Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

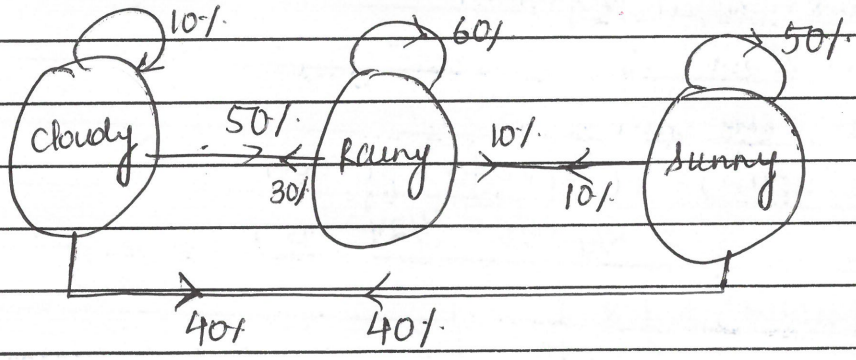
Subject: Natural Language Processing

Numericals for reference on markov chains:-

A markov chain consist of three important components:-

- Initial probability distribution: An initial probability distribution over the states, πi is the probability that the markov chain will start in a certain state i. Some state j may have $\pi j = 0$ meaning they cannot be initial states.
- One or more states
- Transition probability distribution: A transition probability matrix A where each transition represents the probability of moving from state i to state j.

The diagram below represents a markov chain where there are three states representing the weather of the day.



There are three different states such as:-

Cloudy, rain and sunny, the following represents the transition probabilities ie

If it is sunny today, then tomorrow

- 50% probability for sunny
- 10% probability for rainy
- 40% probability for cloudy

and so on using this markov chain, what is the probability that wednesday will be cloudy if today is sunny (Today is Monday)?

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

→ Sunny–Sunny (Tuesday) – cloudy (wednesday) : The probability to a cloudy wednesday can be calculated as $0.5 \times 0.4 = 0.2$.

→ Sunny – Rainy (Tuesday) – cloudy (wednesday) : The probability to a cloudy wednesday can be $0.1 \times 0.3 = 0.03$.

→ Sunny – cloudy (Tuesday) – ~~Rain~~ cloudy (wednesday) $= 0.4 \times 0.1 = 0.04$

The total probability ⇒ $0.2 + 0.03 + 0.04$
of a cloudy wednesday $= 0.27$.

Conditional Random Field :– [CRF]

→ Conditional random fields (CRF's) are a probabilistic framework for labelling and segmenting sequential data based on the conditional approach described in the HMM.

→ A CRF is a form of undirected graphical model that defines a single log linear distribution over label sequences given a particular observation sequence.

→ The primary advantage of CRF's over hidden markov model is their conditional nature, resulting in the relaxation of the independence assumptions, required by HMM in order to ensure tractable inference.

→ Additionally CRF's avoid to label bias problem, a weakness exhibited by maximum entropy markov model (MEMM). and other conditional markov models based on directed graphical models.

→ The graphical structure of a conditional random field may be used to factorize the joint distribution over elements

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

CRF

- Conditional random fields (CRFs) are a class of statistical modelling methods. They are applied in pattern recognition and they are used for structured prediction.
- A classifier predicts a label for a single sample without considering neighbouring samples, a CRF can take context into account.
- To achieve this predictions are modelled as a graphical model. And they represent the presence of dependencies between the predictions.
- Depending on the application the type of graph is used. In natural language processing "linear chain" CRF's are popular, because each prediction is dependent only on the immediate neighbours.
- In image processing the graph connects locations to nearby similar locations to enforce that they receive similar predictions.
- other examples where CRF's are used: labelling or parsing
  1) for sequential data of natural language processing
  2) for biological sequences.
  3) for POS tagging.

Describe undirected probabilistic graphical method:—

- CRF's are a type of discriminative undirected probabilistic graphical model.
- We define CRF on observations X and random variables Y as follows:

Let $G = (V, E)$ be a graph such that

$Y = (Y_v)_{v \in V}$ so that Y is indexed by a vertices of G.

Then (X, Y) is a conditional random field where each random variable Y_v conditioned on X obeys a markov property with respect to the graph; that is, its probability is dependent only on its neighbours in G: $P(Y_v | X, \{Y_w : w \neq v\}) = P(Y_v | X, \{Y_w : w \sim v\})$

where $w \sim v$ mean that w and v are neighbours in G.

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

Inference for CRF:

For general graphs, the problem of exact inference in CRF's is not possible. But there exists special cases for which exact inference is feasible.

