Lecture 6: Markov models

Statistical Methods for Natural Language Processing

Fredrik Engström

February 16, 2011

Summary of lecture 5

- $I(X;Y) = H(X) - H(X|Y)$
- The cross-entropy between p and q is

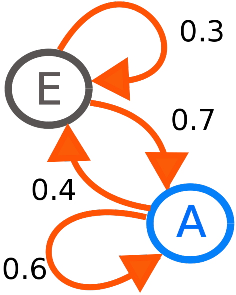
$$\sum_x p(x) \log \frac{1}{q(x)}.$$

- MLE: Maximize $P(x_1, \ldots, x_k|\theta)$
- Needs smoothing with sparse data.
- Bayesian: Maximize $P(\theta|x_1, \ldots, x_k)$. Same as maximizing $P(x_1, \ldots, x_k|\theta)P(\theta)$

Markov models

Intuition: Some random process changing over time with the following properties:

- **Memoryless**, meaning that only the present state and **not** the past affects the future states.
- **Stationary**, meaning that it is time homogeneous.



Examples: Bigram analysis. Fia med knuff. Trigram(?). Cards(?).

Definition

Definition

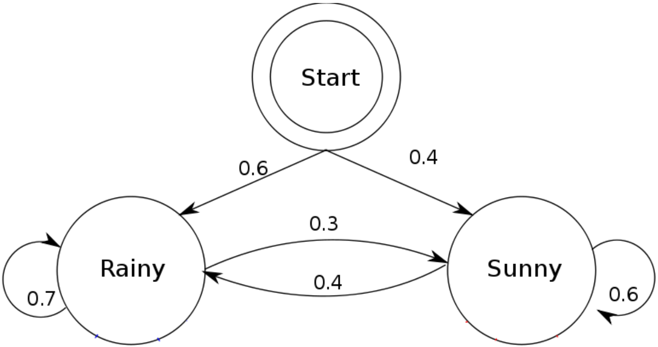
A **Markov model** is an (infinite) sequence of random variables $X_1, X_2, \ldots$ such that

- $P(X_{k+1} = y | X_1 = x_1, \ldots X_k = x_k) = P(X_{k+1} = y | X_k = x_k)$
- $P(X_{k+1} = y | X_k = x) = P(X_2 = y | X_1 = x)$

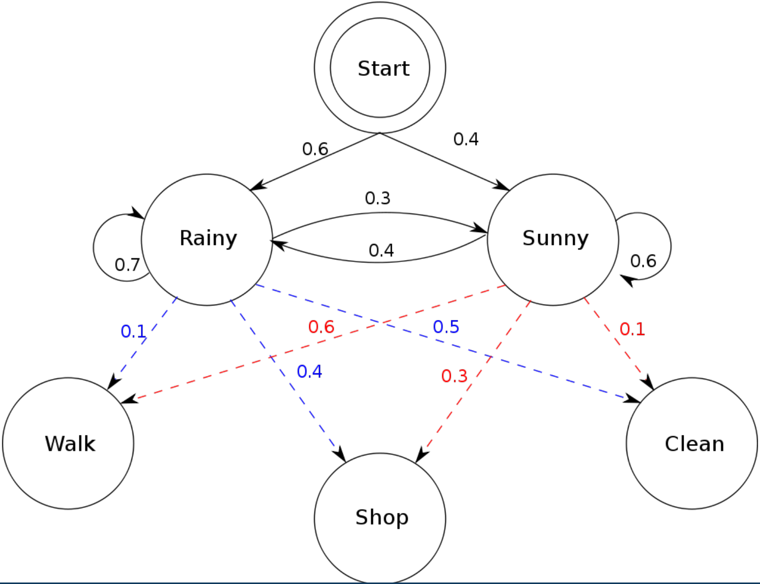
Alternative definition as a **finite state machine**:

- Set of **states** S.
- **Initial state probabilities** π_i for $i \in S$.
- **State transition probabilities** a_{ij} for $i, j \in S$.

Example: Weather



Hidden Markov model



Hidden Markov model

Definition of a **(state-emitting) hidden Markov model**:

- Set of **states** S.
- Set of **outputs** K.
- **Initial state probabilities** π_i for $i \in S$.
- **State transition probabilities** a_{ij} for $i, j \in S$.
- **Output emission probabilities** b_{ik} for $i \in S, k \in K$.

