Hidden Markov Model

The i.i.d assumption allows us to express the likelihood function as the product over all datapoints of the probability distribution evaluated at each data point. But in some circumstances, for example, in sequential data, data generation is not independent, previous data can have an impact on the next data generation, which violates the i.i.d assumption. For example, sequential data with time series.

It's useful to disguish stationary and nonstationary sequential distributions. In the stationary case, only the data evolves in time, but the distribution from which the data is generated remains the same. While in the nonstationary case, the generative distribution itself is evolving with time, that is, the parameter θ that defines the probabilistic model will change along with time. Here we shall focus on the stationary case.

Given a sequential data with observations ordered in time, we expect that recent observations are more informative than more historical observations. And considering a future observation is dependent on all of the history observations is impractical, as the parameter of such a model would grow wiout limit. So **Markov Models** is defined as the future observations are independent of all but the most recent observations which leads to **M**-order Markov Chain.

The joint distribution for a sequence of observations is given by :

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(\mathbf{x}_n | \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1})$$

On the right hand of the above equation, It is the product of conditional probability, If we assume each conditional probability is independent of all previous observations except the most recent ones, we obtain the **first-order Markov Chain**, i.e., $p(\mathbf{x}_n|\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n-1}) = p(\mathbf{x}_n|\mathbf{x}_{n-1})$. So the joint distribution for a sequence of N observations can be written as :

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = p(\mathbf{x}_1) \prod_{n=2}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1})$$

In most applications of this model, the conditionale distribution $p(\mathbf{x}_n|\mathbf{x}_{n-1})$ is constrained to have the same parameters is this probability model following the assumption of stationary distribution as above. Then this **first-order Markov Chain** is known as the **homogeneous Markov Chain**. Similaly we can extend to **M-th order Markov Chain**, which is defined as follows:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{x}_{n-1}, \mathbf{x}_{n-2}, \ldots, \mathbf{x}_{n-M})$$

However, this **M-th order Markov Chain** parameter will grow exponentially with M. So to achieve the goal that we build a model that is not limited to the Markov assumption to any order and in the mean time, can be specified using a limited number of free parameters, we can introduce additional latent variables as the GMM model does, leading to **state space models**.

For each observation $\mathbf{x}_n$, there is a corresponding latent variable $\mathbf{z}_n$ (which may be different type or dimensionality to the observed variable). The joint distribution for this model is given by:

$$p(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N, \mathbf{z}_1, \mathbf{z}_2, \mathbf{z}_N) = p(\mathbf{z}_1) \prod_{n=2}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1}) \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{z}_n)$$

We can see there is always a path between two observed variables $\mathbf{x}_n$,$\mathbf{x}_m$. And our future predictions depends on all observed variables but don't satisfy any order Markov Chain.

Hidden Markov Models(HMM) is a specific instance of state space model in which the latent variables are discrete. It has a wide applications of modeling sequential data, such as speech recognition , natural language modeling, 分词、命名实体识别、词性标注