

UNIT–2 (8 Marks and 16 Marks)

1. How to handle uncertain knowledge with example? And How to represent knowledge in an uncertain domain? (Dec-2013) and Define uncertain knowledge, prior probability and conditional probability. State the Baye's theorem. How it is useful for decision making under uncertainty? (May-2014)

PROBABILISTICREASONING

Uncertainty:

Till now, we have learned knowledge representation using first-order logic and propositional logic with certainty, which means we were sure about the predicates. With this knowledge representation, we might write $A \rightarrow B$, which means if A is true then B is true, but consider a situation where we are not sure about whether A is true or not then we cannot express this statement, this situation is called uncertainty. So to represent uncertain knowledge, where we are not sure about the predicates, we need uncertain reasoning or probabilistic reasoning.

Causes of uncertainty:

Following are some leading causes of uncertainty to occur in the real world.

1. Information occurred from unreliable sources.
2. Experimental Errors
3. Equipment fault
4. Temperature variation
5. Climate change.

Probabilistic reasoning:

Probabilistic reasoning is a way of knowledge representation where we apply the concept of probability to indicate the uncertainty in knowledge. In probabilistic reasoning, we combine probability theory with logic to handle the uncertainty.

- We use probability in probabilistic reasoning because it provides a way to handle the uncertainty that is the result of someone's laziness and ignorance.
- In the real world, there are lots of scenarios, where the certainty of something is not confirmed, such as "It will rain today," "behavior of someone for some situations," "A match between two teams or two players."
- These are probable sentences for which we can assume that it will happen but not sure about it, so here we use probabilistic reasoning.

Need of probabilistic reasoning in AI:

- When there are unpredictable outcomes.
- When specifications or possibilities of predicates becomes too large to handle.
- When an unknown error occurs during an experiment.

In probabilistic reasoning, there are two ways to solve problems with uncertain knowledge:

- **Bayes' rule**
- **Bayesian Statistics**

As probabilistic reasoning uses probability and related terms, so before understanding probabilistic reasoning, let's understand some common terms:

Probability: Probability can be defined as a chance that an uncertain event will occur. It is the numerical measure of the likelihood that an event will occur. The value of probability always remains between 0 and 1 that represent ideal uncertainties.

$0 \leq P(A) \leq 1$, where $P(A)$ is the probability of an event A .

$P(A) = 0$, indicates total uncertainty in an event A .

$P(A) = 1$, indicates total certainty in an event A .

We can find the probability of an uncertain event by using the below formula.

$$\text{Probability of occurrence} = \frac{\text{Number of desired outcomes}}{\text{Total number of outcomes}}$$

- $P(\neg A)$ = probability of a not happening event.
- $P(\neg A) + P(A) = 1$.

Event: Each possible outcome of a variable is called an event.

Sample space: The collection of all possible events is called sample space.

Random variables: Random variables are used to represent the events and objects in the real world.

Prior probability: The prior probability of an event is probability computed before observing new information.

Posterior Probability: The probability that is calculated after all evidence or information has taken into account. It is a combination of prior probability and new information.

Conditional probability:

Conditional probability is a probability of occurring an event when another event has already happened.

Let's suppose, we want to calculate the event A when event B has already occurred, "the probability of A under the conditions of B", it can be written as:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)}$$

Where $P(A \wedge B)$ = Joint probability of a and B
 $P(B)$ = Marginal probability of B.

If the probability of A is given and we need to find the probability of B, then it will be given as:

$$P(B|A) = \frac{P(A \wedge B)}{P(A)}$$

It can be explained by using the below Venn diagram, where B is occurred event, so sample space will be reduced to set B, and now we can only calculate event A when event B is already occurred by dividing the probability of $P(A \wedge B)$ by $P(B)$.

Bayes't heorem in Artificial intelligence

Bayes' theorem:

Bayes' theorem is also known as **Bayes' rule**, **Bayes' law**, or **Bayesian reasoning**, which determines the probability of an event with uncertain knowledge. In probability theory, it relates the conditional probability and marginal probabilities of two random events. Baye's theorem was named after the British mathematician **Thomas Bayes**. The **Bayesian inference** is an application of Baye's theorem, which is fundamental to Bayesian statistics. It is a way to calculate the value of $P(B|A)$ with the knowledge of $P(A|B)$.

