Module-4

7 a. Explain Naïve Bayes classifier.
b. Explain brute force MAP learning algorithm.

NAIVE BAYES CLASSIFIER

- The naive Bayes classifier applies to learning tasks where each instance x is described by a conjunction of attribute values and where the target function f (x) can take on any value from some finite set V.
- A set of training examples of the target function is provided, and a new instance is presented, described by the tuple of attribute values (a_1, a_2.. .a_m).
- The learner is asked to predict the target value, or classification, for this new instance.

The Bayesian approach to classifying the new instance is to assign the most probable target value, V_{MAP}, given the attribute values (a_1, a_2.. .a_m) that describe the instance

$$v_{MAP} = \underset{v_j \in V}{\text{argmax}}\, P(v_j|a_1, a_2 \ldots a_n)$$

Use Bayes theorem to rewrite this expression as

$$v_{MAP} = \underset{v_j \in V}{\text{argmax}} \frac{P(a_1, a_2 \ldots a_n|v_j)P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \underset{v_j \in V}{\text{argmax}}\, P(a_1, a_2 \ldots a_n|v_j)P(v_j) \qquad \textbf{equ (1)}$$

- The naive Bayes classifier is based on the assumption that the attribute values are conditionally independent given the target value. Means, the assumption is that given the target value of the instance, the probability of observing the conjunction (a_1, a_2.. .a_m), is just the product of the probabilities for the individual attributes:

$$P(a_1, a_2 \ldots a_n|v_j) = \prod_i P(a_i|v_j)$$

Substituting this into Equation (1),

Naive Bayes classifier:

$$\mathbf{V_{NB} = \underset{v_j \in V}{\text{argmax}}\, P(v_j) \prod_i P(a_i|v_j)} \qquad \text{equ (2)}$$

Where, V_{NB} denotes the target value output by the naive Bayes classifier

BRUTE-FORCE MAP LEARNING algorithm:

1. For each hypothesis h in H, calculate the posterior probability

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

2. Output the hypothesis h_{MAP} with the highest posterior probability

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

In order specify a learning problem for the BRUTE-FORCE MAP LEARNING algorithm we must specify what values are to be used for P(h) and for P(D|h) ?

Let's choose P(h) and for P(D|h) to be consistent with the following assumptions:

- The training data D is noise free (i.e., $d_i = c(x_i)$)
- The target concept c is contained in the hypothesis space H
- Do not have a priori reason to believe that any hypothesis is more probable than any other.

What values should we specify for P(h)?

- Given no prior knowledge that one hypothesis is more likely than another, it is reasonable to assign the same prior probability to every hypothesis h in H.
- Assume the target concept is contained in H and require that these prior probabilities sum to 1.

$$P(h) = \frac{1}{|H|} \text{ for all } h \in H$$

What choice shall we make for P(D|h)?

- P(D|h) is the probability of observing the target values $D = (d_1 \ldots d_m)$ for the fixed set of instances $(x_1 \ldots x_m)$, given a world in which hypothesis h holds
- Since we assume noise-free training data, the probability of observing classification d_i given h is just 1 if $d_i = h(x_i)$ and 0 if $d_i \neq h(x_i)$. Therefore,

$$P(D|h) = \begin{cases} 1 & \text{if } d_i = h(x_i) \text{ for all } d_i \in D \\ 0 & \text{otherwise} \end{cases}$$

Given these choices for P(h) and for P(D|h) we now have a fully-defined problem for the above BRUTE-FORCE MAP LEARNING algorithm.

Recalling Bayes theorem, we have

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

Consider the case where h is inconsistent with the training data D

$$P(h|D) = \frac{0 \cdot P(h)}{P(D)} = 0$$

The posterior probability of a hypothesis inconsistent with D is zero

Consider the case where h is consistent with D

$$P(h|D) = \frac{1 \cdot \frac{1}{|H|}}{P(D)} = \frac{1 \cdot \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} = \frac{1}{|VS_{H,D}|}$$

Where, $VS_{H,D}$ is the subset of hypotheses from H that are consistent with D

To summarize, Bayes theorem implies that the posterior probability P(h|D) under our assumed P(h) and P(D|h) is

$$P(D|h) = \begin{cases} \frac{1}{|V S_{H,D}|} & \text{if h is consistent with D} \\ 0 & \text{otherwise} \end{cases}$$

Discuss Minimum Description Length principle in brief.
Explain Bayesian belief networks and conditional independence with example.

