UNIT-IV

BAYESIAN LEARNING

1. INTRODUCTION

Bayesian learning provides the basis for learning algorithms that directly manipulate probabilities.

Features of Bayesian learning methods

- Each observed training example can incrementally decrease or increase the estimated probability of the correct hypothesis.
- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
- Bayesian methods can accommodate hypotheses that make probabilistic predictions (e.g., hypotheses such as "this pneumonia patient has a 93% chance of complete recovery").
- New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
- Even in cases where Bayesian methods prove computationally intractable, they can provide a standard of optimal decision making against which other practical methods can be measured.

2. BAYES THEOREM

Bayes theorem states that

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

Where $P(h|D)$ denotes the posteriori probability of h given D $P(D|h)$ denotes the posteriori probability of D given h $P(h)$ denotes the prior probability of h and $P(D)$ denotes the prior probability of D

- In many learning scenarios, the learner considers some set of candidate hypotheses **H** and is interested in finding the most probable hypothesis ***h ϵ H*** given the observed data ***D.***
- The maximally probable hypothesis is called a ***maximum*** a ***posteriori*** (MAP) hypothesis. We can determine the MAP hypotheses (***hMAP)*** by using Bayes theorem as follows:

$$h_{MAP} \equiv \underset{h \in H}{\operatorname{argmax}} P(h|D)$$

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

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

Notice in the final step we dropped the term ***P(D)*** because it is a constant for all ***h.***

In some cases, we will assume that every hypothesis in H is equally probable. In this case we can further simplify Equation (2) and need only consider the term ***P(D|h)*** to find the most probable hypothesis. ***P(D|h)*** is often called the ***likelihood*** of the data ***D*** given ***h,*** and any hypothesis that maximizes ***P(D|h)*** is called a ***maximum likelihood*** (ML) hypothesis, h_{ML}.

$$h_{ML} \equiv \underset{h \in H}{\operatorname{argmax}} P(D|h)$$

An Example:

To illustrate Bayes theorm, consider a medical diagnosis problem in which there are two alternative hypotheses: (1) the patient has cancer and (2) the patient has no cancer. The avaiiable data is from a particular laborator test with two possible outcomes: (positive) and (negative). We have prior knowledge that over the entire population of people only 0.008 have this disease. Furthermore, the test returns a correct positive result in only 98% of the cases in which the disease is actually present and a correct negative

result in only 97% of the cases in which the disease is not present. In other cases, the test returns the opposite result. Suppose we now observe a new patient for whom the lab test returns a positive result. Should we diagnose the patient as having cancer or not? **Sol:** The above situation can be summarized by the following probabilities:

$$P(cancer) = .008, \quad P(\neg cancer) = .992$$
$$P(\oplus|cancer) = .98, \quad P(\ominus|cancer) = .02$$
$$P(\oplus|\neg cancer) = .03, \quad P(\ominus|\neg cancer) = .97$$

The maximum a posteriori hypothesis can be found using Equation (2):

$$P(\oplus|cancer)P(cancer) = (.98).008 = .0078$$
$$P(\oplus|\neg cancer)P(\neg cancer) = (.03).992 = .0298$$

$$\text{Thus, } h_{MAP} = \neg cancer.$$

This means that, the patient has no cancer.

3. MAXIMUM LIKELIHOOD AND LEAST-SQUARED ERROR HYPOTHESES

In the previous section we saw the maximum likelihood hypothesis which is given as:

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}}\, p(D|h)$$

Here, we assume that the probability distribution is a normal distribution. A Normal distribution is a smooth, bell-shaped distribution that can be completely characterized by its mean μ and its standard deviation σ^2.

