A type-safe introduction to

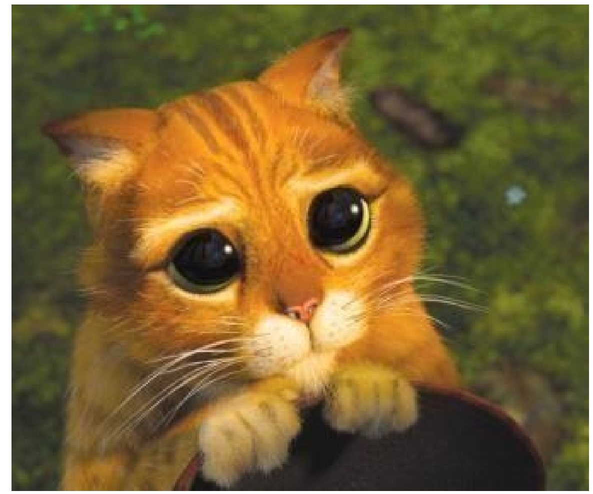
Probability Theory

Fabian M. Suchanek

Probability theory

starring主演

and



Shrek

Puss

In all of the following, if there is a definition followed by a "special case", it is fully sufficient to know and understand the special case for the purpose of most applications in Information Extraction, including Markov Random Fields, Hidden Markov Models, Markov Chains, and Conditional Random Fields.

Overview

Introduction to Probabilities

	Chains	Complex dependencies and/or feature functions
only visible variables	Markov Chains	Markov Random Fields
visible and invisible variables	Hidden Markov Models	Conditional Random Fields

Overview

Introduction to Probabilities

	Chains	Complex dependencies and/or feature functions
only visible variables	Markov Chains	Markov Random Fields
visible and invisible variables	Hidden Markov Models	Conditional Random Fields

Def: Universe

A universe (also: sample space) is the set of all possible worlds (also: possible states of the world, outcomes of a random trial).

$$\Omega = \{w_1, ..., w_n\}$$

Special case: Universe

Given n named sets $X_1, ..., X_n$, we use as universe $\Omega = X_1 \times ... \times X_n$.

Example: X1={Shrek,Puss},
X2={shouts,purrs}

world	X1	X2
w1:	Shrek	shouts
w2:	Puss	purrs
w3:	Shrek	purrs
w4:	Puss	shouts

possible world

universe

Set of events / Sigma algebra

A sigma algebra (also: set of events) over a universe

is a set of subsets of the universe that fulfills certain conditions.

$$F \subseteq 2^{\Omega}$$

It is the set of

"allowable events"

Special case: Sigma algebra

We use as sigma-algebra the set of all subsets of the universe:

$$F = 2^{\Omega}$$

Def: Probability

A probability (also: probability measure, probability distribution) is a function

$$P : F \to [0,1]$$

such that

$$P(\emptyset) = 0$$

$$P(\Omega) = 1$$

$P(\cup_i E_i) = \sum_i P(E_i)$ for a finite number of disjoint sets E_i .

For general (not necessarily disjoint) sets E_i, we have:

$P(\cup_i E_i) \leq \sum_i P(E_i)$ "union bound"

...because the E_i may overlap, in which case $|\cup_i E_i| < \sum_i |E_i|$.

Special case: Probability

We use as probability a function that maps every possible world to [0,1].

$$P : \Omega \rightarrow [0, 1]$$

	X1	X2	
w1:	Shrek	shouts	P(w1)=0.4
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1
w4:	Puss	shouts	P(w4)=0.2

Special case: Probability

For our probability function, we define
$P(\{w_1, ..., w_n\}) = \sum_i P(w_i)$ for $w_i \in \Omega$.

P({w1,w2}) = P(w1)+P(w2) = 0.7

	X1	X2	
w1:	Shrek	shouts	P(w1)=0.4
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1
w4:	Puss	shouts	P(w4)=0.2

Def: Probability space

A probability space is a triple of a universe, a sigma algebra, and a probability measure: (Ω, F, P)

We assume a fixed probability space from now on.

