Lecture Notes

Probabilistic Graphical Models

This lecture notes covers the 3rd module of the course Probabilistic Graphical Models.

Undirected Graphical Models

Until now, you saw how to figure out independencies in directed graphical models. You have seen causal relationships and the case of a collision which is fundamentally different from the rest.

The only difference in the graph structure in directed and undirected graphical models is that the edges don't have any direction in undirected graphical models.

The notion of d-separation in undirected graphical models is really simple. $X \perp Y | Z$ states that the subset of nodes X and Y are independent conditioned on Z when upon removal of Z there exists no connection between the subsets of nodes.

You also saw that the Hammersley-Clifford theorem states that a graph and a joint probability distribution are Markov Equivalent if and only if

$$p(x1,x2,...,xn)=1/Z\prod Nci=1\psi(xij:xik\in ci)$$

where,
Nc is the number of cliques.
xij:xik are the nodes that belong to the clique ci
Z is the normalising constant
ψ is the potential attached to the clique ci

A **clique** is a part of the graph in which all the nodes or vertices are fully connected to each other. A **maximal clique** is the one which has the highest number of nodes that satisfy the definition of a clique.

You know that while factorising a joint probability, you need to look for maximal cliques. The product of the potential functions of these maximal cliques will give you the joint probability (after normalising using Z). Now, while looking for maximal cliques, you should make sure that you include every node of the graph and don't leave any of the nodes.

Professor Raghavan has stated that for the Hammersley-Clifford theorem to hold, no combination of the nodes should have the probability value zero. It is equivalent to remove the nodes causing the probability value to become zero and then applying the Hammersley-Clifford theorem.

You have seen that the expression of the joint probability distribution can be written by taking the log on both sides as follows:

$$\log(p(x1,x2,...,xn))=\sum Nci=1\log(\psi(xij:xik\in ci)-\log(Z)$$

It is visible that the log of the normalizing factor, Z, is an inconvenient thing to calculate as well as optimise. We're not going to go in the calculation of log(Z) because calculation of Z is out of scope of this module.

Now, the log values can become messy for this part: $\log(\psi(x_{ij}:x_{ik} \in c_i)$. Let's see how this problem is solved by introducing the energy function in the following lecture.

Note: In the lectures, the professor has used log() and ln() both. These are used interchangeably and are trivial because we're not getting into the precise calculations.

The transformation that we apply to tackle the problem of log values is as follows:

$$\psi(c_i) = e^{-E(c_i)}$$

The potential function is written as the negative exponential of the energy function. Upon applying this transformation, you can see that the sum of logs of potential functions become the sum of the energy functions with a negative sign.

$$\log(p(x_1, x_2, \ldots, x_n)) = -\sum_{i=1}^{N_c} E(C_i) - \log(Z)$$

Even though we have simplified this part of the expression, the log of the normalising factor remains a problem which we'll have to deal with anyway in undirected graphical models.

The energy function is used quite often in thermodynamics and professor Raghavan has given an analogy of the interaction within the individuals of a colony and inter-colony interactions as interaction within a maximal clique and inter-clique interactions.

Please note that an inter-clique interaction is also maximal clique. Hence, there is an energy function to quantify the interactions between the maximal cliques.

An energy function for any two molecules measures the discord between the pair of molecules in a thermodynamic setting. If both have the same state, the energy function has low values and if they are in different states, the energy function will have a high value.

Such an analogy is used in the case of image segmentation where pixels act as molecules. So, if we want to segment the image into foreground and background, if a pixel and all its neighbouring pixels belong to the same class, that is, foreground or background, then the energy between this pixel and its neighbours will be low.

This also means that if the neighbouring pixels belong to a particular class, it is highly likely that the pixel also belongs to the same class. And we use this notion to segment a given image in the next segment.

In an image segmentation problem, you start off by **assigning labels** to the pixels. You can see that the maximal clique here is an edge between any two pixels. Hence, according to the Hammersley-Clifford theorem, the joint probability distribution is represented as a multiplication of potentials associated with each pair of neighbouring pixels. Consequentially, the log of the probability distribution is the negative of the sum of the energy function associated with each pair of neighbouring pixels along with the log of the normalizing factor as shown below.

$$\log(p(x1,x2,...,xn)) = -\sum_{i=1}^{Nc} E(ci) - \log(Z)$$

It is visible from the expression that the probability for certain assignment of labels will be high when the energy function for that assignment is low. In other words, you are looking for the maximum likelihood solution.

You have seen that image segmentation can be done in different ways one of which is demonstrated in the lectures using CRFs. CRFs are a modification of MRFs. Let's revise MRFs and their properties:

Markov random fields are a type of undirected graphical models which are a compact and efficient probabilistic representation and are generally used to model spatial data like pixels in an image. The key properties of a **Markov Random Field** are

1. Global Markov Property
2. Local Markov Property
3. Pairwise Markov Property

Let's discuss all the properties of Markov Random Fields through an informative story.

You all know about Vladimir Putin. Putin has revolutionized politics in Russia for the last twenty-five years and is a cult figure in world politics. Let's look at the social graph of Russia under Putin's leadership.