- Bayes'theorem allows updating the probability prediction of an event by observing new information of the real world.
- **Example:** If cancer corresponds to one's age then by using Bayes' theorem, we can determine the probability of cancer more accurately with the help of age.
- Bayes'theorem can be derived using product rule and conditional probability of event A with known event B: As from product rule we can write:

$$P(A \cap B) = P(A|B) P(B) \text{ or}$$

Similarly, the probability of event B with known event A:

$$P(A \cap B) = P(B|A) P(A)$$

Equating right hand side of both the equations, we will get:

$$P(A|B) = \frac{P(B|A) P(A)}{P(B)} \quad \dots(a)$$

The above equation(a) is called as **Bayes' rule** or **Bayes' theorem**. This equation is basic of most modern AI systems for **probabilistic inference**.

It shows the simple relationship between joint and conditional probabilities. Here,

- **P(A|B)** is known as **posterior**, which we need to calculate, and it will be read as Probability of hypothesis A when we have occurred an evidence B.
- **P(B|A)** is called the **likelihood**, in which we consider that hypothesis is true, then we calculate the probability of evidence.
- **P(A)** is called the **prior probability**, probability of hypothesis before considering the evidence
- **P(B)** is called **marginal probability**, pure probability of evidence.

In the equation (a), in general, we can write $P(B) = \sum P(A_i) * P(B|A_i)$, hence the Bayes' rule can be written as:

$$P(A_i|B) = \frac{P(A_i) * P(B|A_i)}{\sum_{i=1}^k P(A_i) * P(B|A_i)}$$

Where $A_1, A_2, A_3, \dots, A_n$ is a set of mutually exclusive and exhaustive events.

Applying Bayes' rule:

Bayes' rule allows us to compute the single term $P(B|A)$ in terms of $P(A|B)$, $P(B)$, and $P(A)$. This is very useful in cases where we have a good probability of these three terms and want to determine the fourth one. Suppose we want to perceive the effect of some unknown cause, and want to compute that cause, then the Bayes' rule becomes:

$$P(\text{cause} | \text{effect}) = \frac{P(\text{effect} | \text{cause}) P(\text{cause})}{P(\text{effect})}$$

Application of Bayes' theorem:

- It is used to calculate the next step of the robot when the already executed step is given.
- Bayes' theorem is helpful in weather forecasting.
- It can solve the Monty Hall problem.

2. Discuss about Bayesian theory and Bayesian network. (Dec2017)

Bayesian network

- "A Bayesian network is a probabilistic graphical model which represents a set of variables and their conditional dependencies using a directed acyclic graph."
- It is also called a **Bayes network, belief network, decision network, or Bayesian model.**
- Bayesian networks are probabilistic, because these networks are built from a **probability distribution**, and also use probability theory for prediction and anomaly detection.

Real world applications are probabilistic in nature, and to represent the relationship between multiple events, we need a Bayesian network. It can also be used in various tasks including **prediction, anomaly detection, diagnostics, automated insight, reasoning, time series prediction, and decision making under uncertainty.**

➤ Bayesian Network can be used for building models from data and experts opinions, and it consists of two parts:

- **Directed Acyclic Graph**
- **Table of conditional probabilities.**

The generalized form of Bayesian network that represents and solve decision problems under uncertain knowledge is known as an Influence diagram.

A Bayesian network graph is made up of nodes and Arcs (directed links), where:

- Each node corresponds to the random variables, and a variable can be continuous or discrete.
- Arc or directed arrows represent the causal relationship or conditional probabilities between random variables. These directed links or arrows connect the pair of nodes in the graph. These links represent that one node directly influence the other node, and if there is no directed link that means that nodes are independent with each other
 - In the above diagram, A, B, C, and D are random variables represented by the nodes of the network graph.
 - If we are considering node B, which is connected with node A by a directed arrow, then node A is called the parent of Node B.
 - Node C is independent of node A.

The Bayesian network has mainly two components:

- **Causal Component**
- **Actual numbers**

Each node in the Bayesian network has condition probability distribution $P(X_i|Parent(X_i))$, which determines the effect of the parent on that node. Bayesian network is based on Joint probability distribution and conditional probability. So let's first understand the joint probability distribution:

Joint probability distribution:

If we have variables $x_1, x_2, x_3, \dots, x_n$, then the probabilities of a different combination of $x_1, x_2, x_3, \dots, x_n$, are known as Joint probability distribution.