We assume a fixed set of training instances ***(x1 . . . xm)*** and therefore consider the data D to be the corresponding sequence of target values ***D = (d1 . . .dm).*** Then we can write ***P(D|h)*** as the product of the various ***p(di|hi)***

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} p(d_i|h)$$

Since, we are assuming the normal distribution for the probabilities, the above equation can be written as

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-\mu)^2}$$

$$= \underset{h \in H}{\operatorname{argmax}} \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(d_i-h(x_i))^2}$$

Applying logarithm to the above equation, we get

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} \ln \frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

The first term in this expression is a constant independent of ***h,*** and can therefore be discarded, yielding

$$h_{ML} = \underset{h \in H}{\operatorname{argmax}} \sum_{i=1}^{m} -\frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

Maximizing this negative quantity is equivalent to minimizing the corresponding
positive quantity. Hence, above equation becomes

$$h_{ML} = \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} \frac{1}{2\sigma^2}(d_i - h(x_i))^2$$

Finally, we can again discard constants that are independent of ***h*** to get

$$h_{ML} = \underset{h \in H}{\operatorname{argmin}} \sum_{i=1}^{m} (d_i - h(x_i))^2$$

Thus, the above equation shows that the maximum likelihood hypothesis ***hML*** the one that minimizes the sum of the squared errors between the observed training values ***di*** and the hypothesis predictions ***h(xi).***

4. BAYES OPTIMAL CLASSIFIER

- Before defining the Bayes optimal classifier, let us consider a hypothesis space containing three hypotheses, hl, h2, and h3. Suppose that the posterior probabilities of these hypotheses given the training data are **0.**4, **0.3,** and **0.3** respectively.
- Thus, hl is the MAP hypothesis.
- Suppose a new instance x is encountered, which is classified positive by ***h1,*** but negative by ***h2*** and ***h3***.
- Taking all hypotheses into account, the probability that x is positive is 0.4 (the probability associated with ***h1)***, and the probability that it is negative is therefore 0.6. The most probable classification (negative) in this case is different from the classification generated by the MAP hypothesis.
- In general, the most probable classification of the new instance is obtained by combining the predictions of all hypotheses, weighted by their posterior probabilities.
- If the possible classification of the new example can take on any value ***vj*** from some set V, then the probability P(vj | D) is given as

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D)$$

The optimal classification of the new instance is the value ***vj,*** for which P (vj | D) is maximum.

Thus Bayes optimal classifier is defined as

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D)$$

To illustrate in terms of the above example, the set of possible classifications of the new instance is

$$V = \{\oplus, \ominus\}$$

and

$$P(h_1|D) = .4, \quad P(\ominus|h_1) = 0, \quad P(\oplus|h_1) = 1$$

$$P(h_2|D) = .3, \quad P(\ominus|h_2) = 1, \quad P(\oplus|h_2) = 0$$

$$P(h_3|D) = .3, \quad P(\ominus|h_3) = 1, \quad P(\oplus|h_3) = 0$$

Therefore

$$\sum_{h_i \in H} P(\oplus|h_i) P(h_i|D) = .4$$

$$\sum_{h_i \in H} P(\ominus|h_i) P(h_i|D) = .6$$

And

$$\underset{v_j \in \{\oplus, \ominus\}}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i) P(h_i|D) = \ominus$$

Any system that classifies new instances according to Equation (5.1) is called a ***Bayes*** optimal ***classifier,*** or Bayes optimal learner.

