Lecture Notes

Probabilistic Graphical Models

This lecture notes covers the 1st module of the course Probabilistic Graphical Models.

Introduction to Probabilistic Graphical Models

Basics of Probability

Probability is a measure of the likelihood of the occurrence of an event. Measures of probability range from 0 to 1; zero means an impossible event and one means a certain event.

Terminology:

1. **Trial or experiment**: An action whose result is uncertain. For example, throwing a dice, tossing a coin, etc.
2. **Event**: A single result of an experiment
3. **Sample space**: The total number of possible outcomes of an experiment
4. **Sample point**: One of the possible outcomes

The probability of an event = Number of favourable outcomes/Total number of possible outcomes

Steps to find probability:

1: List every possible outcome of the experiment being performed.

2: Count every possible outcome of the experiment.

3: Count all the favourable outcomes.

4: Use the probability formula to determine the probability of the occurrence of an event.

Joint probability is the probability of two events taking place simultaneously. Simply put, it is the probability that event X occurs at the same time as event Y.

The joint probability of two events X and Y is $P(X \cap Y)$ also represented as $P(X, Y)$.

When the events are independent of each other, $P(X \cap Y) = P(X, Y) = P(X).P(Y)$

Note that independence between X and Y states that the occurrence of X does not affect the occurrence of Y.

When the 3 events X, Y and Z are independent of each other,

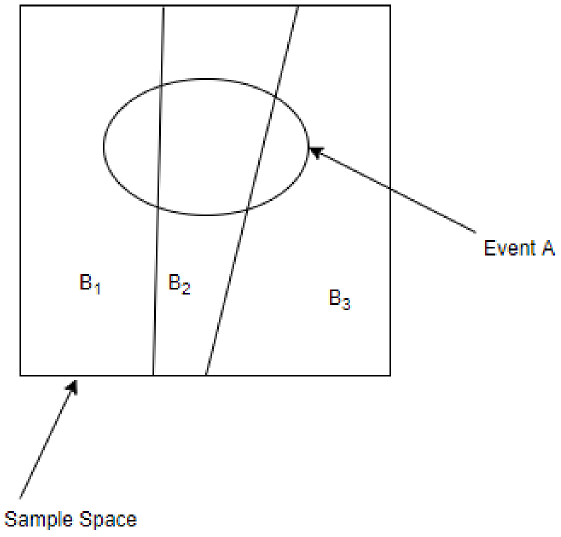
$P(X, Y, Z) = P(X).P(Y).P(Z)$

The joint probability can often be represented as a joint probability distribution table.

Multiplication rule states that

$P(A \cap B) = P(A/B).\ P(B)$

Now let's consider a situation where the sample space is divided into 3 disjoint regions as shown below:



Disjoint regions imply that there is no intersection between the regions, i.e

$B1 \cap B2 \cap B3=0$

Hence, we can write P(A) as

$P(A)=P(A \cap B1)+P(A \cap B2)+P(A \cap B3)$

By the multiplication rule, we can write

$P(A)=P(A|B1).P(B1)+P(A \cap B2).P(B2)+P(A \cap B3).P(B3)$

Bayes' theorem describes the probability of an event, based on prior knowledge of conditions that might be related to the event. If you know the conditional probability P(B|A), you can use Bayes' rule to find out the reverse probability P(A|B).

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

Introduction to Probabilistic Graphical Models

As the name suggests, Probabilistic Graphical Models has two components: Probabilities and Graphs. You must have already figured out that a Probabilistic Graphical Model, often referred to as PGM, is an interplay of probabilistic and graphical models. In simple terms, any model that outputs probabilities is referred to as a probabilistic model.

A key difference between a probabilistic and a non-probabilistic model is that in probabilistic models, the final decision is taken by the user of the model which is not the case with non-probabilistic models. In a classification problem over three classes, a non-probabilistic or deterministic model like SVM would give the decision of the particular class to which the data point would belong to while a probabilistic model like logistic regression would give the probability of the data point belonging to the different classes. It is important to note here that the confidence of the model in predicting the class of the data point is very different from the correctness of the model. The correctness of the model only measures the prediction error made by the model.

Let's understand the difference between generative and discriminative class of models. Deterministic models primarily belong to the discriminative class of models while probabilistic models can be either generative or discriminative.