MINIMUM DESCRIPTION LENGTH PRINCIPLE

- A Bayesian perspective on Occam's razor
- Motivated by interpreting the definition of h_{MAP} in the light of basic concepts from information theory.

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

which can be equivalently expressed in terms of maximizing the $\log_2$

$$h_{MAP} = \underset{h \in H}{argmax}\ \log_2 P(D|h) + \log_2 P(h)$$

or alternatively, minimizing the negative of this quantity

$$h_{MAP} = \underset{h \in H}{argmin}\ -\log_2 P(D|h) - \log_2 P(h) \qquad \text{equ (1)}$$

This equation (1) can be interpreted as a statement that short hypotheses are preferred, assuming a particular representation scheme for encoding hypotheses and data

- $-\log_2 P(h)$: the description length of h under the optimal encoding for the hypothesis space H, $L_{C_H}(h) = -\log_2 P(h)$, where C_H is the optimal code for hypothesis space H.
- $-\log_2 P(D \mid h)$: the description length of the training data D given hypothesis h, under the optimal encoding from the hypothesis space H: $L_{C_H}(D|h) = -\log_2 P(D|\ h)$, where $C_{D|h}$ is the optimal code for describing data D assuming that both the sender and receiver know the hypothesis h.
- Rewrite Equation (1) to show that h_{MAP} is the hypothesis h that minimizes the sum given by the description length of the hypothesis plus the description length of the data given the hypothesis.

$$h_{MAP} = \underset{h \in H}{argmin}\ L_{C_H}(h) + L_{C_{D|h}}(D|h)$$

Where, C_H and $C_{D|h}$ are the optimal encodings for H and for D given h

The Minimum Description Length (MDL) principle recommends choosing the hypothesis that minimizes the sum of these two description lengths of equ.

$$h_{MAP} = \underset{h \in H}{argmin}\ L_{C_H}(h) + L_{C_{D|h}}(D|h)$$

Minimum Description Length principle:

$$h_{MDL} = \underset{h \in H}{\text{argmin}}\ L_{C_1}(h) + L_{C_2}(D \mid h)$$

Where, codes C_1 and C_2 to represent the hypothesis and the data given the hypothesis

The above analysis shows that if we choose C_1 to be the optimal encoding of hypotheses C_H, and if we choose C_2 to be the optimal encoding $C_{D|h}$, then $h_{MDL} = h_{MAP}$

.....

b)

BAYESIAN BELIEF NETWORKS

- The naive Bayes classifier makes significant use of the assumption that the values of the attributes $a_1 \ldots a_n$ are conditionally independent given the target value v.
- This assumption dramatically reduces the complexity of learning the target function

A Bayesian belief network describes the probability distribution governing a set of variables by specifying a set of conditional independence assumptions along with a set of conditional probabilities
Bayesian belief networks allow stating conditional independence assumptions that apply to subsets of the variables

Notation

- Consider an arbitrary set of random variables $Y_1 \ldots Y_n$, where each variable Yi can take on the set of possible values $V(Y_i)$.
- The joint space of the set of variables Y to be the cross product $V(Y_1)$ x $V(Y_2)$ x. . . $V(Y_n)$.
- In other words, each item in the joint space corresponds to one of the possible assignments of values to the tuple of variables $(Y_1 \ldots Y_n)$. The probability distribution over this joint' space is called the joint probability distribution.
- The joint probability distribution specifies the probability for each of the possible variable bindings for the tuple $(Y_1 \ldots Y_n)$.
- A Bayesian belief network describes the joint probability distribution for a set of variables.