5. BAYESIAN BELIEF NETWORKS (BBN)

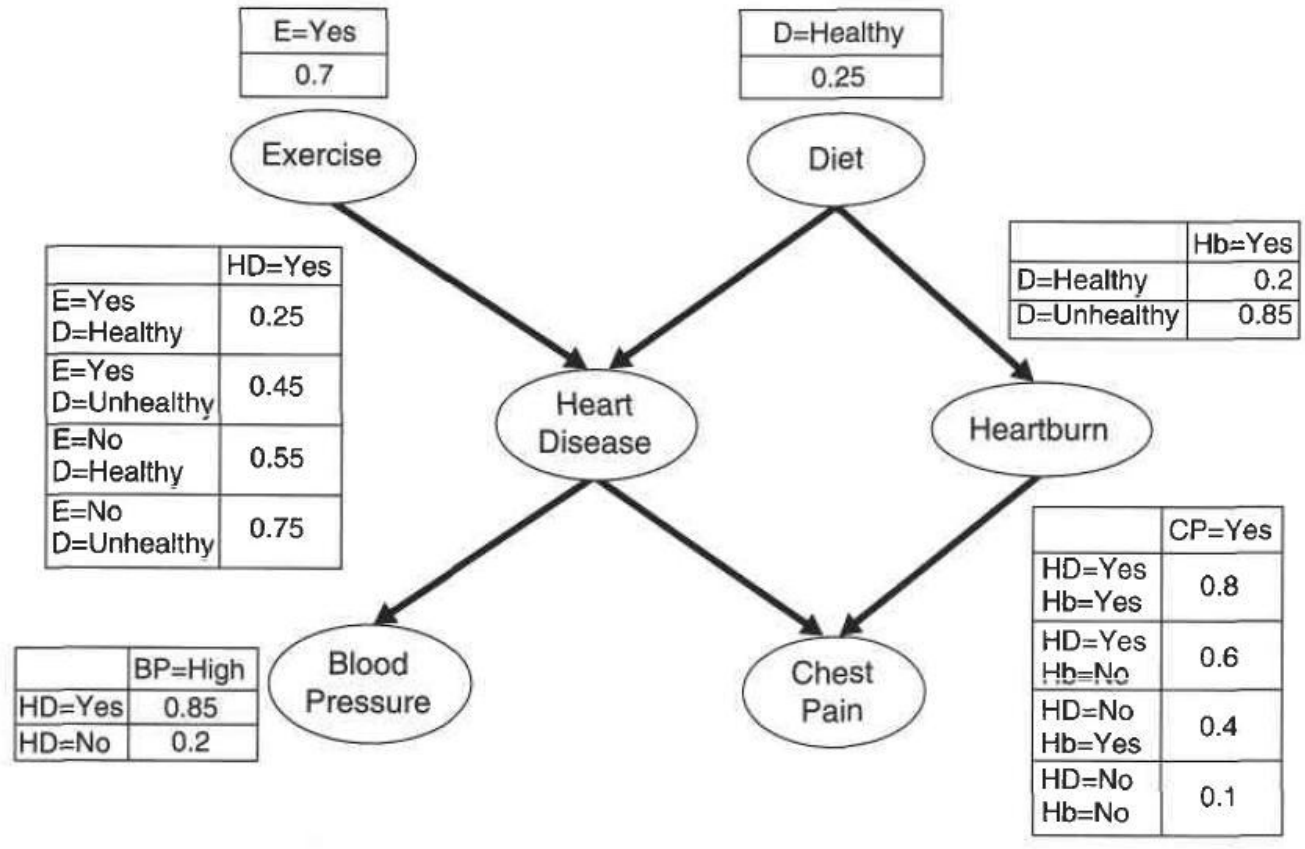
- The conditional independence assumption made by naive Bayes classifiers may seem too rigid, especially for classification problems in which the attributes are somewhat correlated (i.e. depended).
- Instead of requiring all the attributes to be conditionally independent given the class, this approach allows us to specify which pair of attributes is conditionally independent and which pair of attributes is conditionally dependent.

A Bayesian belief network (Bayesian network for short) represents the joint probability distribution for a set of variables.

Model Representation

- A Bayesian belief network (BBN), or simply, Bayesian network, provides a graphical representation of the probabilistic relationships among a set of random variables. There are two key elements of a Bayesian network:

1. A directed acyclic graph (dag) encoding the dependence relationships among a set of variables. 2. A conditional probability table (CPT) associating each variable in the network.



