National Technical University of Ukraine
“Igor Sikorsky Kyiv Polytechnic Institute”
Institute of Physics and Technology

Lecture 12
Generative Models for Discrete Data. Markov Models

Dmytro Progonov,
PhD, Associate Professor,
Department Of Physics and Information Security Systems

Content

- Markov models;
- Transition matrix;
- Stationary distribution of a Markov chain:
  - Computing the stationary distribution;
  - Conditions of stationary distribution;
  - Detailed balance
- Hidden Markov models;
- Types of inference problems in HMMs;
- Generalizations of HMMs.

Markov models

The *basic idea* behind a ***Markov chain*** is to assume that X_t captures all relevant information for predicting the future, e.g. we assume it is a sufficient statistics. If we assume discrete time steps, we can write the joint distribution as follows

$$p(X_{1:T}) = p(X_1)p(X_2|X_1)p(X_3|X_2)\cdots = p(X_1)\prod_{t=2}^{T} p(X_t|X_{t-1})$$

If we assume the transition function $p(X_t|X_{t-1})$ is independent on time, the chain is called ***homogeneous***, ***stationary*** or ***time-invariant***.

If we assume that observed variables are discrete, so $X_t \in \{1,2,\cdots K\}$, this is called a ***discrete-state*** or ***finite-state Markov chain***.

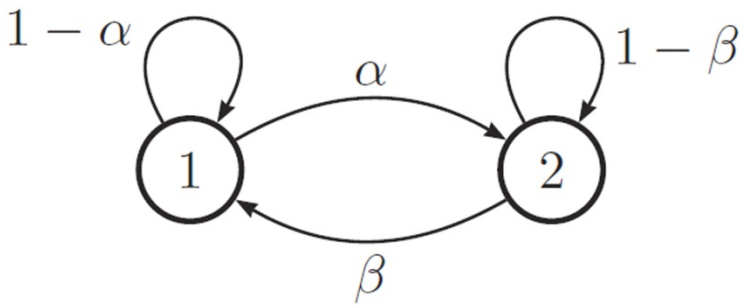
Transition matrix (1/2)

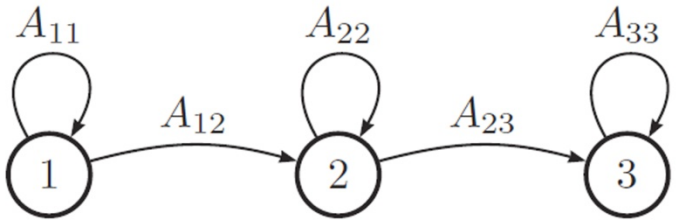
When X_t is discrete , so $X_t \in \{1,2,\cdots K\}$, the conditional distribution $p(X_t|X_{t-1})$ can be written as a $K \times K$ ***transition matrix*** $\mathbf{A}$, where $A_{ij} = p(X_t = j|X_{t-1} = i)$is the probability of going from state i to state j.

Each row of the matrix $\mathbf{A}$ sums to one, $\sum_j A_{ij} = 1$, so this is called a ***stochastic matrix***.

$$\mathbf{A} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$$

$$\mathbf{A} = \begin{pmatrix} A_{11} & A_{12} & 0 \\ 0 & A_{22} & A_{23} \\ 0 & 0 & 1 \end{pmatrix}$$





State transition diagrams for some simple Markov chains. Left: a 2-state chain. Right: a 3-state left-to-right chain.

Transition matrix (2/2)

The n −step transition matrix $\mathbf{A}(n)$ is defined as

$$A_{ij}(n) \triangleq p(X_{t+n} = i | X_t = j)$$

which is the probability of getting from i to j in exactly n steps. Obviously $\mathbf{A}(1) = \mathbf{A}$. The ***Chapman-Kolmogorov equations*** state that

$$A_{ij}(m+n) = \sum_{k=1}^{K} A_{ik}(m) A_{kj}(n)$$

We can write the above as a matrix multiplication

$$\mathbf{A}(m+n) = \mathbf{A}(m)\mathbf{A}(n)$$

$$\mathbf{A}(n) = \mathbf{A}^n$$

Markov model application: Language modelling

The marginal probabilities $p(X_t = k)$ are called ***unigram statistics***.

