ML Probabilistic Modeling 1 by ambedkar@IISc

- Introduction
- Markov Random Fields

On Probabilistic Modeling

On Probabilistic Modeling : Why?

- Let us go back to our fundamental assumption

  The observed data is assumed to be sampled from a (unknown) underlying probability distribution

- Previously we have not used this for modeling but we have used when we defined the risk

On Probabilistic Modeling (Contd. . .)

- Recall the definition of risk in the context of supervised learning :
  Given data $(x_n, y_n)_{n=1}^{N}$ find $f : \mathcal{X} \to \mathcal{Y}$ that best approximates the relation between random variables X and Y.
  Thus the risk is defined as

$$
\begin{aligned}
L(f) &= \mathbb{E}_{(x,y)\sim P}[l(Y, f(X)] \\
&= \int l(y, f(x))d(P(x, y))
\end{aligned}
$$

- What if we just learn $P(X, Y)$ instead of $y = f(x)$?

On Probabilistic Modeling (Contd...)

Advantages

- Machine learning problems intrinsically involve "Uncertainty". Probabilistic models automatically include that is the predictions
- Probabilistic models are generative in nature. Hence one can generate samples.

Disadvantages

- Inference and learning of probabilistic modeling is notoriously difficult.

Solution

- MCMC
- Variational methods

Probabilistic Modeling : Example

- Consider the problems of text classification. Aim is to classify an email is spam or not.
- $X = (X_1, \ldots, X_i, \ldots, X_D)$: is a one hot vector i.e. $X_i \in 0, 1$ denotes whether an email is spam or not.
- $Y \in 0, 1$: whether an email is spam or not.
- $P_\theta(Y, X_1, \ldots, X_D)$: Probabilistic modeling of $Y, X_1, \ldots, X_D$.

Probabilistic Modeling : Example (Cont. . .)

$P_\theta(y = 1|x_1, \ldots, x_D)$: Given an email $x = (x_1, \ldots, x_D)$ this is the probability that x is spam

$P_\theta(y = 0|x_1, \ldots, x_D)$: Given x, probability that x is not spam

Question : What is the size of the set on which P_θ in defined?
Ans : 2^{D+1}

- Considering D is the size of the vocabulary of English, this is a huge set
- Also estimating the parameter θ is very difficulty.

Probabilistic Modeling : Challenges

- We can say machine learning problems involves probabilistic modeling on
  "Sets that are exponentially big"

- Hence knowing the "dependencies" among the random variables is important

Naive Bayes Assumption

Assumption : Given Y, all $X_1, \ldots, X_D$ are independent i.e.

$$P(Y, X_1, \ldots, X_D) = P(Y) \prod_{d=1}^{D} P(X_d|Y)$$

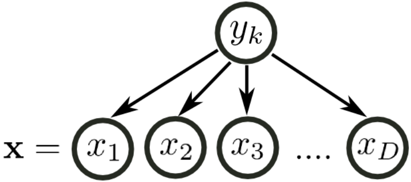
In the text classification problems

- Each $P(X_d|Y)$ can be described by 4 parameters
- The entire distribution is parameterized by O(n) parameter

We have brought this down from exponential sized set to fixed sets.

Naive Bayes Assumption : Graphical Representation

- The graphical representation of data generation
  Story : That is the graph describes how the data is generated



- An email was generated by first choosing email is spam or not (y_k) and thus based on this words for email are sampled.

3 Elements of Probabilistic Modeling

- Reputation : Deals with model selection and assumptions
- Inference : given a Model how to get answers for various questions
- Learning : Parameter estimation

Reputation

- How do we represent underlying probability models
- What kind of independence : Directed Vs Undirected
- Latent variable models?

Inference

Given a model how can we obtain answer to relevant questions?

- Marginal inference : What is the probability distribution of X_1. For example, what is the probability that email is spam if certain word present in the email?

$$P(X_1) = \sum_{X_2} \sum_{X_3} \cdots \sum_{X_D} P(X_1, \ldots, X_D)$$

- MAP Inference : MAP inference answer questions about mostly likely assignment. For example, what is the mostly likely spam message

$$\arg\max_{X_1, \ldots, X_D} P(X_1, \ldots, X_D, Y = 1)$$

Learning

- Given data how to estimate the parameters?
- Learning and inference are tightly connected
- Usually inference algorithm will become part of learning

Bayesian Nets

We will not study this in detail. We give only some definitions

- Given any distribution $P(X_1, \ldots, X_D)$ we can write this using chain rule

$$P(X_1, \ldots, X_D) = P(X_1)P(X_2|X_1)P(X_3|X_2, X_1)\ldots$$

$$\ldots(X_n|X_{n-1}, \ldots, X_1)$$

- Bayesian network represent distributions where right hand side depends only on small number of ancestor variables X_A:

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

Bayesian Networks(Contd. . .)