Conditional Independence

Let X, Y, and Z be three discrete-valued random variables. X is conditionally independent of Y given Z if the probability distribution governing X is independent of the value of Y given a value for Z, that is, if

$$(\forall x_i, y_j, z_k)\ P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Where,

$$x_i \in V(X),\ y_i \in V(Y),\ \text{and}\ z_k \in V(Z).$$

The above expression is written in abbreviated form as

$$P(X \mid Y, Z) = P(X \mid Z)$$

Conditional independence can be extended to sets of variables. The set of variables $X_1 \ldots X_l$ is conditionally independent of the set of variables $Y_1 \ldots Y_m$ given the set of variables $Z_1 \ldots Z_n$ if

$$P(X_1 \ldots X_l | Y_1 \ldots Y_m, Z_1 \ldots Z_n) = P(X_1 \ldots X_l | Z_1 \ldots Z_n)$$

The naive Bayes classifier assumes that the instance attribute A_1 is conditionally independent of instance attribute A_2 given the target value V. This allows the naive Bayes classifier to calculate $P(A_1, A_2 \mid V)$ as follows,

$$\begin{aligned} P(A_1, A_2|V) &= P(A_1|A_2, V)P(A_2|V) \\ &= P(A_1|V)P(A_2|V) \end{aligned}$$

Module-4

7 a. What is Bayes theorem and maximum posterior hypothesis? (04 Marks)

b. Derive an equation for MAP hypothesis using Bayes theorem. (04 Marks)

c. Consider a football game between two rival teams: Team 0 and Team 1. Suppose Team 0 wins 95% of the time and Team 1 wins the remaining matches. Among the games won by team 0, only 30% of them come from playing on teams 1's football field. On the otherhand, 75% of the victories for team 1 are obtained while playing at home. If team 1 is to host the next match between the two teams, which team will most likely emerge as the winner? (08 Marks)

Bayes' theorem is a formula that describes how to update the probabilities of hypotheses when given evidence. It follows simply from the axioms of conditional probability, but can be used to powerfully reason about a wide range of problems involving belief updates.

Given a hypothesis h and evidence D, Bayes' theorem states that the relationship between the probability of the hypothesis before getting the evidence P(h) and the probability of the hypothesis after getting the evidence P(h|D) is

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

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}$$

- P(h) = prior (initial) probability that hypothesis h holds , before we observed any training data.
- P(D) = prior probability of training data D
- P(h|D) = posterior probability of h given D (it holds after we have seen the training data D)
- P(D|h) = probability of observing data D given some world in which hypothesis h holds.

Maximum a Posteriori (MAP) Hypothesis

- In many learning scenarios, the learner considers some set of candidate hypotheses H and is interested in finding the most probable hypothesis h ∈ H given the observed data D. Any such maximally probable hypothesis is called a maximum a posteriori (MAP) hypothesis.
- Bayes theorem to calculate the posterior probability of each candidate hypothesis is h_{MAP} is a MAP hypothesis provided

$$\begin{aligned} h_{MAP} &= \underset{h \in H}{argmax}\ P(h|D) \\ &= \underset{h \in H}{argmax}\ \frac{P(D|h)P(h)}{P(D)} \\ &= \underset{h \in H}{argmax}\ P(D|h)P(h) \end{aligned}$$

- P(D) can be dropped, because it is a constant independent of h

7b<Refer others>

7c

Change 65% to 95%

Let X be the team hosting the match and Y be the winner of the match. Both X and Y can take on values from the set {0,1}. Then:

Probability Team 0 wins is $P(Y = 0) = 0.65$.

Probability Team 1 wins is $P(Y = 1) = 1 - 0.65 = 0.35$.

Probability Team 1 hosted the match it won is

$$P(X = 1|Y = 1) = 0.75.$$

Probability Team 1 hosted the match won by Team 0 is

$$P(X = 1|Y = 0) = 0.3.$$

$$\begin{aligned} P(Y = 1|X = 1) &= \frac{P(X = 1|Y = 1) \times P(Y = 1)}{P(X = 1)} \\ &= \frac{P(X = 1|Y = 1) \times P(Y = 1)}{P(X = 1, Y = 1) + P(X = 1, Y = 0)} \\ &= \frac{P(X = 1|Y = 1) \times P(Y = 1)}{P(X = 1|Y = 1)P(Y = 1) + (X = 1|Y = 0)P(Y = 0)} \\ &= \frac{0.75 \times 0.35}{0.75 \times 0.35 + 0.3 \times 0.65} \\ &= 0.5738 \end{aligned}$$

Hence forth probability of hosting teaming winning chance (0.5738) is more .