Hidden Markov model

Definition of a **(arc-emitting) hidden Markov model**:

- Set of **states** S.
- Set of **outputs** K.
- **Initial state probabilities** π_i for $i \in S$.
- **State transition probabilities** a_{ij} for $i, j \in S$.
- **Output emission probabilities** b_{ijk} for $i, j \in S, k \in K$.

Inferences

The three inferences:

1. Given a HMM μ what is the probability of a certain output sequence O, i.e., what is $P(O|\mu)$?
2. Given an output sequence O and a HMM μ what is the best guess at at a state sequence explaining the output O, i.e. which state sequence X maximizes $P(X|O,\mu)$?
3. Given an output sequence O what is the best guess at a HMM explaining the output?

Next we give solutions to 1 and 2.

The forward algorithm

Given a HMM μ what is the probability of a certain output sequence O, i.e., what is $P(O|\mu)$?

Let $\alpha_i(t)$ be $P(o_1, \ldots, o_t, X_t = s_i)$.

- $\alpha_i(1) = P(X_1 = s_i) b_{io_1} = \pi_i b_{io_1}$.
- $\alpha_i(t+1) = P(o_1, \ldots, o_{t+1}, X_{t+1} = s_i) =$
  $$b_{io_{t+1}} \sum_{j=1}^{N} P(o_1, \ldots, o_t, X_t = s_j) a_{ji} = b_{io_{t+1}} \sum_{j=1}^{N} \alpha_j(t) a_{ji}$$

Example: $O =$('walk', 'shop', 'clean'), $s_1 =$Rainy $s_2 =$Sunny

	t=1	2	3
s_1	$\alpha_1(1)$	$\alpha_1(2)$	$\alpha_1(3)$
s_2	$\alpha_2(1)$	$\alpha_2(2)$	$\alpha_2(3)$

	t=1	2	3
s_1	.06	.0552	.02904
s_2	.24	.0486	.004572

P('walk', 'shop', 'clean')$= .02904 + .004572 = .033612$

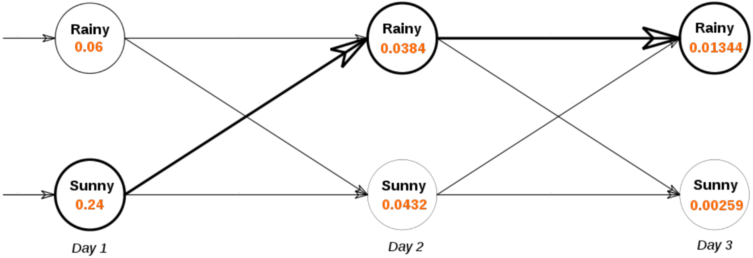
The Viterbi algorithm

Given an output sequence O and a HMM μ what is the best guess at at a state sequence explaining the output $O = o_1, \ldots, o_T$, i.e. which state sequence X maximizes $P(X|O, \mu)$?

Let $\delta_i(t)$ be

$$\max_{x_1,\ldots,x_{t-1}} P(X_1 = x_1, \ldots, X_{t-1} = x_{t-1}, X_t = s_i, o_1, \ldots, o_t).$$

- $\delta_i(1) = P(X_1 = s_i) b_{io_1} = \pi_i b_{io_1}$.
- $\Psi_i(1) = 0$
- $\delta_i(t+1) = \max\limits_{x_1,\ldots,x_t} P(X_1{=}x_1, \ldots, X_t{=}x_t, o_1, \ldots, o_{t+1}) = b_{io_{t+1}} \max\limits_j(\delta_j(t) a_{ji})$
- $\Psi_i(t+1) = \text{argmax}_j \, \delta_j(t) a_{ji}$
- $\hat{X}_T = \text{argmax}_i \, \delta_T(i)$
- $\hat{X}_t = \Psi_{\hat{X}_{t+1}}(t+1)$


Rainy
0.06
Sunny
0.24
Day 1
Rainy
0.0384
Sunny
0.0432
Day 2
Rainy
0.01344
Sunny
0.00259
Day 3

Summary

- Markov models
- Hidden (state / arc emitting) Markov models
- Forward algorithm
- Viterbi algorithm