- Example : Suppose we have $P(X_7|X_6, \ldots, X_2, X_1)$. Say we can approximate this distribution by $P(X_7|X_4, X_2)$. Than $A_{X_7} = X_4, X_2$

Definition : Bayesian network is a directed graph G=(V,E) together with

- A random variable X_i for each node V_i
- One conditional probability distribution $P(X_i|A_{X_i})$ per node

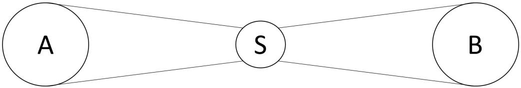
We say probability model $P(X_1, \ldots, X_D)$ over a DAG G if it can be decomposed into product of factor specified by G.

Markov Random Fields

Markov Random Fields generalizes Markovian Property

$$X_n \leftarrow X_{n-1} \leftarrow X_{n-2} \leftarrow \ldots \leftarrow X_D$$

Given X_{n-1} there is no effect of $X_{n-2}, \ldots, X_D$ on X_n i.e.
$P(X_n | X_{n-1}, \ldots, X_D) = P(X_n | X_{n-1})$

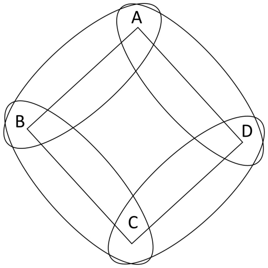


Suppose every node in A has a path to B via S. Thus given S, A and B are independent.

MRF : Example

Example

- Let us study voting preferences of four individuals A,B,C,D. Suppose friendships are (A,B),(B,C),(C,D),(D,A). hence assumption is that they have similar voting patterns.



MRF : Example (contd...)

Aim : Probabilistic modeling of joint voting decision of A,B,C,D.

Define $\tilde{P}(A,B,C,D) = \phi(A,B)\phi(B,C)\phi(C,D)\phi(D,A)$

where $\phi(X,Y)$ is a factor that gives more weight to consistent voter among friends e.g.

$$\phi(X,Y) = \begin{cases} 10 & \text{if X=1, Y=1} \\ 5 & \text{if X=0, Y=0} \\ 1 & \text{otherwise} \end{cases}$$

Now we define probability distribution

$$P(A,B,C,D) = \frac{1}{Z}\tilde{P}(A,B,C,D)$$

$$Where Z = \sum_{A,B,C,D} \tilde{P}(A,B,C,D)$$

- Note that the way we have defined potential function i.e. ϕ, results in the following : P(A,B,C,D) has more value when A and B votes consistently. Similarly same for (B,C),(C,D) and (D,A)

MRF : Some Remarks

- We cannot say anything about how one variable effect othe. We can only say somthing about how two r.v.s are coupled.
- In the case of undirected models we need less knowledge and we do not need generaize story of how one variable is generated.
- We only identify dependent variables and define the strength of their interactions.
  - Energy depending on different configuration
  - thus connect this energy to a probability via normalizing constant.

MRF : Definition

<u>Def</u> : Let G=(V,E) be a graph. Consider the collection of r.v.s $X_{vv \in V}$ and each r.v X_v taken values from the same set $\mathcal{X}$.$X_{vv \in V}$ is said to be Markov Random Field if the joint probability distribution satisfies Markovian property w.r.t G.

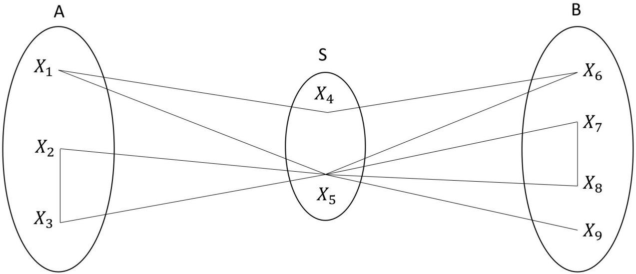
What is Markovian Property w.r.t Graph G?

- $\boxed{\text{A set } A \subset V \text{ separates two vertices } v \in A \text{ and } w \in A\text{, if every path from v to w contains a node from A.}}$

- For any disjoint sets A, B, $S \subset V$ if all vertices in A and B are separated by S thus $\{X_a\}_{a \in A}$ and $\{X_b\}_{b \in B}$ are independent given $\{X_s\}_{s \in S}$

$$i.e.\ P(X_a : a \in A | X_t : t \in S \cup B) = P(X_a : a \in A | X_t : t \in S)$$

Example :



$$P(X_1, X_2, X_3 | X_4, X_5, X_6, X_7, X_8, X_9) = P(X_1, X_2, X_3 | X_4, X_5)$$

Markov Blanket

A set of vertices MB(v) is called Markov blanket of $v \in V$ if for any $B \subset V$ such that $v \notin B$ we have

$$P(X_v | X_t : t \in MB(v), B) = P(X_v | X_t : t \in MB(v))$$

i.e. X_v is conditionally independent from any other r.v.s given $X_t : t \in MB(v)$

In MRF MB(v)=Neighbour(v)

Hammersley–Clifford Theorem

A strictly positives probability distribution P satisfies Makovian property w.r.t our undirected graph G	$\iff$	P factorises our G