$P[x_1, x_2, x_3, \dots, x_n]$, it can be written as the following way in terms of the joint probability distribution.
 $=P[x_1|x_2, x_3, \dots, x_n]P[x_2, x_3, \dots, x_n]$
 $=P[x_1|x_2, x_3, \dots, x_n]P[x_2|x_3, \dots, x_n] \dots P[x_{n-1}|x_n]P[x_n]$.

In general for each variable X_i , we can write the equation as:

$$P(X_i|X_{i-1}, \dots, X_1) = P(X_i|Parents(X_i))$$

3. Explain about Dempster shafer theory.(May-2017)

Dempster–Shafer Theory (DST)

- DST is a mathematical **theory of evidence** based on belief functions and plausible reasoning. It is used to combine separate pieces of information (evidence) to calculate the probability of an event.
- DST offers an alternative to traditional probabilistic theory for the mathematical representation of uncertainty.
- DST can be regarded as, a more general approach to represent uncertainty than the Bayesian approach. Bayesian methods are sometimes inappropriate

Example:

Let **A** represent the proposition "**Moore is attractive**". Then the axioms of probability insist that $P(A) + P(\neg A) = 1$. Now suppose that Andrew does not even know who "**Moore**" is, and then also, it is not fair to say that he disbelieves the proposition. It would therefore be meaningful to denote Andrew's belief **B** of **B(A)** and **B(¬A)** as both being **0**.

Dempster-Shafer Model

The idea is to allocate a number between 0 and 1 to indicate a degree of belief on a proposal as in the probability framework. However, it is not considered a probability but a belief mass. The distribution of masses is called basic belief assignment.

In other words, in this formalism a degree of belief (referred as mass) is represented as a belief function rather than a Bayesian probability distribution.

Example: Belief assignment

Suppose a system has five members, say five independent states, and exactly one of which is actual. If the original set is called S , $|S|=5$, then the set of all subsets (the power set) is called 2^S . If each possible subset is represented as a binary vector (describing any member is present or not by writing **1** or **0**), then 2^5 subsets are possible, ranging from the empty subset **(0,0,0,0,0)** to the "everything" subset **(1,1,1,1,1)**.

The "empty" subset represents a "contradiction", which is not true in any state, and is thus assigned a mass of **one**; The remaining masses are normalized so that their total is **1**. The "everything" subset is labeled as "unknown"; it represents the state where all elements are present **one**, in the sense that you cannot tell which is actual.

Belief and Plausibility

Shafer's framework allows for belief about propositions to be represented as intervals, bounded by two values, belief (or support) and plausibility:

$$\mathbf{belief} \leq \mathbf{plausibility}$$

Belief in a hypothesis is constituted by the sum of the masses of all sets enclosed by it (i.e. the sum of the masses of all subsets of the hypothesis). It is the amount of belief that directly supports a given hypothesis at least in part, forming a lower bound.

Plausibility is 1 minus the sum of the masses of all sets whose intersection with the hypothesis is empty. It is an upper bound on the possibility that the hypothesis could possibly happen, up to that value, because there is only so much evidence that contradicts that hypothesis.

Example:

A proposition say "**the cat in the box is dead.**" Suppose we have **belief of 0.5** and **plausibility of 0.8** for the proposition.

For example,

Suppose we have a belief of 0.5 for a proposition, say "the cat in the box is dead." This means that we have evidence that allows us to state strongly that the proposition is true with a confidence of 0.5. However, the evidence contrary to that hypothesis (i.e. "the cat is alive") only has a confidence of 0.2. The remaining mass of 0.3 (the gap between the 0.5 supporting evidence on the one hand, and the 0.2 contrary evidence on the other) is "indeterminate," meaning that the cat could either be dead or alive. This interval represents the level of uncertainty based on the evidence in the system.

Hypothesis	Mass	Belief	Plausibility
Neither (alive nor dead)	0	0	0
Alive	0.2	0.2	0.5
Dead	0.5	0.5	0.8
Either (alive or dead)	0.3	1.0	1.0

