e-PGPathshala

Subject : Computer Science

Paper: Machine Learning

Module: Bayes Learning

Module No: CS/ML/29

Quadrant I – e-text

Welcome to the e-PG Pathshala Lecture Series on Machine Learning. In this module we will discuss an important concept called Bayes Learning. This concept forms the core of a number of modern machine learning algorithms and applications.

Learning Objectives:

The learning objectives of this module are as follows:

- To understand Probability, Bayes Theorem and Bayes Learning
- To understand Bayes Learning Framework
- To understand Probabilistic Classification

29.1 Introduction

Before we discuss Bayes learning let us consider the important issue of fitting a model to data. This is important in machine learning before based on this model built from data and associated target values, we predict target values for unknown hitherto unseen data. Some of the important criteria when we fit a model are the choice of weights and thresholds. In some cases we need to incorporate prior knowledge and in other cases we need to merge multiple sources of information. Another important factor is the modelling of uncertainty. Bayesian reasoning provides solutions to the above issues. Bayesian reasoning is associated with probability, statistics and data fitting. Figure 29.1 shows the idea of frequentist or orthodox statistics where probability is defined as the frequency of occurrences in possibly infinite number of trials.

Principle #1: The first principle of probability as stated by Pierre-Simon Laplace (Figure 29.2) in 1814 is given below:

“Probability theory is nothing more than common sense reduced to calculation.”

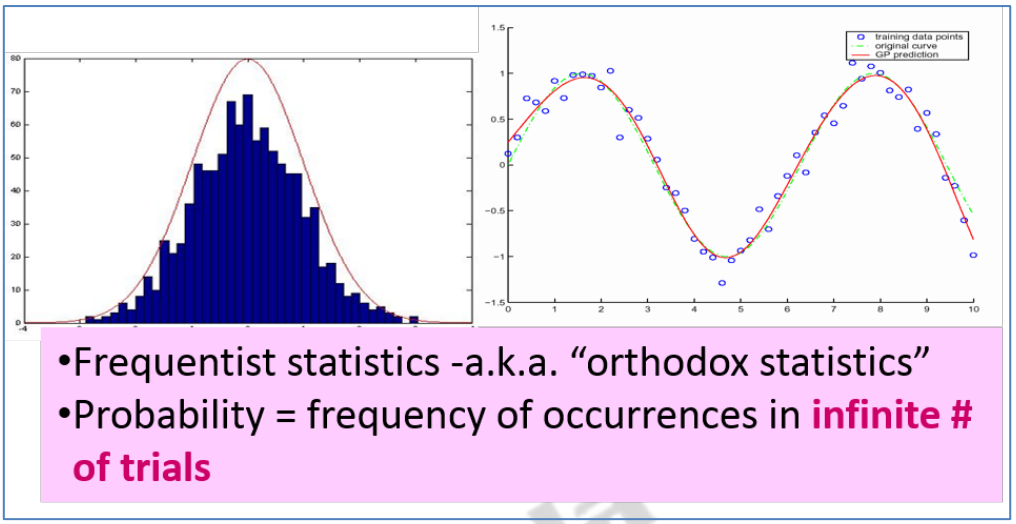


Figure 29.1 Probability as Frequency of Occurences

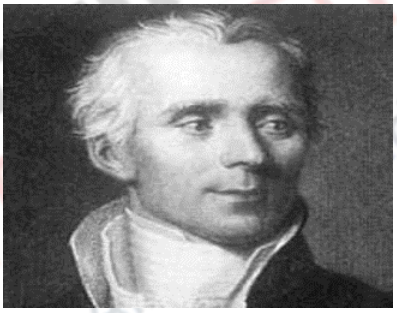


Figure 29.2 Pierre-Simon Laplace in 1814

29.2 Probability Basics

Let us recall the basic laws of probability given in Figure 29.3. We talk about prior or independent probability of a variable X as P(X). Conditional probability $P(X_1|X_2)$ is defined as the probability of X_1 given X_2. Joint probability is defined as the probability of both X_1 and X_2 occurring together. The relation between joint probability and conditional probability is also given. If X_1 and X_2 are independent

