes in particular have been used in many other natural language processing applications such as Part of Speech

Tagging, Statistical Spell Checking, Automatic Speech Recognition, Probabilistic Parsing and Statistical Machine Translation.

6.9.3.2 Naïve Bayes Classification - Example 6

Now let us consider another example to illustrate Naïve Bayes classification. We use the equation given below to find the colour.

$$c^* = \underset{c}{\operatorname{argmax}}\, p(x_1|c)p(x_2|c)\cdots p(x_n|c)p(c)$$

HOT	LIGHT	SOFT	RED
COLD	HEAVY	SOFT	RED
HOT	HEAVY	FIRM	RED
HOT	LIGHT	FIRM	RED
COLD	LIGHT	SOFT	BLUE
COLD	HEAVY	FIRM	BLUE
HOT	HEAVY	FIRM	BLUE
HOT	LIGHT	FIRM	BLUE
HOT	HEAVY	FIRM	?????

Figure 10.10 (a) Data of Example 6

First we find the prior of red and blue:

$$P(c = red) = 0.5$$
$$P(c = blue) = 0.5$$

Next we find the likelihood probabilities of red and blue given each of the attributes independently.

$$P(hot \mid red) = 0.75 P(heavy \mid red) = 0.5 P(firm \mid red) = 0.5$$
$$P(hot \mid blue) = 0.5 P(heavy \mid blue) = 0.5 P(firm \mid blue) = 0.5$$

Finally we find the posterior probabilities for each of the category and we find the category to be red since that category has maximum posterior probability

P(hot | c=red) P(heavy | c=red) P(firm | c=red) P(c=red)
= 0.75*0.5*0.5*0.5 = 0.09375
P(hot | c=blue) P(heavy | c=blue) P(firm | c=blue) P(c=blue)
= 0.5*0.5*0.5*0.5 = 0.0625

The independence hypothesis that assuming that the attributes are independent of each other makes computation possible, yields optimal classifiers when satisfiedbut is seldom satisfied in practice, as attributes (variables) are often correlated.Bayesian networksthat combine Bayesian reasoning with causal relationships between attributes attempts to overcome this limitation. We will study Naïve Bayes Classifier in detail in another module.

Summary

- Described the basics of probability
- Discussed the concepts of Bayes Theorem
- Explained the application of Bayes theorem to Machine Learning

epathshala पाठशाला
A Gateway to All Post Graduate Courses