- The "neither" hypothesis is set to zero by definition (it corresponds to "no solution"). The orthogonal hypotheses "Alive" and "Dead" have probabilities of 0.2 and 0.5, respectively. This could correspond to "Live/Dead Cat Detector" signals, which have respective reliabilities of 0.2 and 0.5.
- Finally, the all-encompassing "Either" hypothesis (which simply acknowledges there is a cat in the box) picks up the slack so that the sum of the masses is 1. The belief for the "Alive" and "Dead" hypotheses matches their corresponding masses because they have no subsets; belief for "Either" consists of the sum of all three masses (Either, Alive, and Dead) because "Alive" and "Dead" are each subsets of "Either".
- The "Alive" plausibility is $1 - m(\text{Dead})$: 0.5 and the "Dead" plausibility is $1 - m(\text{Alive})$: 0.8. In other way, the "Alive" plausibility is $m(\text{Alive}) + m(\text{Either})$ and the "Dead" plausibility is $m(\text{Dead}) + m(\text{Either})$.
- Finally, the "Either" plausibility sums $m(\text{Alive}) + m(\text{Dead}) + m(\text{Either})$. The universal hypothesis ("Either") will always have 100% belief and plausibility—it acts as a [checksum](#) of sorts.

Plausibility in K: It is the sum of masses of set that intersects with K. i.e; $Pl(K) = m(a) + m(b) + m(c) + m(a,b) + m(b,c) + m(a,c) + m(a,b,c)$

Characteristics of Dempster Shafer Theory:

- It will ignore part such that probability of all events aggregate to 1.
- Ignorance is reduced in this theory by adding more and more evidences.
- Combination rule is used to combine various types of possibilities.

Advantages:

- As we add more information, uncertainty interval reduces.
- DST has much lower level of ignorance.
- Diagnose hierarchies can be represented using this.
- Person dealing with such problems is free to think about evidences.

Disadvantages:

- In this, computation effort is high, as we have to deal with 2^n of sets.

4. Explain about the exact inference in Bayesian networks. (May-2015)

7.2 Bayesian Learning and Inference

AU : May-14, Dec.-14

- 1) It calculates the probability of each hypothesis, given the data and makes the predictions on that basis.
- 2) Predictions are made by using all the hypotheses, weighted by their probabilities, rather than a single "best" hypothesis. This way learning is reduced to probabilistic inference.
- 3) Let D represent all the data with observed value d , then the probability of each hypothesis obtained by Bayes' Rule -

$$P(h_i|d) = \alpha P(d|h_i) P(h_i) \quad \dots (7.2.1)$$

If we want to make prediction about an unknown quantity X , then we have,

$$\begin{aligned} P(X|d) &= \sum_i P(X|d, h_i) P(h_i|d) \\ &= \sum_i P(X|h_i) P(h_i|d) \quad \dots (7.2.2) \end{aligned}$$

Where it is assumed that each hypothesis determines a probability distribution over X .

- 4) Equation (7.2.2) shows that predictions are weighted averages over the predictions of the individual hypotheses.

- 5) The important quantities in Bayesian learning are the hypothesis prior - $P(h_i)$ and the likelihood of the data under each hypothesis - $P(d|h_i)$.
- 6) The basic characteristic of Bayesian learning is that "True hypothesis dominates the Bayesian prediction".
- 7) For any fixed prior that does not rule out the true hypothesis, the posterior probability of any false hypothesis will eventually vanish, simply because the probability of generating "uncharacteristic" data indefinitely is vanishingly small.
- 8) The Bayesian prediction is optimal whether the data set be small or large.
- 9) For real learning problems, the hypothesis space is usually very large or infinite.
- 10) Approximation in Bayesian learning : -
 - i) A prediction can be made on the basis of single most probable hypothesis, h_i , that maximizes $P(h_i|d)$. This is called as maximum a posteriori or MAP hypothesis.
 - ii) Predictions made according to an MAP hypothesis h_{map} are approximately Bayesian to the extent that $P(X|d) \approx P(X|h_{\text{map}})$.
 - iii) Finding MAP hypothesis is often much easier than Bayesian learning, because it requires solving an optimization problem instead of a large summation.
 - iv) Overfitting Trade-offs :
 - a) Overfitting can occur when the hypothesis space is too expressive, so that it contains many hypotheses that fit the data set well.
 - b) Rather than placing an arbitrary limit on the hypotheses to be considered, Bayesian and MAP learning methods use the prior to penalize complexity.
 - c) Typically, more complex hypothesis have a lower prior probability-in part because there are usually many more complex hypotheses than simple hypotheses.
 - d) On the other hand, more complex hypotheses have a greater capacity to fit the data.
 - v) Hence, the hypothesis prior embodies a trade-off between the complexity of a hypothesis and its degree of fit to the data.
 - vi) If H contains only deterministic hypothesis, then in that case, $P(d|h_i)$ is 1 if h_i is consistent and 0 otherwise. Looking at equation (7.2.1) we see that h_{MAP} will then be the simplest logical theory that is consistent with the data. Therefore, maximum a posteriori learning provides a natural embodiment of Ockham's razor.

