Hidden Markov Model

July 2020

This note is based on Daniel Jurafsky and James H. Martin's *Speech and Language Processing, 3rd Edition Draft*

1 Markov Chains

Markov chain is a model that tells us something about the probability of transitions from one state to another. It is a stochastic model, but future actions are not affected by the steps that led up to the present state.
Markov Property: For states x_0, ..., x_n, in which n is a positive integer,

$$P(X_n = x_n|X_{n-1} = x_{n-1}) = P(X_n = x_n|X_{n-1} = x_{n-1}, X_{n-2} = x_{n-2}, ..., X_0 = x_0)$$

2 The Hidden Markov Model

A **hidden Markov model (HMM)** allows us to talk about both observed events (like words that we see in the input) and hidden events (like part-of-speech tags) that we think of as causal factors in our probabilistic model. An HMM is specified by following contents:

$Q = q_1 q_2 ... q_N$	A set of N **states**
$A = a_{11} ... A_{ij} ... A_{NN}$	A **transition probability matrix**, each a_{ij} represents the probability of moving from state i to state j
$O = o_1 o_2 ... o_T$	A sequence of T **observations**
$B = b_i(o_t)$	A sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state
$\pi = \pi_1, ..., \pi_n$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i.

3 Viterbi Algorithm

3.1 An Example

We have a sentence *Janet will back the bill*, and its corresponding series of tags are:
Janet/NNP will/MD back/VB the/DT bill/NN

In Fig 8.7 and Fig 8.8 We have two tables defining the HMM.

Here are steps for calculating the best tag for each word, say if we do not know the above series of tags. There are 5 columns (words) in total, and We start from the first word *Janet*. Multiply the transition probability to the observation likelihood, we could get the cell values for each tag:

$$\begin{aligned} P(NNP| < s >) * P(Janet|NNP) &= 0.2767 * 0.000032 \approx 0.000009 \\ P(MD| < s >) * P(Janet|MD) &= 0.0006 * 0 = 0 \\ ... &= 0 \\ P(DT| < s >) * P(Janet|DT) &= 0.2026 * 0 = 0 \end{aligned}$$

	NNP	MD	VB	JJ	NN	RB	DT
$<s>$	0.2767	0.0006	0.0031	0.0453	0.0449	0.0510	0.2026
NNP	0.3777	0.0110	0.0009	0.0084	0.0584	0.0090	0.0025
MD	0.0008	0.0002	0.7968	0.0005	0.0008	0.1698	0.0041
VB	0.0322	0.0005	0.0050	0.0837	0.0615	0.0514	0.2231
JJ	0.0366	0.0004	0.0001	0.0733	0.4509	0.0036	0.0036
NN	0.0096	0.0176	0.0014	0.0086	0.1216	0.0177	0.0068
RB	0.0068	0.0102	0.1011	0.1012	0.0120	0.0728	0.0479
DT	0.1147	0.0021	0.0002	0.2157	0.4744	0.0102	0.0017

Figure 8.7 The A transition probabilities $P(t_i|t_{i-1})$ computed from the WSJ corpus without smoothing. Rows are labeled with the conditioning event; thus $P(VB|MD)$ is 0.7968.

	Janet	will	back	the	bill
NNP	0.000032	0	0	0.000048	0
MD	0	0.308431	0	0	0
VB	0	0.000028	0.000672	0	0.000028
JJ	0	0	0.000340	0	0
NN	0	0.000200	0.000223	0	0.002337
RB	0	0	0.010446	0	0
DT	0	0	0	0.506099	0

Figure 8.8 Observation likelihoods B computed from the WSJ corpus without smoothing, simplified slightly.

Next, for the column *cell*, we could calculate the product similarly:

$$[P(NNP|<s>)*P(Janet|NNP)]*[P(MD|NNP)*P(will|MD)] = 0.000009*(0.0110*0.308431)$$
$$[P(NNP|<s>)*P(Janet|NNP)]*[P(VB|NNP)*P(will|VB)] = 0.000009*(0.0009*0.000028)$$
$$[P(NNP|<s>)*P(Janet|NNP)]*[P(NN|NNP)*P(will|NN)] = 0.000009*(0.0584*0.0002)$$

Similarly, we could calculate the probability for each series of tags:

$$P(t_1^n) \approx \prod_{i=1}^{n} P(t_i|t_{i-1})$$

4 References

https://brilliant.org/wiki/markov-chains/