For example, the above figure shows the Bayesian belief network for detecting HeartDisease and HeartBurn in patients. Each variable in the diagram is assumed to be binary-valued. The parent nodes for heart disease (HD) correspond to risk factors that may affect the disease, such as exercise (E) and diet (D). The child nodes for heart disease correspond to symptoms of the disease, such as chest pain (CP) and high blood pressure (BP). For example, the diagram shows that heartburn (Hb) may result from an unhealthy diet and may lead to chest pain.

Model Building

- Model building in Bayesian networks involves two steps: (1) creating the structure of the network, and (2) estimating the probability values in the tables associated with each node. The network topology can be obtained by encoding the subjective knowledge of domain experts.

Example of Inferencing Using BBN

- Suppose we are interested in using the BBN shown above to diagnose whether a person has heart disease. The following cases illustrate how the diagnosis can be made under different scenarios.

Case 1: No Prior Information Without any prior information, we can determine whether the person is likely to have heart disease by computing the prior probabilities P(HD = Yes) and P(HD = No). To simplify the notation, let α Є{Yes, No} denote the binary values of Exercise and β Є {Healthy, Unhealthy} denote the binary value of Diet.

$$
\begin{aligned}
P(\mathtt{HD} = \mathtt{Yes}) &= \sum_{\alpha}\sum_{\beta} P(\mathtt{HD} = \mathtt{Yes}|E = \alpha, D = \beta)P(E = \alpha, D = \beta) \\
&= \sum_{\alpha}\sum_{\beta} P(\mathtt{HD} = \mathtt{Yes}|E = \alpha, D = \beta)P(E = \alpha)P(D = \beta) \\
&= 0.25 \times 0.7 \times 0.25 + 0.45 \times 0.7 \times 0.75 + 0.55 \times 0.3 \times 0.25 \\
&\quad + 0.75 \times 0.3 \times 0.75 \\
&= 0.49.
\end{aligned}
$$

Since P(HD = No) = 1 - P(HD = Yes) = 0.51, the person has a slightly higher chance of not getting the disease. **Case 2: High Blood Pressure** If the person has high blood pressure) we can make a diagnosis about heart disease by comparing the posterior probabilities, P(HD = Yes | BP = High) against P(HD = No | BP = High). To do this, we must compute P(BP = High):

$$P(\texttt{BP} = \texttt{High}) = \sum_{\gamma} P(\texttt{BP} = \texttt{High}|\texttt{HD} = \gamma)P(\texttt{HD} = \gamma)$$
$$= 0.85 \times 0.49 + 0.2 \times 0.51 = 0.5185.$$

where $\gamma \in \{\texttt{Yes}, \texttt{No}\}$. Therefore, the posterior probability the person has heart disease is

$$P(\texttt{HD} = \texttt{Yes}|\texttt{BP} = \texttt{High}) = \frac{P(\texttt{BP} = \texttt{High}|\texttt{HD} = \texttt{Yes})P(\texttt{HD} = \texttt{Yes})}{P(\texttt{BP} = \texttt{High})}$$
$$= \frac{0.85 \times 0.49}{0.5185} = 0.8033.$$

Similarly, $P(\texttt{HD} = \texttt{No}|\texttt{BP} = \texttt{High}) = 1 - 0.8033 = 0.1967$. Therefore, when a person has high blood pressure, it increases the risk of heart disease.

Characteristics of BBN Following are some of the general characteristics of the BBN method:

1. BBN provides an approach for capturing the prior knowledge of a particular domain using a graphical model. The network can also be used to encode causal dependencies among variables.

2. Constructing the network can be time consuming and requires a large amount of effort. However, once the structure of the network has been determined, adding a new variable is quite straightforward.

3. Bayesian networks are well suited to dealing with incomplete data. Instances with missing attributes can be handled by summing or integrating the probabilities over all possible values of the attribute.