- Prior, conditional and joint probability for random variables
  - Prior probability: $P(X)$
  - Conditional probability: $P(X_1 \mid X_2), P(X_2 \mid X_1)$
  - Joint probability: $\mathbf{X} = (X_1, X_2), P(\mathbf{X}) = P(X_1, X_2)$
  - Relationship: $P(X_1, X_2) = P(X_2 \mid X_1)P(X_1) = P(X_1 \mid X_2)P(X_2)$
  - Independence: $P(X_2 \mid X_1) = P(X_2), P(X_1 \mid X_2) = P(X_1), P(X_1, X_2) = P(X_1)P(X_2)$

Figure 28.3 Basics of Probability Bayesian Rule

then the conditionally probability of X_1 given X_2 reduces to $P(X_1)$ and similarly the conditional probability of X_2 given X_1 reduces to $P(X_2)$. The joint probability when X_1 and X_2 are independent is given by the product of $P(X_1)$ and $P(X_2)$.

29.3 Bayesian Rule

The Baye's Rule or Bayes theorem is a very important concept that is used in many different aspects of machine learning. Figure 29.4 shows a general definition of the rule where the conditional probability of C given X is given by the conditional probability of X given C multiplied by the independent probability of C divided by the independent probability of X. Here the conditional probability of C given X (that is the probability to be determined) is called the posterior probability, the conditional probability of X given C (usually determined from the labelled data) is called likelihood, independent probability of C is called prior probability and the independent probability of X is called evidence. This very important theorem is further explained in Figure 29.5. Here we explain that the posterior probability of the target vector W given the training data D can be determined knowing the prior probability of target vector W, likelihood or probability of observed data D given target vector W and the evidence or the probability of the data D.

.

