# MVE550 2019 Lecture 5

Petter Mostad

Chalmers University

November 19, 2019

# Overview lecture 5: HMMs

- Hidden Markov Models: Ideas and definitions.
- Example: Discriminating between types of DNA sequences using HMMs.
- Inference for HMMs.
- The Forward and Backward algorithms.
- The Multinomial Dirichlet conjugacy.
- Bayesian learning about parameters of Markov chains.

# Hidden Markov Models

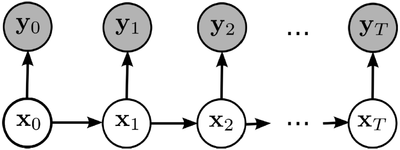


Figure: A hidden Markov model.

- A Hidden Markov Model (HMM) consists of
  - a Markov chain $X_0, \dots, X_n, \dots,$, and
  - another chain $Y_0, \dots, Y_n, \dots$, so that

$$\Pr(Y_k \mid Y_0, \dots, Y_{k-1}, X_0, \dots, X_k) = \Pr(Y_k \mid Y_{k-1}, X_k)$$

- In some models $Y_k$ depends only on $X_k$, not on $Y_{k-1}$.
- Generally, $Y_0, \dots, Y_k \dots$, are *observed*, while $X_0, \dots, X_k \dots$, are *hidden*.
- In our applications, the $X_k$ have a finite state space and the $Y_k$ are discrete.

# Inference questions for HMMs

- If the HMM parameters are given and the $Y_i$ are observed, "find" the values of the $X_i$'s:
  - Find the sequence $X_0, \ldots, X_k$ maximizing the probability of the observed $Y_0, \ldots, Y_k$ in the given model: The *Viterbi algorithm* (not part of course).
  - Find the *joint distribution* of $X_0, \ldots, X_k$ given the observed $Y_0, \ldots, Y_k$ and the model. (In practice: Find a sequence $X_0, \ldots, X_k$ that is a *sample* from this joint distribution).
  - Find the marginal distribution for each $X_i$ given the observed $Y_0, \ldots, Y_k$ and the model: The Forward-Backward algorithm, see below.
- If the HMM parameters are NOT known:
  - If the $X_i$ and $Y_i$ are observed for some "training data", we can infer parameters (see below).
  - More advanced algorithms can be used if the training data is only partially observed.

# The Forward algorithm

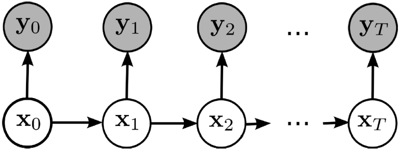


Figure: A hidden Markov model.

- The forward algorithm: For $i = 0, \dots, T$, compute $\pi(X_i \mid Y_0, \dots, Y_{i-1})$.
- A recursive formula is possible to obtain:
  - Obtain $\pi(X_i \mid Y_0, \dots, Y_i)$ from $\pi(X_i \mid Y_0, \dots, Y_{i-1})$ using Bayes formula.
  - Obtain $\pi(X_{i+1} \mid Y_0 \dots, Y_i)$ from $\pi(X_i \mid Y_0 \dots, Y_i)$ using the transition matrix.
- Note: A version of the algorithm can also be made if there is a direct dependency of $Y_i$ on $Y_{i-1}$.

# The Backward algorithm

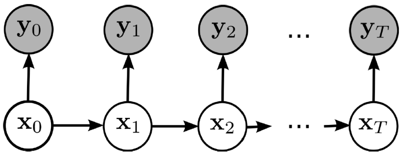


Figure: A hidden Markov model.

- The backward algorithm: For $i = T, \ldots, 0$, compute $\pi(Y_i, \ldots, Y_T \mid X_i)$.
- A recursive formula is possible to obtain:

$$\pi(Y_i, \ldots, Y_T \mid X_i) =$$
$$\pi(Y_i \mid X_i) \sum_{X_{i+1}} \pi(Y_{i+1}, \ldots, Y_T \mid X_{i+1}) \pi(X_{i+1} \mid X_i)$$

- Note: A version of the algorithm can also be made if there is a direct dependency of $Y_i$ on $Y_{i-1}$.

# The Forward Backward algorithm

- Put the two together using

$$\pi(X_i \mid Y_0, \ldots, Y_T) \propto_{X_i} \pi(Y_i, \ldots, Y_T \mid X_i)\pi(X_i \mid Y_0, \ldots, Y_{i-1})$$

- One can use the forward algorithm together with an adaptation of the backward algorithm to find a sequence $x_0, \ldots, x_T$ that is a sample from

$$\pi(X_0, \ldots, X_T \mid Y_0, \ldots, Y_T)$$

- Actual implementation depends on the types of variables involved.

# The Multinomial Dirchlet conjugacy

- A vector $x = (x_1, \ldots, x_k)$ of non-negative integers has a Multinomial distribution with parameters $n$ and $p$, where $n > 0$ is an integer and $p$ is a probability vector of length $k$ if $\sum_{i=1}^{k} x_i = n$ and the probability mass function is given by

$$\pi(x \mid n, p) = \frac{n!}{x_1!x_2!\ldots x_k!} p_1^{x_1} p_2^{x_2} \ldots p_k^{x_k}.$$

- A vector $\theta = (\theta_1, \ldots, \theta_k)$ of non-negative real numbers satisfying $\sum_{i=1}^{k} \theta_i = 1$ has a Dirichlet distribution with parameter vector $\alpha = (\alpha_1, \ldots, \alpha_k)$, if it has probability density function

$$\pi(\theta \mid \alpha) = \frac{\Gamma(\alpha_1 + \alpha_2 + \cdots + \alpha_k)}{\Gamma(\alpha_1)\Gamma(\alpha_2) \cdot \Gamma(\alpha_k)} \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \cdots \theta_k^{\alpha_k - 1}.$$

- We have conjugacy in this case.
- The predictive distribution is given by

$$\pi(x) = \frac{n!}{x_1! \ldots x_k!} \cdot \frac{\Gamma(\alpha_1 + x_1)}{\Gamma(\alpha_1)} \cdots \frac{\Gamma(\alpha_k + x_k)}{\Gamma(\alpha_k)} \cdot \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\Gamma(\sum_{i=1}^{k} \alpha_i + x_i)}.$$

# Learning about the transition matrix from data

- We can use as prior a product of Dirichlet densities, one for each row of the transition matrix.
- The posterior is then also a product of Dirichlet densities.
- Use of *pseudocounts* and comparison to using frequencies.
- Computation of predictions for one or more further steps of the chain.
- Other alternatives for the prior can be used.