- Probabilistic Reasoning
- vii) a) Another trade-off between complexity and degree of fit is obtained by taking the logarithm of equation (7.2.1).
 - b) Choosing h_{MAP} to maximize $P(d|h_i) P(h_i)$ is equivalent to minimizing

$$-\log_2 P(d|h_i) - \log_2 P(h_i)$$

- vii) a) Another trade-off between complexity and degree of fit is obtained by taking the logarithm of equation (7.2.1).
- b) Choosing h_{MAP} to maximize $P(d|h_i) P(h_i)$ is equivalent to minimizing $-\log_2 P(d|h_i) - \log_2 P(h_i)$.
- c) Using the connection between information encoding and probability we see that the $-\log_2 P(h_i)$ term equals the number of bits required to specify the hypothesis h_i .
- d) $\log_2 P(d|h_i)$ is the additional number of bits required to specify the data given the hypothesis.
- e) To see this, consider that no bits are required if the hypothesis predicts the data exactly as with h_5 and the string of lime candies and $\log_2 1 = 0$.
- f) MAP learning is choosing the hypothesis that provides maximum compression of the data.
- g) The same task is addressed more directly by the minimum description length or MDL, learning method, which attempts to minimize the size of hypothesis and data encodings rather than work with probabilities.
- viii) a) Another approximation is provided by assuming a uniform prior over the space of hypotheses. In that case, MAP learning reduces to choosing an h_i that maximizes $P(d|H_i)$. This is called a **maximum-likelihood (ML)** hypothesis, h_{ML} .
- b) Maximum-likelihood learning is very common in statistics, a discipline in which many researchers distrust the subjective nature of hypothesis priors. It is reasonable approach when there is no reason to prefer one hypothesis over another a priori. For example, when all hypotheses are equally complex.
- c) It provides a good approximation to Bayesian and MAP learning when the data set is large, because the data swamps the prior distribution over hypotheses, but it has problems (all we shall see) with small data sets.

7.2.1 Learning with Complete Data

7.2.1.1 Maximum-Likelihood Parameter Learning : (Discrete Models)

Statistical learning methods have important task which is parameter learning with complete data. This task involves finding the numerical parameters for a

Data is said to be complete when each data points contains values for every variable in the mode.

Consider candy-bag example : -

New manufacturer : Then the lime/Cherry proportions is completely unknown.

Parameter : $\theta \in [0, 1]$ (proportion of cherry)

Hypothesis : h_θ

Assumption : All proportions equally likely a priori.

BN's variables : Flavour \in {Cherry, Lime}

N unwrapped candies ; C cherries and $l = N - C$ limes.

Likelihood of this particular data set :

$$P(d | h_\theta) = \prod_{j=1}^N P(d_j | h_\theta) = \theta^c (1 - \theta)^l$$

- Finding the maximum-likelihood hypothesis h_{ML} is then equivalent to maximising the log-likelihood :

$$L(d | h_\theta) = \log P(d | h_\theta)$$

$$= \sum_{j=1}^N \log P(d_j | h_\theta) = c \log \theta + l \log(1 - \theta)$$

To that end : 1) differentiate L with respect to θ and

2) set the resulting expression to 0.

$$\frac{dL(d | h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{l}{1 - \theta} = 0 \Rightarrow \theta = \frac{c}{c + l} = \frac{c}{N}$$

- Previous result is obvious, but the process is important : -

- 1) Write down the expression for the likelihood of the data as a function of the parameter(s).
- 2) Write down the derivative of the log-likelihood with respect to each parameter.
- 3) Find the parameter values such that the derivatives are zero.

The last step can be tricky, using iterative solution algorithms or numerical optimisation.

- **Problem with ML learning :**

If some events have never been observed, h_{ML} assigns them 0 probability.