For first-order Markov model we obtain $p(X_t = k|X_{t-1} = j)$ that is called ***bigram statistics***.

Correspondingly for second-order Markov model $p(X_t = k|X_{t-1} = j, X_{t-2} = i)$ we obtain ***trigram statistics***.

These statistics can be used for *following applications*:

- Sentence completion;
- Data compression;
- Text classification;
- Automatic essay writing.

Stationary distribution of a Markov chain (1/2)

We can interpret Markov models as stochastic dynamical systems, where we "hop" from one state to another at each time step. In this case, we are often interested in the long term distribution over states, which is known as the ***stationary distribution*** of the chain.

Let $A_{ij} = p(X_t = j | X_{t-1} = i)$ be the one-step transition matrix, and let $\pi_t(j) = p(X_t = j)$ be the probability of being in state j at time t. It is conventional in this context to assume that $\boldsymbol{\pi}$ is a *row* vector.

If we have an initial distribution over states of $\boldsymbol{\pi}_0$, then at time 1 we have

$$\pi_1(j) = \sum_i \pi_0(i) A_{ij}$$

$$\boldsymbol{\pi}_1 = \boldsymbol{\pi}_0 \mathbf{A}$$

Stationary distribution of a Markov chain (2/2)

We can imagine iterating these equations. IF we ever reach a stage where

$$\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{A}$$

then we say we have reached the ***stationary distribution***, also called the ***invariant*** or ***equilibrium distribution***. Once we enter the stationary distribution, we will never leave.

In general, we have

$$\pi_i \sum_{j \neq i} A_{ij} = \sum_{j \neq i} \pi_j A_{ij}$$

In other words, the probability of being in state i times the net flow out of state i must equal the probability of being in each other state j times the net flow from that state into i. These are called the ***global balance equations***.

Computing the stationary distribution

To find the stationary distribution, we can just solve the eigenvector equation $\mathbf{A}^T\mathbf{v} = \mathbf{v}$, and then to set $\boldsymbol{\pi} = \mathbf{v}^T$, where $\mathbf{v}$ is an eigenvector with eigenvalue 1.

Note, that the *eigenvectors are only guaranteed to be real-valued if the matrix is positive*, $A_{ij} > 0$ (and hence $A_{ij} < 1$, due to the sum-to-one constrain). A more general approach, which can handle chains where some transition probabilities are 0 or 1 is as follows.

We have K constraints from $\boldsymbol{\pi}(\mathbf{I} - \mathbf{A}) = \mathbf{0}_{K\times 1}$ and 1 constrain from $\boldsymbol{\pi}\mathbf{1}_{K\times 1} = 0$. Let us replace any column, e.g. the last, of $\mathbf{I} - \mathbf{A}$ with $\mathbf{1}$, to get a new matrix $\mathbf{M}$. Next we define $\mathbf{r} = [0,0,\cdots,1]$ where the 1 in the last position corresponds to the column of all 1 in $\mathbf{M}$. We then solve $\boldsymbol{\pi}\mathbf{M} = \mathbf{r}$.

Conditions of stationary distribution (1/2)

A Markov chain is called ***irreducible*** when we can get from any state to any other state.

A chain has a ***limiting distributions*** if $\pi_j = \lim_{n \to +\infty} {A_{ij}}^n$ exists and is independent of i, for all j. If this holds, then the long-run distribution over states will be independent of the starting state:

$$p(X_t = j) = \sum_i p(X_0 = i) A_{ij}(t) \to \pi_j \; as \; t \to +\infty$$

Define the ***period*** of state i to be

$$d(i) = \gcd\{t : A_{ii}(t) > 0\}$$

Where gcd stands for *greatest common divisor*. We say a ***state*** i ***is aperiodic*** if $d(i) = 1$.

We say a ***chain is aperiodic*** if all its states are aperiodic.

Conditions of stationary distribution (2/2)

Recurrent state means that you will return to that state with probability 1.

Regular chain is one whose transition matrix satisfies ${A_{ij}}^n > 0$ for some integer n and all i, j.