Figure 29.4 Bayesian Rule

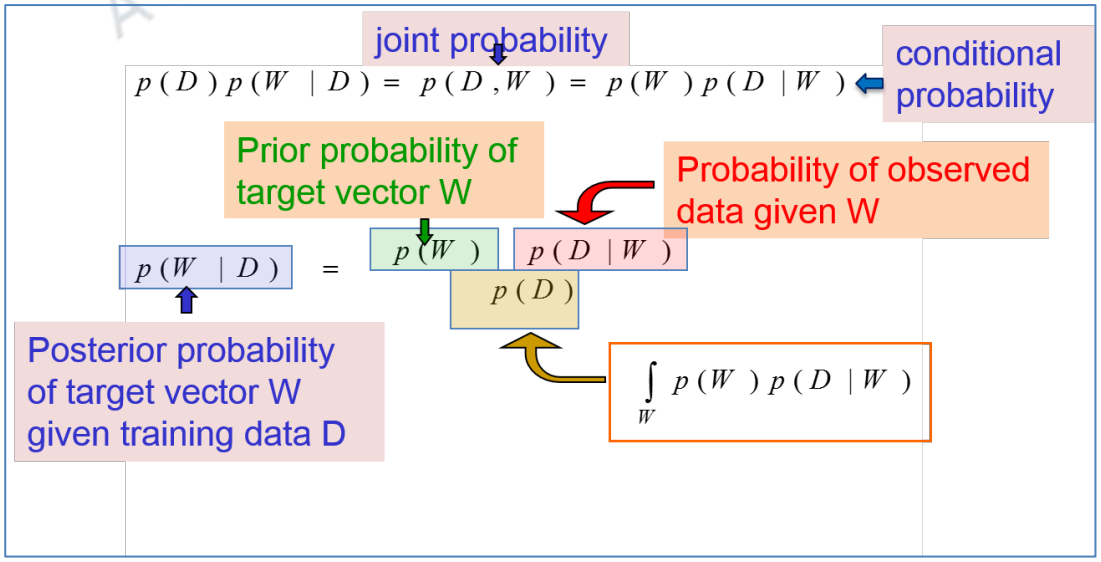


Figure 29.5 Explanation of Bayes Theorem

29.4 Discrete and Continuous Variables – Modelling the World

Let us first discuss probabilities over discrete variables. A probability function maps the possible values of x against their respective probabilities of occurrence, P(x). P(x) is a number from 0 to 1.0. Let us assume the simple case where the discrete variable C can take on two values Heads or Tails. We know that in such cases the probability of C=Heads and C=Tails i.e P(C=Heads) = P(C=Tails) is 0.5 and the P(C=Heads) + P(C=Tails) = 1.

C - { Heads, Tails }
P(C=Heads) = .5

P(C=Heads) + P(C=Tails) = 1

Now let consider the case of continuous variables. The probability function that accompanies a continuous random variable is a continuous mathematical function that integrates to 1. Let x be a real number. Then we need to describe beliefs over x where x may be joint angles, price of stock etc..

Let x - R^N
How do we describe beliefs over x?
e.g., x is a face, joint angles, ...

In the case of continuous variables we describe the data in terms of a probability distribution function (PDF) or marginal probability. One common distribution that is used to describe the data is the Gaussian or normal distribution. This distribution can be defined by the parameters mean μ and standard deviation σ (Figure 29.6). This is a bell shaped curve with different centers and spreads depending on μ and σ.

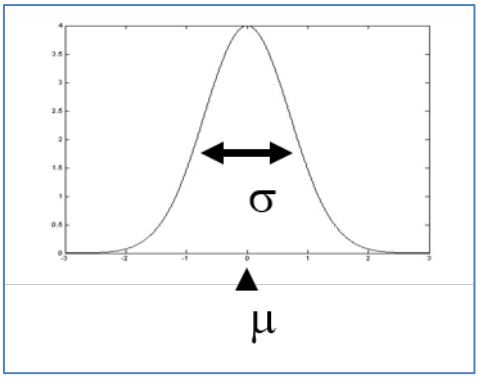


Figure 29.6 Gaussian Distribution

Now the question is why do we use Gaussians? This is due to the fact that it has convenient analytic properties, it is governed by the central limit theorem, works reasonably well for most real data. Though it does not suit all types of data, it acts as a good building block. The values of of the data point x is given by the function (Figure 29.7).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$$

Figure 29.7 Gaussian Distribution Function

29.5 Inference

How do we reason about the world from observations? In this context there are three important sets of variables namely observations, unknowns and auxiliary ("nuisance") variables. Given the observations, what are the probabilities of the unknowns? The first step is parameter estimation. Here we are given lots of data, and using this data we need to determine unknown parameters. The various estimators used to determine these parameters include Maximum A Posteriori (MAP), Maximum likelihood and Unbiased estimators. We will discuss these estimators later.

A word about finding a model that fits the data. A model that fits the data well but does not generalize is said to be over fitting. This does not augur well for machine learning since these type of models do not predict well on unseen data. This occurs when an estimate is obtained from a "spread-out" posterior.

29.6 Probability "Principles"

29.6.1 Basic Probability:

The following are some of the facts about probability:

1. Probability theory is common sense reduced to calculation.
2. Given a model, we can derive any probability
3. Describe a model of the world, and then compute the probabilities of the unknowns with Bayes' Rule
4. Parameter estimation leads to over-fitting when the posterior isn't "peaked." However, it is easier than Bayesian prediction.
5. Least-squares estimation is a special case of MAP, and can suffer from over- and under-fitting

6. You can learn (or marginalize out) all parameters.

29.6.2 Benefits of the Bayesian approach

The following are some of the benefits of the general Bayesian approach: This approach allows principled modeling of noise and uncertainty. It defines a unified model for learning and synthesis where essentially all parameters can be learnt and it is possible to get reasonably good results from simple models. This is an area where lots of good research is being carried out and several good algorithms are available.