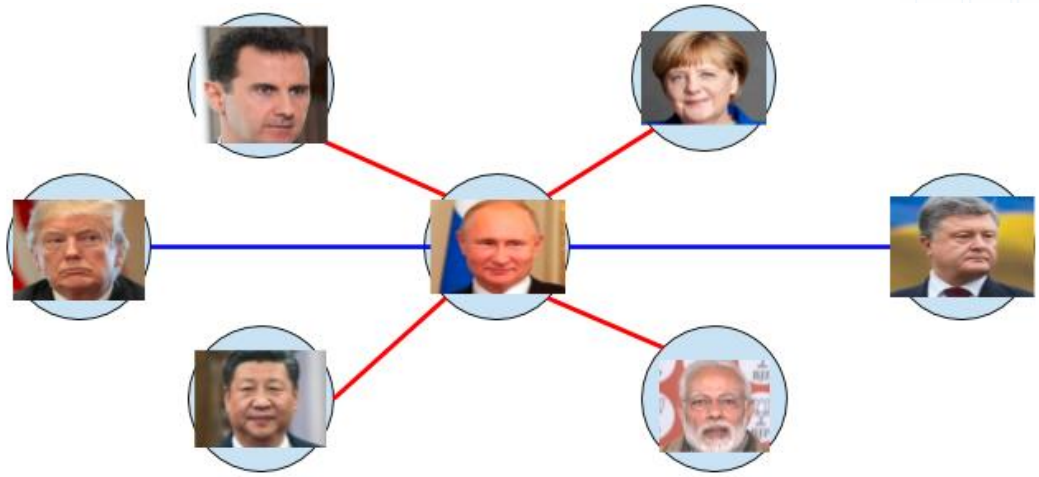
Friends

1. **India**: Putin's Russia considers India as its top ally. Who could forget Russia's help in the Indo-Pak war of 1971, when Russia sent his troops to prevent the combined attack of UK and US. India won the war and was able to free Bangladesh because of Russia's help.
2. **China**: Russia considers China as its close ally and fortunately for Russia, China reciprocates the friendship. (Trump, are you listening???). Chinese President Xi Jinping offered Putin, the first friendship medal during his visit to China.
3. **Germany**: Germany being the world's industrious leader is dependant on Russia for 60 percent of its oil imports. This fact was brought to prominence by Trump as he criticised Germany for its dependency on Russia. Trump is mad with the progress of Germany and Russia's gas pipeline Nord Stream 2.
4. **Syria**: Russia's support for President Assad of Syria is well known and the two countries enjoy a historically strong, stable, and friendly relationship.

Foes

1. **USA**: Trump has openly expressed his intentions to strengthen the relationship with Russia. But it remains to be seen if Russia will reciprocate.
2. **Ukraine**: Russia's relationship with Ukraine deteriorated after the Crimean crisis where Ukraine accused Russia of meddling with Ukraine's internal affairs.

This can be pictorially described using the following graph.



Now, let's understand the properties using the above graph.

1. **Global Markov Property**:
   Global Markov Property states that for sets of nodes A, B, and C, A ⊥ B| C iff C separates A from B in the graph G. In other words, when we remove all the nodes in C, if there are no paths connecting any node in A to any node in B, then the conditional independence property holds.
   In Putin's social graph, based on the Global Markov Property, we can say that
   India ⊥ Syria | Russia
   (India is independent of Syria given Russia).
2. **Local Markov Property**:
   The smallest set of nodes that renders a node conditionally independent of all the other nodes in the graph is called its Markov blanket. In an undirected graphical model, a node's **Markov blanket** is its set of immediate neighbours. This is called the undirected local Markov property.
   In Putin's social graph, based on the Local Markov Property, we can say that
   Russia ⊥ Rest of the world | {India, China, Germany, Syria, Ukraine, USA}
   (Russia is independent of the rest of the world given its social graph neighbours India, Germany, Syria, Ukraine, USA and China)
3. **Pairwise Markov Property**:
   According to the local Markov property, two nodes are conditionally independent given the rest of the nodes if there is no direct edge between them. This is called the pairwise Markov property.
   In Putin's social graph, based on the pairwise Markov Property, we can say that
   India ⊥ Syria | {Russia, China, Germany, Ukraine, USA}
   (India is independent of Syria given the rest of social graph members Russia, Germany, Ukraine, USA and China)

You saw that a **conditional random field** is a Markov random field with an additional condition. In the case of image segmentation in the lectures, the **label is conditioned on the values of the pixels**, specifically, on the **neighbouring pixels** because they are the ones that comprise of the Markov blanket for the central pixel. Also notice that we're trying to segment a grayscale image, hence, we have only one value corresponding to each pixel rather than 3 values that we would have had for an RGB image. This is one of the methods of performing image segmentation which you revise from the lectures.

Let's end with some key points that shall help in deciding the graph to be used for a particular use case. There are causal relationships and collision structures in directed graphs which cannot be modelled by using undirected graphs. Similarly, some conditional independencies present in undirected graphs cannot be modelled by directed graphs. One of the key things to keep in mind is the relationship between the

variables. If the causal relationship between the variables is symmetric as in the case of pixels in images, you should use an undirected graph. On the other hand, if you see that a certain variable is the cause of another variable, like in case of medical diagnosis, you use a directed graph.

----------THE END----------