4. Because the data is combined probabilistically with prior knowledge, the method is quite robust to model overfitting.

UNIT-IV
SECTION-A

Objective Questions

1. Bayes error is the_____bound of probability of classification error. [ ]
A. Lower B. Upper

2. Bayes decision rule is the theoretically …….. classifier that minimize probability of classification error. [ ]
A. Best B. Worst C. Average

3. In Bayes Theorem, unconditional probability is called as [ ]
A. Evidence b. Likelihood c. Prior d. Posterior

4. In Bayes Theorem, Class conditional probability is called as [ ]
A. Evidence B. Likelihood c. Prior d. Posterior

5. Bayesian reasoning provides a _____approach to inference. [ ]
A. deterministic B. Probabilistic C. both A and B D. none of the above

6. Previous probabilities in Bayes Theorem that are changed with help of new available information are classified as [ ]
A. independent probabilities B. posterior probabilities
C. independent probabilities D. dependent probabilities

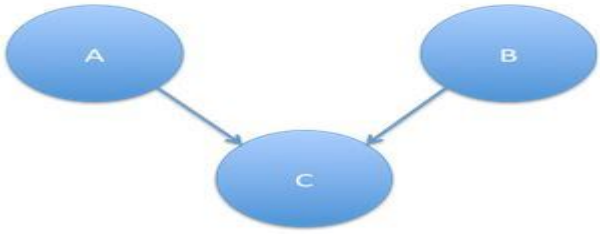
7. Three components of Bayes decision rule are class prior, likelihood and[ ]
A. Evidence B. Instance C. Confidence D. Salience

8. $P(\vec{X})P(wi/\vec{X})$ = [ ]

A. $P(1-\vec{X})P(wi/\vec{X})$ B. $P(\vec{X})P(1-wi/\vec{X})$ C. $P(\vec{X}|wi)P(wi)$ D. $P(\vec{X}-wi)P(wi/\vec{X})$

9. A and B are Boolean random variables. Given: P(A=True) = 0.3, P(A=False) = 0.7, P(B=True|A=True) = 0.4, P(B=False|A=True) = 0.6, P(B=True|A=False) = 0.6, P(B=False|A=False) = 0.4 . Calculate P(A=True|B=False) by Bayes rule.
[ ]
A.0.49 B. 0.39 C. 0.37 D.0.28

10. In the following Bayesian network A, B and C are Boolean random variables taking values in {True, False}.



Which of the following statements is true? [ ]

A. The value of C is not given. If the value of B changes from True to False, the conditional probability of A, P(A|B) changes.

B. The value of C is given to be True. If the value of B changes from True to False, the conditional probability of A, P(A|B) changes.

C. Neither A nor B D. Both A and B

11. Diabetic Retinopathy is a disease that affects 80% people who have diabetes for more than 10 years. 5% of the Indian population has been suffering from diabetes for more than 10 years. Answer the following questions. What is the joint probability of finding an Indian suffering from Diabetes for more than 10 years and also has Diabetic Retinopathy? [ ]

A. 0.024 B. 0.040 C. 0.076 D. 0.005

12. Which of the following properties is false in the case of a Bayesian Network: [ ]

A. The edges are directed
B. Contains cycles
C. Represents conditional independence relations among random variables
D. All of the above

SECTION-B

Descriptive Questions

1. Define the concept of Conditional Independence.
2. What is Bayes theorem? Explain how this is used in computing MAP and Maximum likelihood hypothesis?
3. Write the features of Bayesian learning methods.
3. Explain Naive Bayes Classifier with example.
4. Write a short note on Bayesian Belief Networks.
5. How is Naive Bayes algorithm useful for learning and classifying text?
6. Describe maximum likelihood and least-squared error hypotheses

7. Explain minimum description length principle.
8. How the gradient search can be performed to maximize likelihood in a neural net.
9. Illustrate the steps for Brute-force MAP learning algorithm
10. Explain about posteriori probability in Bayes theorem