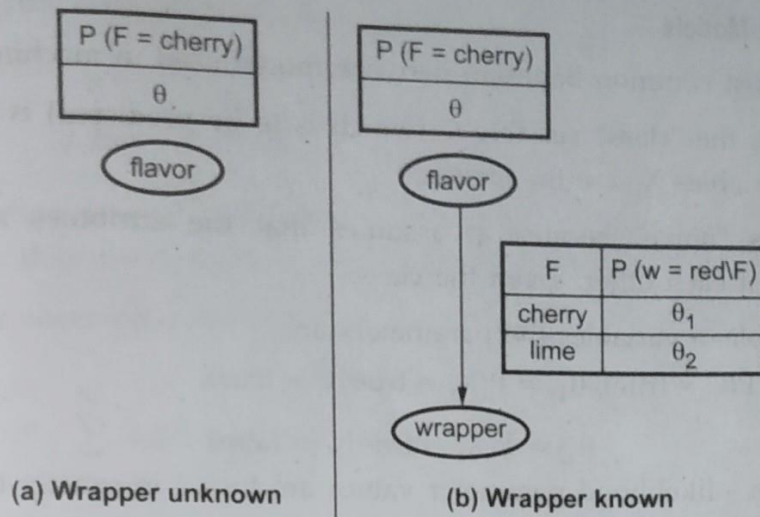


Fig. 7.2.1 Problem in ML Learning

New pb(b) : Wrappers red or blue. Colors selected probabilistically (see new BN)

$P(\text{Flavor} = \text{Cherry}, \text{Wrapper} = \text{green} \mid h_{\theta, \theta_1, \theta_2})$

$= P(\text{Flavor} = \text{cherry} \mid h_{\theta, \theta_1, \theta_2}) P(\text{Wrapper} = \text{green} \mid \text{Flavor} = \text{Cherry}, h_{\theta, \theta_1, \theta_2})$

$= \theta \cdot (1 - \theta_1)$

Experiment :

$$N = c + l = (r_c + g_c) + (r_l + g_l)$$

Likelihood :

$$P(d \mid h_{\theta, \theta_1, \theta_2}) = \theta^c (1 - \theta)^l \cdot \theta_1^{r_c} (1 - \theta_1)^{g_c} \cdot \theta_2^{r_l} (1 - \theta_2)^{g_l}$$

log-likelihood :

$$L = [\log \theta + l \log (1 - \theta)] + [r_c \log \theta_1 + g_c \log (1 - \theta_1)] + [r_l \log \theta_2 + g_l \log (1 - \theta_2)]$$

Derivatives :

$$\frac{dL}{d\theta} = \frac{c}{\theta} - \frac{l}{1 - \theta} = 0 \quad \Rightarrow \theta = \frac{c}{c + l}$$

$$\frac{dL}{d\theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1 - \theta_1} = 0 \quad \Rightarrow \theta_1 = \frac{r_c}{r_c + g_c}$$

$$\frac{dL}{d\theta_2} = \frac{r_l}{\theta_2} - \frac{g_l}{1 - \theta_2} = 0 \quad \Rightarrow \theta_2 = \frac{r_l}{r_l + g_l}$$

With complete data, ML parameter learning for a BN decomposes into separate

7.2.1.2 Naive Bayes Models

- 1) This is the most common Bayesian network model used in machine learning.
- 2) In this model, the "class" variable C (which is to be predicted) is the root and the "attribute" variables X_i are the leaves.
- 3) The model is "naive" because it assumes that the attributes are conditionally independent of each other, given the class.
- 4) Assuming Boolean variables the parameters are,

$$\theta = P(C = \text{true}), \theta_{i1} = P(X_i = \text{true} | C = \text{true}),$$

$$\theta_{i2} = P(X_i = \text{true} | C = \text{false})$$

The maximum - likelihood parameter values are found in exactly the same way as shown in Fig. 7.2.1 (b).

- 5) Once the model has been trained in this way, it can be used to classify new examples for which the class variable C is unobserved. With observed attribute values, x_1, x_2, \dots, x_n , the probability of each class is given by,

$$P(C | x_1, x_2, \dots, x_n) = \alpha P(C) \prod_i P(x_i | C)$$

- 6) A deterministic prediction can be obtained by choosing the most likely class.
- 7) The method learns fairly well but not as well as decision-tree learning; this is presumably because the true hypothesis - which is a decision tree - is not representable exactly using a naive Bayes model.
- 8) Naive Bayes learning do well in a wide range of applications; the boosted version is one of the most effective general-purpose learning algorithm.
- 9) Naive Bayes learning scales well to very large problems : with n Boolean attributes, there are just $2n+1$ parameters, and no search is required to find h_{ML} , the maximum - likelihood naive Bayes hypothesis.
- 10) Naive Bayes learning has no difficulty with noisy data and can give probabilistic predictions when appropriate.