8 a. Describe Brute-force MAP learning algorithm. **(04 Marks)**

b. Discuss the Naïve Bayees classifier. **(04 Marks)**

c. The following table gives data set about stolen vehicles. Using Naïve bayes classifier classify the new data (Red, SUV, Domestic) **(08 Marks)**

Table

Color	Type	Origin	Stolen
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

8a 8b repeated

8c

The Bayes Naive classifier selects the most likely classification V_{nb} given the attribute values $a_1, a_2, \ldots a_n$. This results in:

$$V_{nb} = \text{argmax}_{v_j \in V} P(v_j) \prod P(a_i|v_j) \quad (1)$$

We generally estimate $P(a_i|v_j)$ using m-estimates:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m} \quad (2)$$

where:

n = the number of training examples for which $v = v_j$
n_c = number of examples for which $v = v_j$ and $a = a_i$
p = a priori estimate for $P(a_i|v_j)$
m = the equivalent sample size

For a given problem We need to estimate:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m}$$

We want to classify a Red Domestic SUV. Note there is no example of a Red Domestic SUV in our data set. Looking back at equation (2) we can see how to compute this. We need to calculate the probabilities

`P(Red|Yes), P(SUV|Yes), P(Domestic|Yes) ,`

`P(Red|No) , P(SUV|No), and P(Domestic|No)`

and multiply them by P(Yes) and P(No) respectively . We can estimate these values using equation

```
Yes:                    No:
    Red:                    Red:
        n = 5                   n = 5
        n_c= 3                  n_c = 2
        p = .5                  p = .5
        m = 3                   m = 3
    SUV:                    SUV:
        n = 5                   n = 5
        n_c = 1                 n_c = 3
        p = .5                  p = .5
        m = 3                   m = 3
    Domestic:               Domestic:
        n = 5                   n = 5
        n_c = 2                 n_c = 3
        p = .5                  p = .5
        m = 3                   m =3
```

Looking at $P(Red|Yes)$, we have 5 cases where v_j = Yes , and in 3 of those cases a_i = Red. So for $P(Red|Yes)$, $n = 5$ and $n_c = 3$. Note that all attribute are binary (two possible values). We are assuming no other information so, $p = 1$ / (number-of-attribute-values) = 0.5 for all of our attributes. Our m value is arbitrary, (We will use $m = 3$) but consistent for all attributes. Now we simply apply eqauation **(2)** using the precomputed values of n , n_c, p, and m.

$$P(Red|Yes) = \frac{3+3*.5}{5+3} = .56 \qquad P(Red|No) = \frac{2+3*.5}{5+3} = .43$$

$$P(SUV|Yes) = \frac{1+3*.5}{5+3} = .31 \qquad P(SUV|No) = \frac{3+3*.5}{5+3} = .56$$

$$P(Domestic|Yes) = \frac{2+3*.5}{5+3} = .43 \qquad P(Domestic|No) = \frac{3+3*.5}{5+3} = .56$$

We have $P(Yes) = .5$ and $P(No) = .5$, so we can apply equation (2). For $v = Yes$, we have

```
P(Yes) * P(Red | Yes) * P(SUV | Yes) * P(Domestic|Yes)

    =  .5 * .56 * .31 * .43   = .037
```

and for $v = No$, we have

```
P(No) * P(Red | No) * P(SUV | No) * P (Domestic | No)

    = .5 * .43 * .56 * .56 = .069
```

Since $0.069 > 0.037$, our example gets classified as 'NO'

Module-4

7 a. Explain maximum a posteriori (MAP) hypothesis using Bayes theorem. **(06 Marks)**

b. Estimate conditional probabilities of each attributes {colour, legs, height, smelly} for the species classes: {M, H} using the data given in the table. Using these probabilities estimate the probability values for the new instance – (Colour = Green, Legs = 2, Height = Tall and Smelly = No) **(10 Marks)**

No	Colour	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

OR

8 a. Explain Naive Bayes classifier and Bayseian belief networks. **(10 Marks)**

b. Prove that how maximum likelihood (Bayesian learning) can be used in any learning algorithms that are used to minimize the squared error between actual output hypothesis and predicted output hypothesis. **(06 Marks)**

7a repeated

7b https://youtu.be/z8K-598fqSo

8 a repeated

8b Refer others