Proof can be found in Koller and Friedman : Probabilistic Graphical Models 2009

Factorization of a P.D over a graph

A distribution P is said to factorize about our undirected graph G with set g maximal cliques τ is there exists a set of non-negative functions

$\{\Psi_c\}_{c\in\tau}, \quad \Psi_c : \mathcal{X}^D \to \mathbb{R}$ satisfying

$x, y \in \mathcal{X}^D$ thus $(x_i)_{i\in c} = (y_j)_{j\in c}$

$$\implies \Psi_c(x) = \Psi_c(y)$$

$$and \;\; P(x) = \frac{1}{Z}\prod_{c\in\tau}\Psi_c(x_c)$$

$$Where \;\; Z = \sum_{x\in\mathcal{X}}\prod_{c\in\tau}\Psi_c(x_c)$$

Factorization of a P.D. over a Graph

If P is strictly positive same holds for potential function. Thus

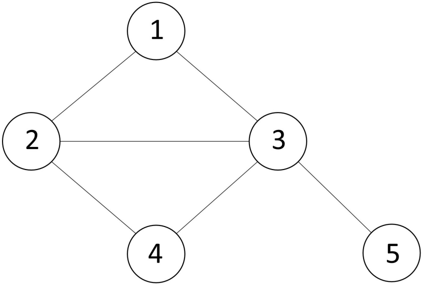
$$\begin{aligned} P(x) &= \frac{1}{z}\prod_{c\in\tau}\Psi_c(x_c) \\ &= \frac{1}{z}\exp\big(\sum_{c\in\tau}\ln\Psi(x_c)\big) \\ P(x) &= \frac{1}{Z}e^{-E(x)} \end{aligned}$$

$$Where \;\; E(x) = -\sum_{c\in\tau}\ln\Psi_c(x_c)$$

- Here P(x) represent Gibbs distribution and E(x) is called energy function.
- Probability distribution of any model can be represented in the form of Gibbs distribution.

Factorization of a P.D. over a Graph

If the distribution P is Markovian w.r.t this graph thus it can be represented as



$$P(X) = \frac{1}{Z}\Psi_{123}(x_1, x_2, x_3)\Psi_{234}(x_2, x_3, x_4)\Psi_{35}(x_3, x_5)$$

representation of Potential Functions

Note that though we ignored parameters θ, undirected graphical models involve parameter. We can write this as

$$\begin{aligned} P(X|\theta) &= \frac{1}{Z(\theta)} \prod_{c \in \tau} \Psi_c(y_c|0_c) \\ &= \frac{1}{Z(\theta)} \exp(\sum_{c \in \tau} \ln \Psi_c(x_c)) \end{aligned}$$

Define $\log$ potential as linear function of the parameter

$$\log \Psi_c(x_c) = \Phi_c(x_c)^T \theta_c,$$

Where $\Phi_c(x_c)$ is a feature vector defined from the values of the variable x_c

Representation of Potential Functions (Contd. . .)

The resultant model is

$$
\begin{aligned}
P(x|\theta) &= \frac{1}{Z(\theta)} \exp(\sum_c \Phi_c(x_c^T \theta_c) \\
or & \\
\log P(x|\theta) &= \sum_c \Phi_c(x_c)^T \theta_c - \log Z(\theta)
\end{aligned}
$$

These are called maximum entropy or log-linear models

Examples of MRFs

1. Ising Models : Modeling behaviour of marginals, which is a 3d or 2d lather with $x_0 \in [-1, +1]$

2. Hopfield Network : Fully connected ising model

3. Boltzmann Machine : This is a generalization of Hopfield network, where oenwould introduce hidden nodes that makes the model more powerfull.

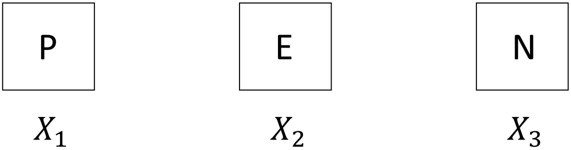
   We will study something called Restricted Boltzmann Machines

Remarks on MRFs

- Computing Z requires summation over exponential numbers of configurations. This is intractable and requires approximate approaches
- Undirected graphical models can be difficult to interpret as causality in last
- It is easier to generate data from directed models rather than undirected models

Conditional Random Fields

- This is a special case of MRFs and is applicable in supervised learning
- Aim is to model distribution of type P(Y|X), where X and Y are vector values
- Application
  Given a sequence of images of English alphabet recognize the letters or a word



Conditional Random Fields

Definition : CRF is a MRF over variables

$X = (X_1, \dots, X_N)$
$Y = (Y_1, \dots, Y_M)$

defines conditional distribution

$$P(y|x) = \frac{1}{Z(x)} \prod_{c \in \tau} \Phi_c(x_c, y_c)$$

$$Where \;\; Z(x) = \sum_{y \in \mathcal{Y}} \prod_{c \in \tau} \Phi_c(x_c, y_c)$$