Some of the applications where Bayesian methods have been employed include data mining, robotics, signal processing, bioinformatics, text analysis, etc.

29.7 Bayesian Framework

The Bayesian framework allows us to combine observed data and prior knowledge. It provides practical learning algorithms. Moreover it is a generative (model based) approach, which offers a useful conceptual framework. This essentially means that any kind of object (e.g. time series, trees, etc.) can be classified, based on a probabilistic model specification.

The Bayesian framework assumes that we always have a prior distribution for everything. However the usage of prior knowledge alone may not be effective. When we observe some data, we combine our prior distribution with a likelihood term to get a posterior distribution. The likelihood term takes into account how probable the observed data is given the parameters of the model. The framework favors parameter settings that make the data likely. It in a way changes the estimate based solely on the prior and fights the prior. With sufficient data, the likelihood terms always win.

29.8 What is Bayesian Learning?

Bayesian learning uses **probability to model** data and quantify uncertainty of predictions. As already discussed it facilitates incorporation of prior knowledge and allows for decision-theoretic reasoning and a probabilistic approach to inference. Some of the assumptions are that the quantities of interest are governed by probability distribution. The optimal decisions are based on reasoning about probabilities and observations. It provides quantitative approach to weighing how evidence supports alternative hypotheses

29.8.1 Features of Bayesian Learning

Each observed training data can incrementally decrease or increase the estimated probability of a hypothesis rather than completely eliminating a hypothesis if it is found to be inconsistent with a single example. **Prior knowledge** can be combined with observed data to determine the final

probability of a hypothesis. Moreover new instances can be classified by combining predictions of multiple hypotheses. Even in computationally intractable cases, Bayesian optimal classifier provides a standard of optimal decision against which other practical methods can be compared.

29.8.2 Context of Bayesian Learning

Bayesian Decision Theory came long before Version Spaces, Decision Tree Learning and Neural Networks. It was studied in the field of Statistical Theory and more specifically, in the field of ***Pattern Recognition***. Bayesian Decision Theory is at the basis of important learning schemes such as the ***Naïve Bayes Classifier***, ***Bayesian Belief Networks*** and the ***Expectation Maximization (EM) Algorithm.*** Bayesian Decision Theory is also useful as it provides a framework within which many non-Bayesian classifiers can be studied.

29.8.3 Advantages of Bayesian Learning

Bayesian approaches, including the Naive Bayes classifier, are among the most common and practical ones in machine learning. Bayesian decision theory allows us to revise probabilities based on new evidence. Bayesian methods provide a useful perspective for understanding many learning algorithms that do not manipulate probabilities.

29.9 Relook at Bayes Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Now let us relook at the Bayes theorem from the viewpoint of Bayesian Learning. The goal is to determine the most probable hypothesis, given the data *D* plus any initial knowledge about the prior probabilities of the various hypotheses in *H*. With this premise, ***Prior probability of h, P(h)*** reflects any background knowledge we have about the chance that *h* is a correct hypothesis. This is before having observed the data. ***Prior probability of D, P(D)*** reflects the probability that training data *D* will be observed given no knowledge about which hypothesis *h* holds. Then we have ***Conditional Probability of observation D, P(D|h)*** which denotes the probability of observing data *D* given some world in which hypothesis *h* holds. Finally we have ***Posterior probability of h, P(h|D)*** which represents the probability that *h* holds given the observed training data *D*. It reflects our confidence that *h* holds after we have seen the training data *D* and it is the quantity that Machine Learning researchers are interested in.

Bayes Theorem allows us to compute ***P(h|D).*** We will now discuss the different ways to carry out this computation. We will discuss two methods namely

M**aximum A Posteriori (MAP) & Maximum Likelihood (ML).** The goal is to find the most probable hypothesis *h* from a set of candidate hypotheses *H* given the observed data *D*.