We say a ***state is ergodic*** if it is aperiodic, recurrent and not-null, and a ***chain is ergodic*** if all its states are ergodic.

Theorem: Every irreducible (singly connected), aperiodic finite state Markov chain has a limiting distribution, which is equal to $\boldsymbol{\pi}$, its unique stationary distribution.

Theorem: Every irreducible (singly connected), ergodic Markov chain has a limiting distribution, which is equal to $\boldsymbol{\pi}$, its unique stationary distribution.

Detailed balance

Establishing ergodicity can be difficult. We now give an alternative condition that is easier to verify.

We say that a Markov chain $\mathbf{A}$ is ***time reversible*** if there exists a distribution $\boldsymbol{\pi}$ such that

$$\pi_i A_{ij} = \pi_j A_{ji}$$

These are called the ***detailed balance equations***. This says that the flow from i to j must equal the flow from$j$ to i, weighted by the appropriate source probabilities.

Theorem: If a Markov chain with transition matrix A is regular and satisfies detailed balance wrt distribution π, then π is a stationary distribution of the chain.

Hidden Markov models

A ***hidden Markov model*** (***HMM***) consists of a discrete-time, discrete state Markov chain, with hidden state $z_t \in \{1,2 \cdots K\}$, plus an ***observation model*** $p(\mathbf{x}_y|z_t)$. The corresponding distribution has the form:

$$p(\mathbf{z}_{1:T}, \mathbf{x}_{1:T}) = p(\mathbf{z}_{1:T})p(\mathbf{x}_{1:T}|\mathbf{z}_{1:T}) = \left[p(z_1)\prod_{t=2}^{T} p(z_t|z_{t-1})\right]\left[\prod_{t=1}^{T} p(\mathbf{x}_t|z_t)\right]$$

The observation in an HMM can be discrete or continuous. If they are discrete, it is common for the observation model to be an observation matrix:

$$p(\mathbf{x}_t = l|z_t = k, \boldsymbol{\theta}) = B(k, l)$$

If the observations are continuous, it is common for the observation model to be a conditional Gaussian:

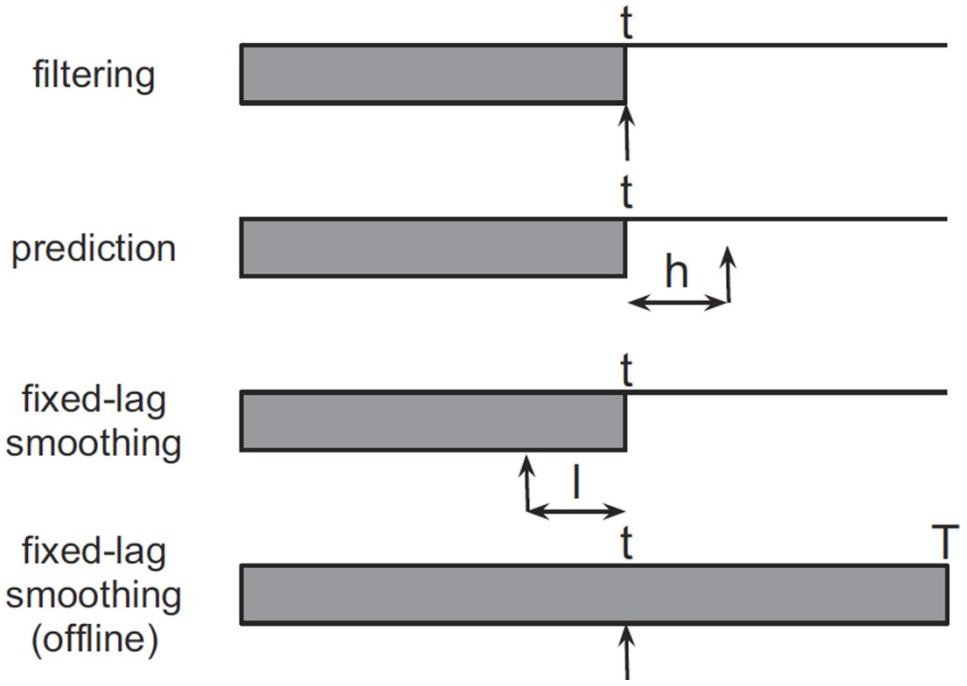
$$p(\mathbf{x}_t|z_t = k, \boldsymbol{\theta}) = \mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Applications of HMMs

- Automatic speech recognition;
- Activity recognition;
- Part of speech tagging;
- Gene finding;
- Protein sequence alignment.