1) If the graph is a chain or a tree, message passing algorithms yield exact solutions. The algorithms that are used in cases are similar to the forward, backward and Viterbi algorithm for the case of HMM's.

2) If the CRF only contain pair-wise potentials and the energy is submodular.

→ CRF's are trained using maximum likelihood estimation which involves optimizing the parametres of the model in order to maximize the probability of the correct output sequence given the input features.

→ The formula for CRF is similar to that of markov random field (MRF) but with addition of input features that condition the probability distribution over output sequences. Let X be the input features and Y be the output sequence. The joint probability distribution of CRF is given by:—

$$P(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x_i)\right)$$

where: 1) $Z(X)$ is the normalization factor that ensures the distribution sums to 1 over all the possible output sequences.

2) λ_k are the learned model features.

3) $f_k(y_{i-1}, y_i, x_i)$ are the feature functions that take as input the current output state y_i, the previous output state y_{i-1} and input features x_i.

4) These functions are binary or real valued and capture dependencies.

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject:Natural Language Processing

CRF VS HMM:

→ CRF can define much larger set of features.
→ CRF uses more global features.
→ CRF can have arbitary weights.

MEMM: [Maximum Entropy markov model]

Maximum entropy modelling is a framework for integrating information from many heterogeneous information sources for classificaton. Maximum entropy refers to the optimization framework in which the goal is to find a probability model that maximizes entropy over set of model.

Maximum entropy classifier is a probabilistic classifier which belongs to the class of exponential models.

Theoretical background of maximum entropy:

our target is to use contextual information of the document (unigrams bigrams, trigrams etc) in order to categorize it in a particular class. each document is represented using standard bag of words framework and each document is represented using a sparse array with 0's and 1's that indicates whether a particular word exists or not.

Target is to construct a stochastic model which accurately represents the behaviour of the random process.

Steps of MEMM:

1] The first step is to collect large number of training data (xi, yi)
where [xi → contextual information [sparse array]
yi → class.

Parshvanath Charitable Trust's
A. P. SHAH INSTITUTE OF TECHNOLOGY
(Approved by AICTE New Delhi & Govt. of Maharashtra, Affiliated to University of Mumbai)
(Religious Jain Minority)

Subject: Natural Language Processing

2) Second step is to summarize the training sample in form of empirical probability distribution.

$$p(x,y) = \frac{1}{N} \times \text{no of times tag appears in a sample.}$$

N → training set.

3) Model: Suppose we have the sequence of observations $O_1, O_2 \ldots O_n$ that we seek to tag with labels $S_1, S_2 \ldots S_n$ that maximize the conditional probability $P(S_1 \ldots S_n | O_1 \ldots O_n)$

In MEMM, this probability is factored into markov transition probabilities. The probability of transitioning to a particular label depends on the observation at that position and the previous position label.

$$P(S_1 \ldots S_n | O_1 \ldots O_n) = \prod_{t=1}^{n} P(S_t | S_{t-1}, O_t)$$

Advantages of MEMM:—

1) MEMM is a discriminative model it uses conditional probability conditioned on the previous tag and current word.
2) In MEMM a distribution is made by adding features which can be picked out by training.
3) The idea is to select the maximum entropy distribution given the constraints specified by features.
4) MEMM is more flexible as more features can be added.

→ University Questions of Module 3.

Q1. Explain HMM (Hidden markov model)? [MU, Dec 2023] (ADS)
Q2. Explain different approaches of POS Tagging? [MU, May 2022] (ADS)
Q3 Explain difference between open word class and closed word class? [MU, May 2023].

— X — X — X — X — X —

Short note on Penn Tree Bank (PTB):—

The Penn Tree Bank is widely used resource in computational linguistics and natural language processing (NLP).
It is developed by the university of Pennsylvania, it consist of a large corpus of text that has been annotated with syntatical and semantic information.
The key points about Penn Tree Bank are as follows:—

1. Corpus composition: The PTB includes a variety of texts, such as articles from Wallstreet, IBM computer manuals, transcribed telephone conversations.
2. Annotations: ① The PTB has been instrumental in the development & evaluation of NLP algorithms, particularly those involving syntatical parsing.
② It has served as a benchmark for many linguistic tasks, helping researchers to compare the performance of diffrent models and methods. The PTB has been annotated with part of speech tags which indicate grammatical categories of words.
3. Syntatical parsing: The syntatical annotations in PTB have been used to train and evaluate parsers that automatically generate parse tree from raw text.
4. Extensions and variations: ① Various extensions and adaptations of the PTB have been created for diffrent languages and for more specific linguistic tasks.