In the M**aximum A Posteriori (MAP)** method we need to find the hypothesis h among the set of hypothesis H that maximizes the posterior probability ***P(h|D).*** Now by applying Bayes theorem, we need to find the hypothesis h that maximizes the likelihood of the data D for a given h and the prior probability of h. The prior probability of the data D does not affect the maximization for finding h so that is not considered. Hence we have the MAP hypothesis as given below:

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

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

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

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).*** This gives rise to h_{ML} or the ***Maximum Likelihood***, as follows:

$$h_{ML} = \textbf{argmax}_{h \in H} P(D|h)$$

26.9.1 Bayes Example

Now let us look at an example of Bayes Theorem. We have data on past student performance. For each student we know the following:

- If student's GPA > 3.0 (G)
- If student had a strong math background (M)
- If student is a hard worker (H)
- If student passed or failed course

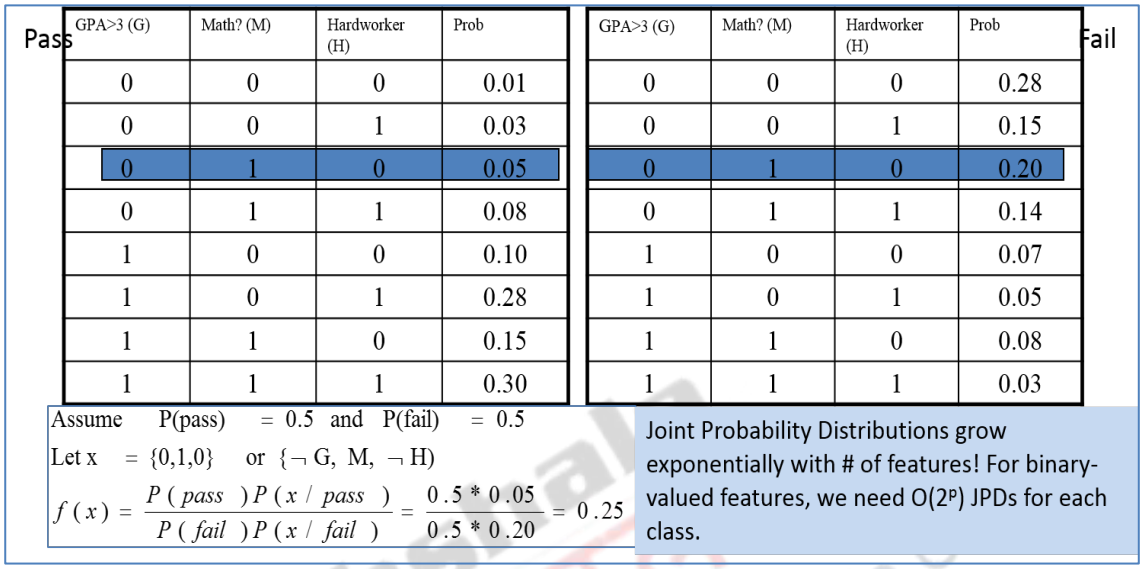
A new student comes along with values $G = g, M = m,$ and $H = h$ and wants to know if they will likely pass or fail the course.

$$f(g,m,h) = \frac{P(g,m,h,pass)}{P(g,m,h,fail)}$$

If $f(g,m,h) \geq 1$, then classifier predicts $pass$; otherwise $fail$.

We are given the data as in the Table of Figure 29.8 where p is the number of binary features. Let us assume that the probability to pass or fail are equal i.e

P(pass)=P(fail)=0.5. Now if we are given that the student has GPA,3, has Maths, and is not a Hard worker that is the value of the features are x={0,1,0}. From the table we know that for this value of x the probability of passing is 0.05 and the probability of failing is 0.20.Therefore we can find f(x)=0.25 (Figure 29.8) which means that the classification gives the value as fail.