Types of inference problems in HMMs (1/2)

- ***Filtering*** means to compute the ***belief state*** $p(z_t|\mathbf{x}_{1:t})$ online, or recursively, as the data streams in;
- Smoothing means to compute $p(z_t|\mathbf{x}_{1:T})$ offline, given all the evidence;
- ***Fixed lag smoothing*** is an interesting compromise between online and offline estimation – it involves computing $p(z_{t-l}|\mathbf{x}_{1:t})$ where $l > 0$ is called the lag;
- ***Prediction*** – we want to predict the future given the past, i.e. to compute $p(z_{t+h}|\mathbf{x}_{1:t})$, where $h > 0$ is called the ***prediction horizon***.



Types of inference problems in HMMs (2/2)

- ***MAP estimation*** means computing $\operatorname{argmax}_{\mathbf{z}_{1:T}} p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$, which is a most probable state sequence;
- ***Posterior samples*** $\mathbf{z}_{1:T} \sim p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$;
- ***Probability of the evidence*** means summing up over all hidden paths, $p(\mathbf{x}_{1:T}) = \sum_{\mathbf{z}_{1:T}} p(\mathbf{z}_{1:T}, \mathbf{x}_{1:T})$

The forwards algorithm

We now describe the *online algorithm* to recursively compute the filtered marginals, $p(z_t|\mathbf{x}_{1:t})$ in an HMM.

First comes the *prediction step*, in which we compute the ***one-step-ahead predictive density*** (this acts as the new prior for time t):

$$p(z_t = j|\mathbf{x}_{1:t-1}) = \sum_i p(z_t = j|z_{t-1} = i)p(z_{t-1} = i|\mathbf{x}_{1:t-1})$$

Next comes the *update step*, in which we absorb the observed data from time t using Bayes rule:

$$\alpha_t(j) \triangleq p(z_t = j|\mathbf{x}_{1:t}) = \frac{p(\mathbf{x}_t|z_t = j)p(z_t = j|\mathbf{x}_{1:t-1})}{\sum_j p(\mathbf{x}_t|z_t = j)p(z_t = j|\mathbf{x}_{1:t-1})}$$

This process is known as the ***predict-update cycle***. The distribution $p(z_t|\mathbf{x}_{1:t})$ is called the (filtered) ***belief state*** at time t.

The forwards-backwards algorithm (1/2)

We now discuss how to compute the smoothed marginals, $p(z_t|\mathbf{x}_{1:t})$ using *offline inference*.

The key decomposition relies on the fact that we can break the chain into two parts, the past and the future, by conditioning on z_t:

$$p(z_t = j|\mathbf{x}_{1:T}) \propto p(z_t = j, \mathbf{x}_{t+1:T}|\mathbf{x}_{1:t}) \propto p(z_t = j|\mathbf{x}_{1:t})p(\mathbf{x}_{t+1:T}|z_t = j)$$

Let $\alpha_t(j) \triangleq p(z_t = j|\mathbf{x}_{1:t})$ be the filtered belief state as before. Also, define:

$$\beta_t(j) \triangleq p(\mathbf{x}_{t+1:T}|z_t = j)$$

as the conditional likelihood of the future evidence given that the hidden state at time t is j. Note that $\beta_t(j)$ is not probability distribution over states, since it does not need to satisfy $\sum_j \beta_t(j) = 1$.

The forwards-backwards algorithm (2/2)

Finally, we define

$$\gamma_t(j) \triangleq p(z_t = j|\mathbf{x}_{1:T})$$

as the desired smoothed posterior marginal.