Models can be classified into two broad categories:

- Discriminative models: These models are trained to answer a specific question, such as predicting the label given some features.
- Generative models: These models are trained to model the underlying source from where the data is generated (this point will be more clear going forward).

An interesting point to note here is that while training a generative model, there is no explicit need to define a 'target variable'. This is because the model is trained to learn the dependencies between all variables (i.e. it learns the underlying source of data), and thus, it will anyway learn the dependencies between all the variables. Predicting the target label is then simply a matter of asking the specific question like - Given the values of the variables are x1,x2,......xn (the 'features' in the discriminative model lingo), what is the probability that y (the 'target variable') is red/green/blue for a sample problem.

To show the difference between a discriminative and a generative model, Prof. has considered a simple example in which the discriminative model classifies data points (x,y) as red(R) or green(G). Hence, for a discriminative model, the point coordinates x and y are the features while the colour (R/G) is the target variable. On the contrary, for a generative model, the three variables x, y and the colour act as features which are used to learn the **full joint probability distribution**: p(x,y,c).

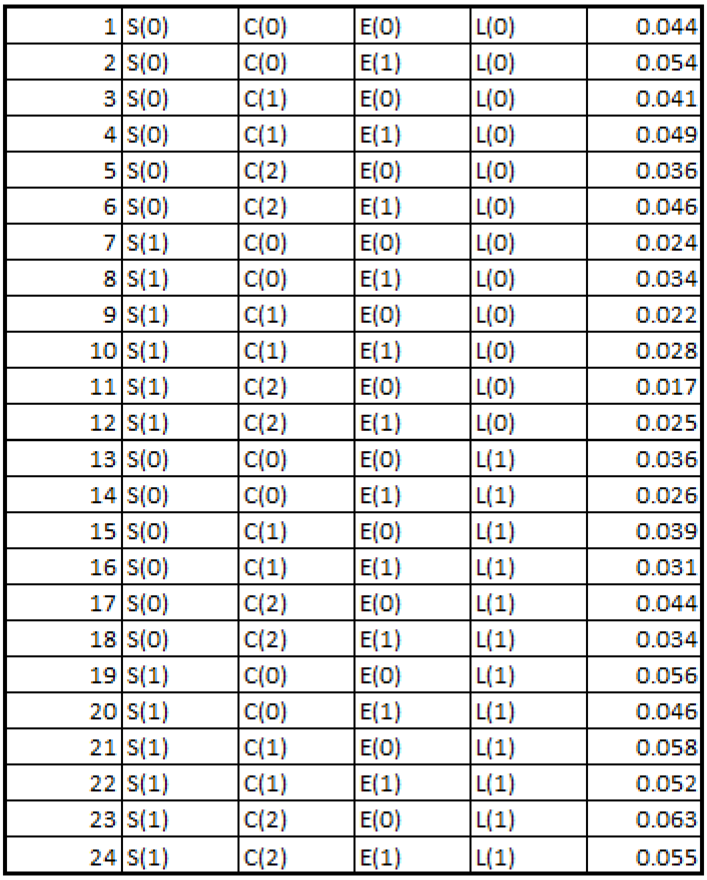
Generative models are trained such that they are able to answer any question about the data. Learning such a model is really difficult as it involves a lot of parameters. For this reason, the number of generative models is very less as compared to discriminative.

You also learnt that a discriminative model needs the values of all the features to make a prediction. Missing values would cause an error in prediction while this is not the case of the generative model. In short, generative models can do with partial data.

Another interesting feature of generative models is that it is able to predict well for outliers also, which has applications in the health sector as some of the numerous patients often belong to extreme ends of the measurable parameters and the prediction for such patients need to be correct.

Generative models are unsupervised learning algorithms, but they are very different from the Machine Learning unsupervised models. The ML unsupervised algorithms are discriminative since they are meant to answer some specific questions like outlier detection while the Generative class of unsupervised models is not meant to answer specific questions, rather model the underlying source of data. These can answer any question that you throw at it.