Pass

GPA>3 (G)	Math? (M)	Hardworker (H)	Prob
0	0	0	0.01
0	0	1	0.03
0	1	0	0.05
0	1	1	0.08
1	0	0	0.10
1	0	1	0.28
1	1	0	0.15
1	1	1	0.30

Fail

GPA>3 (G)	Math? (M)	Hardworker (H)	Prob
0	0	0	0.28
0	0	1	0.15
0	1	0	0.20
0	1	1	0.14
1	0	0	0.07
1	0	1	0.05
1	1	0	0.08
1	1	1	0.03

Assume P(pass) = 0.5 and P(fail) = 0.5

Let x = {0,1,0} or {¬ G, M, ¬ H)

$$f(x) = \frac{P(pass)\,P(x/pass)}{P(fail)\,P(x/fail)} = \frac{0.5 * 0.05}{0.5 * 0.20} = 0.25$$

Joint Probability Distributions grow exponentially with # of features! For binary-valued features, we need $O(2^p)$ JPDs for each class.

Figure 29.8 Example for Bayes Theorem

29.10 Bayesian Classification: Why?

The following are the reasons why Bayesian classification s important.

- Probabilistic learning: We can calculate explicit probabilities for hypothesis, and is among the most practical approaches to certain types of learning problems.
- Incremental: Each training example can incrementally increase/decrease the probability that a hypothesis is correct. Prior knowledge can also be combined with observed data.
- Probabilistic prediction: It is possible to predict multiple hypotheses, weighted by their probabilities.
- Standard: Even when Bayesian methods are computationally intractable, they can provide a standard of optimal decision making against which other methods can be measured

29.10.1 Bayes Optimal Classifier

We will now briefly outline the different types of Bayesian learning. The forthcoming modules will discuss these methods in detail. One great advantage of Bayesian Decision Theory is that it gives us a lower bound on the classification error that can be obtained for a given problem.

In **Bayes optimal classification** the most probable classification of a new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities as given below:

$$\textbf{\textit{argmax}}_{v_j \in V} \Sigma_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

where V is the set of all the values a classification can take and v_j is one of the possible classifications from this set. Unfortunately, Bayes Optimal Classifier is usually too costly to apply! The compromise is the ***Naïve Bayes Classifier*** which we will discuss in detail later.

29.10.2 Bayesian Belief Networks

The ***Bayes Optimal Classifier*** is often too costly to apply. The ***Naïve Bayes Classifier*** uses the conditional independence assumption to defray these costs. However, in many cases, such an assumption is overly restrictive. ***Bayesian belief networks*** provide an ***intermediate*** approach which allows stating conditional independence assumptions that apply to ***subsets*** of the variable.

29.10.3 Expectation-Maximization (EM)

Now we will have a brief look at Expectation-Maximization approach. Consider learning a naïve Bayes classifier using *unlabeled* data. How can we estimate *e.g.* P(A|C)? (Figure 29.9)

Initialization: randomly assign values to P(C), P(A|C), P(B|C)

repeat {

E-step: Here we compute P(C|A,B):

M-step: Then we re-compute maximum likelihood estimation of P(C), P(A|C), P(B|C)

Calculate log likelihood of data

} **until** (likelihood of data does not improve)

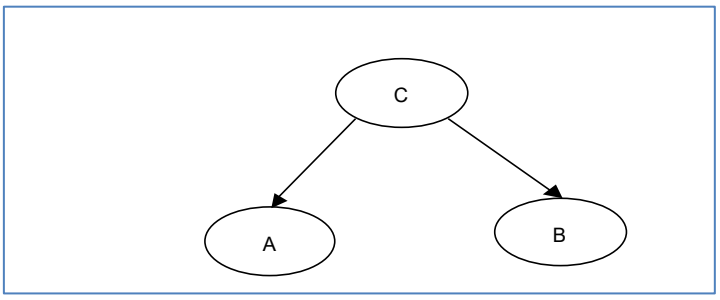


Figure 29.9 Dependencies of the variable A,B & C

Summary

- Explained Probability, Bayes Theorem and Bayes Learning
- Discussed Bayes Learning Framework
- Outlined Maximum A Posteriori (MAP) & Maximum Likelihood (ML)