	X1	X2	
w1:	Shrek	shouts	P(w1)=0.4
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1
w4:	Puss	shouts	P(w4)=0.2

Def: Random variable

A random variable is a function that takes a possible world and returns a value (also: state).

$$X : \Omega \rightarrow \{s_1, ..., s_m\}$$

(The random variable “extracts” a feature from the state of the world)

Example: Random variable

The Random Variable A

$$A : \Omega \rightarrow \{scary, cosy\}$$

describes the atmosphere of a world.

	X1	X2	
w1:	Shrek	shouts	A(w1)=scary
w2:	Puss	purrs	A(w2)=cosy
w3:	Shrek	purrs	A(w3)=cosy
w4:	Puss	shouts	A(w4)=cosy

Are Random Variables random?

Random variables are

- neither random
- nor variables

They take a possible world and
return a characteristic of that world.

$$A(w_1) = cosy$$

In our special case: They are just
the components of the universe.

$$X_2(< Shrek, shouts >) = shouts$$

Special case: Random variable

In our universe, the named sets can be considered random variables.

We consider only these as random variables:

$$X_i(< w_1, ..., w_n >) = x_i$$

	X1	X2	
w1:	Shrek	shouts	X1(w1)=Shrek
w2:	Puss	purrs	X1(w2)=Puss
w3:	Shrek	purrs	X1(w3)=Shrek
w4:	Puss	shouts	X1(w4)=Puss

Def: Events

An event is a subset of the universe.

Event where someone purrs: $\{$w2, w3$\}$

	X1	X2
w1:	Shrek	shouts
w2:	Puss	purrs
w3:	Shrek	purrs
w4:	Puss	shouts

Special case: Events

We define an event as "$X = x$" $:= \{w : w \in \Omega, X(w) = x\}$.

"X2=purrs" = $\{$w2, w3$\}$

	X1	X2
w1:	Shrek	shouts
w2:	Puss	purrs
w3:	Shrek	purrs
w4:	Puss	shouts

Note that
X=x
is not a
statement,
but a shorthand
notation for a set
of possible worlds.

Events

An event has a probability:

$$P(X = x) = P(\{w_i : X(w_i) = x\}) = \sum_{i, X(w_i)=x} P(w_i)$$

P(X1=Shrek) = P({w1,w3}) = 0.5

	X1	X2	
w1:	Shrek	shouts	P(w1)=0.4
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1
w4:	Puss	shouts	P(w4)=0.2

Set operations on events

$$P(X_1 = Shrek \cap X_2 = purrs)$$

$$= P(\{w_1, w_3\} \cap \{w_2, w_3\})$$

$$= P(\{w_3\})$$

$$= P(w_3) = 0.1$$

X=x is a set,
and can hence
undergo set
operations such as

$\cap, \cup$

	X1	X2	
w1:	Shrek	shouts	P(w1)=0.4
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1
w4:	Puss	shouts	P(w4)=0.2

Syntax: Intersections

For the intersection, we write

- a comma-separated list

$$P(X_1 = x_1, ..., X_n = x_n) := P(X_1 = x_1 \cap ... \cap X_n = x_n)$$

- vectors

$$\vec{X} = \vec{x} := X_1 = x_1 \cap ... \cap X_n = x_n$$

- single values

$$P(x_1, ..., x_n) := P(X_1 = x_1 \cap ... \cap X_n = x_n)$$

Def: Conditional probability

The conditional probability of an event A given an event B is

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

i.e.: look only at the worlds where B happens. In these cases, compute the ratio of the probabilities where also A happens.

With random variables:

$$P(X = x|Y = y) := \frac{P(X=x \cap Y=y)}{P(Y=y)}$$

This entails:

$$P(X = x \cap Y = y) = P(Y = y) \times P(X = x|Y = y)$$

Example: Conditional probability

$$P(X_1 = Shrek | X_2 = purrs)$$

	X1	X2	
w1:	Shrek	shouts	P(w1)=0.4
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1
w4:	Puss	shouts	P(w4)=0.2

Example: Conditional probability

$$P(X_1 = Shrek | X_2 = purrs)$$

Look at only cases where someone purrs

	X1	X2	
w1:	Shrek	shouts	P(w1)=0.4
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1
w4:	Puss	shouts	P(w4)=0.2