Prof. Raghavan also mentioned that generative models can perform prediction even when some of the feature values are not available. For a given joint probability distribution,



1	S(0)	C(0)	E(0)	L(0)	0.044
2	S(0)	C(0)	E(1)	L(0)	0.054
3	S(0)	C(1)	E(0)	L(0)	0.041
4	S(0)	C(1)	E(1)	L(0)	0.049
5	S(0)	C(2)	E(0)	L(0)	0.036
6	S(0)	C(2)	E(1)	L(0)	0.046
7	S(1)	C(0)	E(0)	L(0)	0.024
8	S(1)	C(0)	E(1)	L(0)	0.034
9	S(1)	C(1)	E(0)	L(0)	0.022
10	S(1)	C(1)	E(1)	L(0)	0.028
11	S(1)	C(2)	E(0)	L(0)	0.017
12	S(1)	C(2)	E(1)	L(0)	0.025
13	S(0)	C(0)	E(0)	L(1)	0.036
14	S(0)	C(0)	E(1)	L(1)	0.026
15	S(0)	C(1)	E(0)	L(1)	0.039
16	S(0)	C(1)	E(1)	L(1)	0.031
17	S(0)	C(2)	E(0)	L(1)	0.044
18	S(0)	C(2)	E(1)	L(1)	0.034
19	S(1)	C(0)	E(0)	L(1)	0.056
20	S(1)	C(0)	E(1)	L(1)	0.046
21	S(1)	C(1)	E(0)	L(1)	0.058
22	S(1)	C(1)	E(1)	L(1)	0.052
23	S(1)	C(2)	E(0)	L(1)	0.063
24	S(1)	C(2)	E(1)	L(1)	0.055

In order to calculate P(L(1)), you will sum up rows 13 - 24 which according to the law of total probability is essentially,

$$\sum_E \sum_C \sum_S p(S, C, E, L(1))$$

In other words, you are marginalizing the joint probability distribution over S, C and E. You can see, you do not need values of S, C and E to make an inference.

You also saw that marginalization is a computationally expensive process to make an inference/prediction. You have seen the number of summations one needs to perform to make an inference over partial data. This problem shall be solved with the help of graphs.

You have seen that if we have a joint probability distribution over 10 variables each of which have 5 levels each, the joint probability distribution table will have 5^{10} rows which is a huge number and performing marginalisation on such a huge table is inefficient. Hence, the key to overcoming the computation problem of joint probability distribution lies in understanding the independencies between the variables. By looking at the independencies, you can break down the joint probability distribution into a multiplication of smaller independent tables which makes the computations tractable.

Graphs help in figuring out the independencies between the variables. A graph and the probability distribution should be equivalent to each other in a Probabilistic Graphical Model. You also got a brief overview of the components of a graph:

- Node/ Vertex: Each variable in the probability distribution is a node in the graph
- Edge: An edge is a line connecting two nodes which signifies dependence between the two nodes representing the variables

There is an important aspect to note here that the graphs used in Probabilistic Graphical Models are meant to capture pairwise interactions. The interaction between two nodes is often referred to as one node/variable influencing the other.

Now, a graph is Markov equivalent to a joint probability distribution if all the conditional independencies present in the joint probability distribution is captured by the graph. This is known as Markov equivalence.

You already know that the two variables x1 and x2 are independent if their probability distributions satisfy the following equation

$$p(x1, x2) = p(x1).p(x2)$$

Now, let's similarly define **conditional independence**. If two variables x1 and x2 are conditionally independent on x3, then:

$$p(x1, x2|x3) = p(x1|x3).p(x2|x3)$$

In other words, the probability distribution of x1 and x2 conditioned on x3 is a multiplication of the probability distribution of x1 conditioned on x3 and the probability distribution of x2 conditioned on x3.

Note that you need to verify the conditional independence for both the values of x3 (0 and 1, if x3 is a binary variable) to state that x1 and x2 are independent conditioned on x3.

You have seen examples of daily life conditional independence and dependencies in the lectures. There are some key points to note here:

1. Dependence between variables can be figured out using the correlation values
2. Correlation between two columns having all the values as 1 is 0.
3. Covariance is defined as E[(x1−μ1)(x2−μ2)] while correlation is defined as $\frac{E[(x_1-\mu_1)(x_2-\mu_2)]}{\sqrt{E[(x_1-\mu_1)^2]E[(x_2-\mu_2)^2]}}$
4. In a joint probability distribution, to check independence between 2 variables, you can calculate the correlation or check if they satisfy the independence relationship.