7.2.1.3 Maximum-Likelihood Parameter Learning : (Continuous Models)

- 1) Continuous probability model such as the linear-Gaussian model is used for maximum - likelihood parameter learning.
- 2) Because continuous variables are ubiquitous in real-world applications, it is important to know how to learn continuous models from data.
- 3) The principles for maximum likelihood learning are identical to those of the discrete case.
- 4) a) Let us begin with a very simple case : Learning the parameters of a Gaussian

- b) That is, the data are generated as follows :-

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

The parameters of this model are the mean μ and the standard deviation σ . (Notice that the normalizing "constant" depends on σ , so we cannot ignore it.)

- c) Let the observed values be x_1, \dots, x_N . Then the log likelihood is,

$$\begin{aligned} L &= \sum_{j=1}^N \log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_j-\mu)^2}{2\sigma^2}} \\ &= N(-\log \sqrt{2\pi} - \log \sigma) - \sum_{j=1}^N \frac{(x_j-\mu)^2}{2\sigma^2} \end{aligned}$$

- d) Setting the derivatives to zero as usual, we obtain,

$$\frac{\delta L}{\delta \mu} = -\frac{1}{\sigma^2} \sum_{j=1}^N (x_j - \mu) = 0 \quad \Rightarrow \quad \mu = \frac{\sum_j x_j}{N}$$

$$\frac{\delta L}{\delta \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3} \sum_{j=1}^N (x_j - \mu)^2 = 0 \quad \Rightarrow \quad \sigma = \sqrt{\frac{\sum_j (x_j - \mu)^2}{N}}$$

- 5) The maximum-likelihood value of the mean is the sample average and the maximum-likelihood value of the standard deviation is the square root of the sample variance. Again, these are comforting results that confirm "commonsense" practice.
- 6) a) Now consider a linear Gaussian model with one continuous parent X and a continuous child Y .
- b) Y has a Gaussian distribution whose mean depends linearly on the value of X and whose standard deviation is fixed.
- c) To learn the conditional distribution $P(Y|X)$, we can maximize the conditional likelihood.

$$P(y|x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(y - (\theta_1 x + \theta_2))^2}{2\sigma^2}}$$

- d) Here, the parameters are θ_1 , θ_2 and σ . The data are a collection of (x_j, y_j) pairs, as shown in Fig. 7.2.4.

find the maximum-likelihood values of the

7.2.1.4 Bayesian Parameter Learning

ML learning is simple, but not appropriate for small data sets.

Example -

If only cherries have been observed, $h_{ML} \rightarrow \theta = 1.0$

- Bayesian Approach :

- It uses hypothesis prior over possible values of the parameters.
- Update of this distribution is used as the data arrives.

- Candy example with Bayesian view : -

- θ : unknown value of a variable Θ .
- Hypothesis prior : $P(\Theta)$. (Continuous over $[0, 1]$ and integrating to 1).