Example: Conditional probability

$$P(X_1 = Shrek | X_2 = purrs)$$

Look at only cases where someone purrs

	X1	X2	
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1

Example: Conditional probability

$$P(X_1 = Shrek | X_2 = purrs)$$

$$= \frac{P(X_1 = Shrek \cap X_2 = purrs)}{P(X_2 = purrs)}$$

	X1	X2	
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1

Example: Conditional probability

$$P(X_1 = Shrek | X_2 = purrs)$$

$$= \frac{P(X_1 = Shrek \cap X_2 = purrs)}{P(X_2 = purrs)}$$

$$= \frac{P(\{w_1, w_3\} \cap \{w_2, w_3\})}{P(\{w_2, w_3\})} = \frac{P(w_3)}{P(w_2) + P(w_3)} = \frac{0.1}{0.3+0.1} = 0.25$$

	X1	X2	
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1

Example: Conditional probability

$$P(X_1 = Shrek | X_2 = purrs)$$

$$= \frac{P(X_1 = Shrek \cap X_2 = purrs)}{P(X_2 = purrs)}$$

$$= \frac{P(\{w_1, w_3\} \cap \{w_2, w_3\})}{P(\{w_2, w_3\})} = \frac{P(w_3)}{P(w_2) + P(w_3)} = \frac{0.1}{0.3 + 0.1} = 0.25$$

	X1	X2
w2:	Puss	purrs
w3:	Shrek	purrs

P(w2)=0.3

P(w3)=0.1

Example: Conditional probability

$$P(X_1 = Shrek | X_2 = purrs)$$

$$= \frac{P(X_1 = Shrek \cap X_2 = purrs)}{P(X_2 = purrs)}$$

$$= \frac{P(\{w_1, w_3\} \cap \{w_2, w_3\})}{P(\{w_2, w_3\})} = \frac{P(w_3)}{P(w_2) + P(w_3)} = \frac{0.1}{0.3 + 0.1} = 0.25$$

If someone purrs, it's unlikely that it's Shrek

	X1	X2	
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1

Def: Independence

Two events A and B are independent if

$P(A \cap B) = P(A) \times P(B)$

This entails $P(A|B) = P(A), P(B|A) = P(B)$

Knowing B does not influence the probability of A, and vice versa.

Example: Independence

Two events A and B are independent if

$P(A \cap B) = P(A) \times P(B)$

Events are not independent (shouting means it's Shrek)

$$P(X_1 = Shrek \cap X_2 = shouts) = 0.4$$

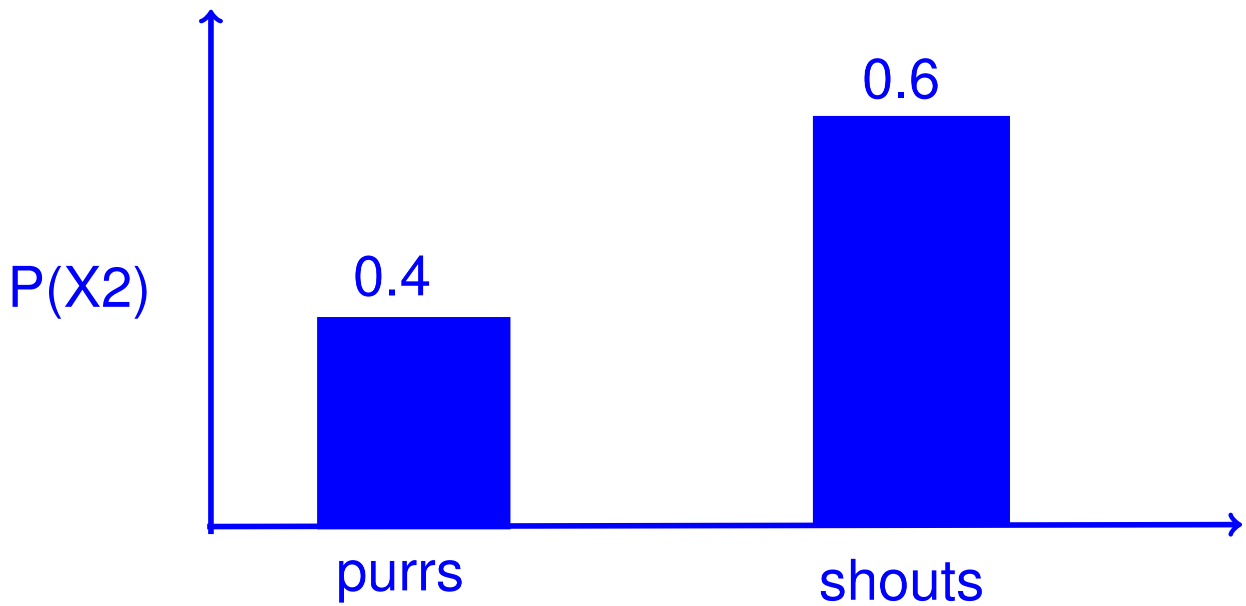
$$\neq P(X_1 = Shrek) \times P(X_2 = shouts) = 0.5 \times 0.6 = 0.3$$