We have already described how to recursively compute the α's in the *forwards algorithm*. We now describe how to recursively compute the β's. If we have already computed β_t, we can compute β_{t-1} as follows:

$$\boldsymbol{\beta}_{t-1} = \boldsymbol{\Psi}(\boldsymbol{\psi}_t \odot \boldsymbol{\beta}_t)$$

where $\boldsymbol{\psi}_t(j) = p(\mathbf{x}_t|z_t = j)$ is the local evidence at time t; $\boldsymbol{\Psi}(i,j)$ is the transition matrix and $\odot$ is the Hadamard product.

Having computed the forwards and backwards messages, we can combine them to compute $\gamma_t(j) \propto \alpha_t(j)\beta_t(j)$

Viterbi algorithm

The Viterbi algorithm can be used to compute the most probable sequence of states in a chain-structured graphical model, i.e. it can compute

$$\mathbf{z}^* = \underset{\mathbf{z}_{1:T}}{\operatorname{argmax}}\, p(\mathbf{z}_{1:T}|\mathbf{x}_{1:T})$$

This is equivalent to computing a shortest path through the trellis diagram, where the nodes are possible states at each time step, and the node and edge weights are log probabilities. That is, the weight of a path $z_1, z_2, \cdots z_T$ is given by

$$\left\{ \log[\pi_1(z_1)] + \log[\phi_1(z_1)] + \sum_{t=2}^{T} [\log[\psi(z_{t-1}, z_t)] + \log[\phi_t(z_t)]] \right\} \rightarrow min$$

EM for HMMs (Baum-Welch algorithm, 1/2)

If the z_t variables are not observed, we are in a situation analogous to fitting a mixture model. The most common approach is to use the EM algorithm to find the MLE or MAP parameters. When applied to HMMs, this is also known as the ***Baum-Welch algorithm***.

E step. The expected complete data log likelihood is given by

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{old}) = \sum_{k=1}^{K} \mathbb{E}[N_k{}^1] \log[\pi_k] + \sum_{j=1}^{K} \sum_{k=1}^{K} \mathbb{E}[N_{jk}] \log[A_{jk}] + \sum_{i=1}^{N} \sum_{t=1}^{T_i} \sum_{k=1}^{K} p(z_t = k|\mathbf{x}_i, \boldsymbol{\theta}^{old}) \log[p(\mathbf{x}_{i,t}|\phi_k)]$$

EM for HMMs (Baum-Welch algorithm, 2/2)

E step (continuation).

$$\mathbb{E}\left[N_k{}^1\right] = \sum_{i=1}^{N} p\left(z_{i1} = k | \mathbf{x}_i, \boldsymbol{\theta}^{old}\right)$$

$$\mathbb{E}\left[N_j\right] = \sum_{i=1}^{N} \sum_{t=2}^{T_i} p\left(z_{i,t} = j | \mathbf{x}_i, \boldsymbol{\theta}^{old}\right)$$

$$\mathbb{E}\left[N_{jk}\right] = \sum_{i=1}^{N} \sum_{t=2}^{T_i} p\left(z_{i,t-1} = j, z_{i,t} = k | \mathbf{x}_i, \boldsymbol{\theta}^{old}\right)$$

M step. According to results, obtained before for mixture models, we have that the M step for $\mathbf{A}$ and $\boldsymbol{\pi}$ is to just normalize the expected counts:

$$\hat{A}_{jk} = \frac{\mathbb{E}\left[N_{jk}\right]}{\sum_{\acute{k}} \mathbb{E}\left[N_{j\acute{k}}\right]}$$

$$\hat{\pi}_k = \frac{\mathbb{E}\left[N_k{}^1\right]}{N}$$

Generalizations of HMMs

- Variable duration (semi-Markov) HMMs;
- Hierarchical HMMs;
- Input-output HMMs;
- Auto-regressive and buried HMMs;
- Factorial HMMs;
- Coupled HMM and the influence model;
- Dynamic Bayesian networks (DBNs).

Conclusion

- Definition of Markov and Hidden Markov models was presented;
- Methods for Markov model parameters estimation were considered;
- Conditions for revealing the stationary distribution of a Markov chain were shown;
- Generalizations of HMMs were considered.