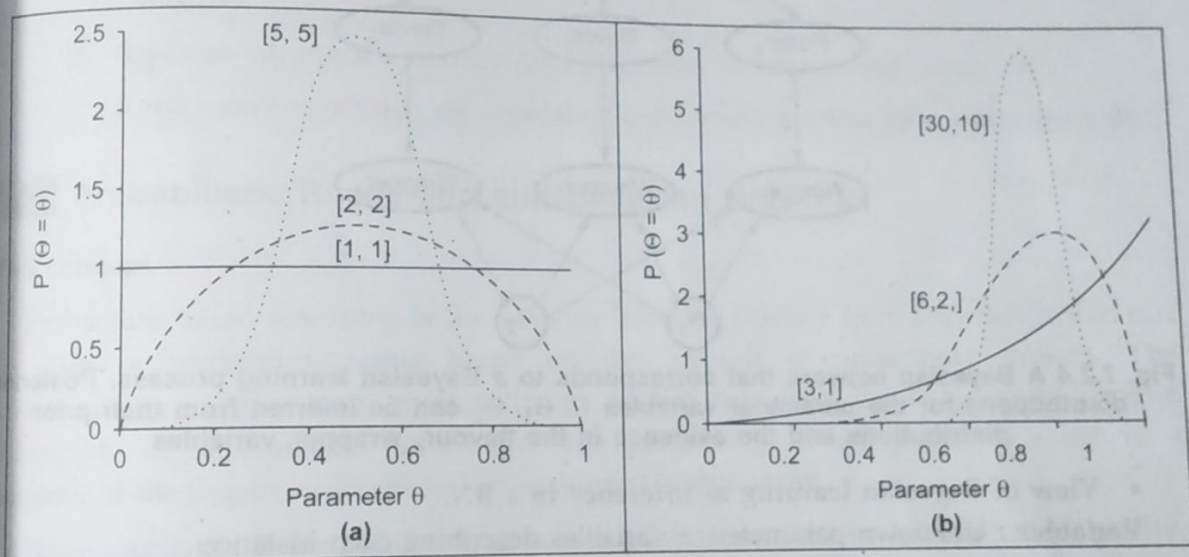


Fig. 7.2.3 (a) and (b) Examples of the beta $[a, b]$ distribution for different values of $[a, b]$

- Candidates :

beta distributions, defined by 2 hyperparameters a and b

such that :

$$\text{beta}[a, b](\theta) = \alpha \theta^{a-1} (1-\theta)^{b-1}$$

- Nice property of the beta family :

If Θ has prior beta $[a, b]$ and a data point is observed, then the posterior for Θ is also a beta distribution.

Beta family :

A beta family is called as the **conjugate prior** for the family of distributions for a Boolean variable.

$$\begin{aligned}
 P(\theta \mid D_1 = \text{Cherry}) &= \alpha P(D_1 = \text{Cherry} \mid \theta) P(\theta) \\
 &= \alpha' \theta \cdot \text{beta}[a, b](\theta) = \alpha' \theta \cdot \theta^{a-1} (1-\theta)^{b-1} \\
 &= \alpha' \theta^a (1-\theta)^{b-1} = \text{beta}[a+1, b](\theta)
 \end{aligned}$$

Note : a and b are virtual counts (starting with beta [1, 1]).

With wrappers : 3 parameters. It need to specify $P(\theta, \theta_1, \theta_2)$.

Assuming parameter independence : $P(\theta, \theta_1, \theta_2) = P(\theta) P(\theta_1) P(\theta_2)$

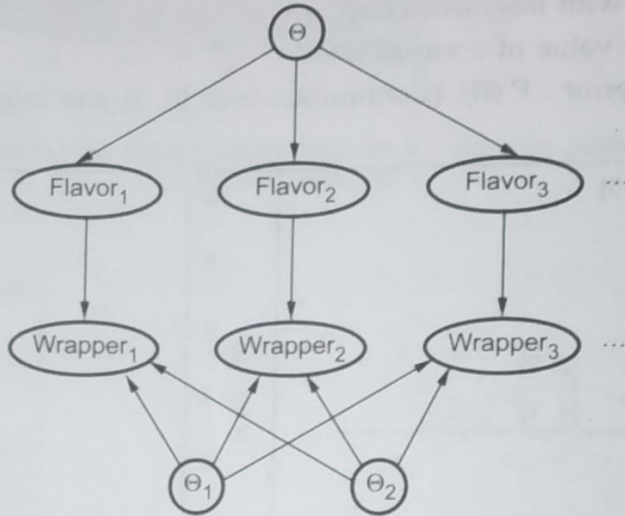


Fig. 7.2.4 A Bayesian network that corresponds to a Bayesian learning process. Posterior distributions for the parameter variables $\theta, \theta_1, \theta_2$ can be inferred from their prior distributions and the evidence in the flavour_i, wrapper_i variables

- View of Bayesian learning as inference in a BN.

Variables : Unknown parameters + variables describing each instance.

$$P(\text{Flavor}_i = \text{Cherry} \mid \theta = \theta) = \theta$$

$$P(\text{Wrapper}_i = \text{red} \mid \text{Flavor}_i = \text{Cherry}, \theta_1 = \theta_1) = \theta_1$$

7.2.1.5 Learning Bayes Net Structures

Often the Bayes net structures are easy to get from expert knowledge. But sometimes the causality relationships are debatable.

For example :

Smoking \Rightarrow Cancer ?

too Much TV \Rightarrow Bad at school ?

- To search for a good model :
 - Start with a linkless model and add parents to each node, then learn parameters and measure accuracy of the resulting model.
 - Start with an initial guess of the structure, and use hill-climbing or simulated