	X1	X2	
w1:	Shrek	shouts	P(w1)=0.4
w2:	Puss	purrs	P(w2)=0.3
w3:	Shrek	purrs	P(w3)=0.1
w4:	Puss	shouts	P(w4)=0.2

Syntax: Distribution

We define for a random variable X: $P(X) := \lambda v : P(X = v)$
i.e., $P(X)(v) = P(X = v)$.

$P(X)$ is a function that takes a value v
and returns the probability $P(X = v)$.

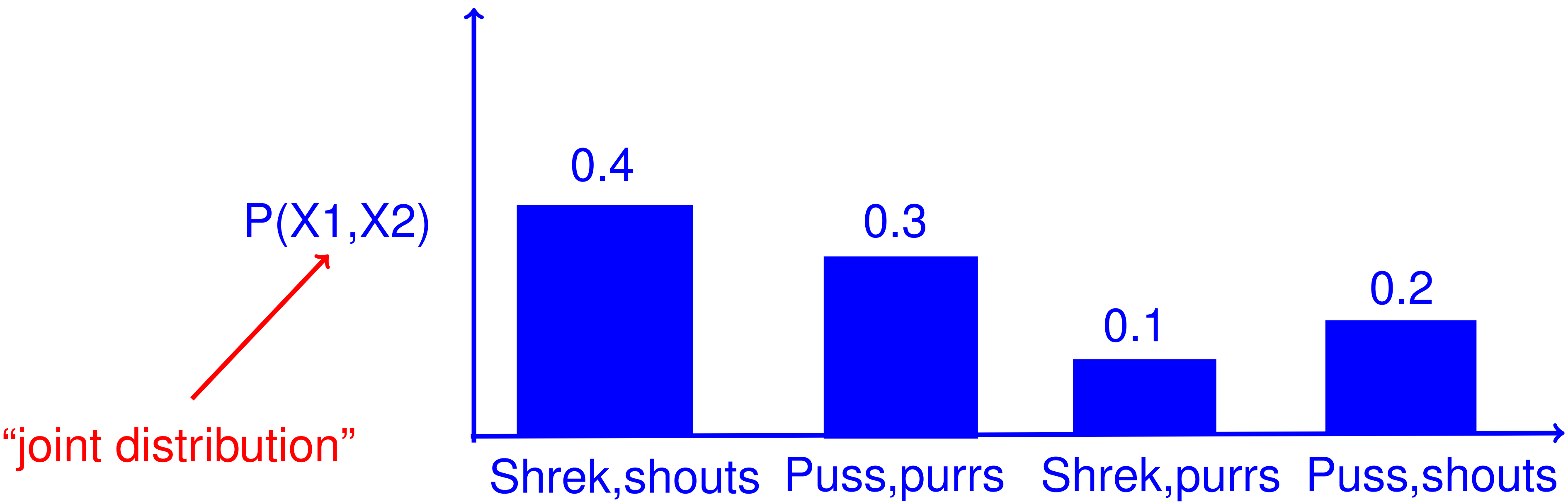


Syntax: Joint Distribution

We define for random variables $X_1, ... X_n$

$$P(X_1, ..., X_n) := \lambda < v_1, ..., v_n >: P(X_1 = v_1 \cap ... \cap X_n = v_n)$$

$P(X, Y)$ is a function that takes a value v for X and a value w for Y and returns $P(X = v \cap Y = w)$.



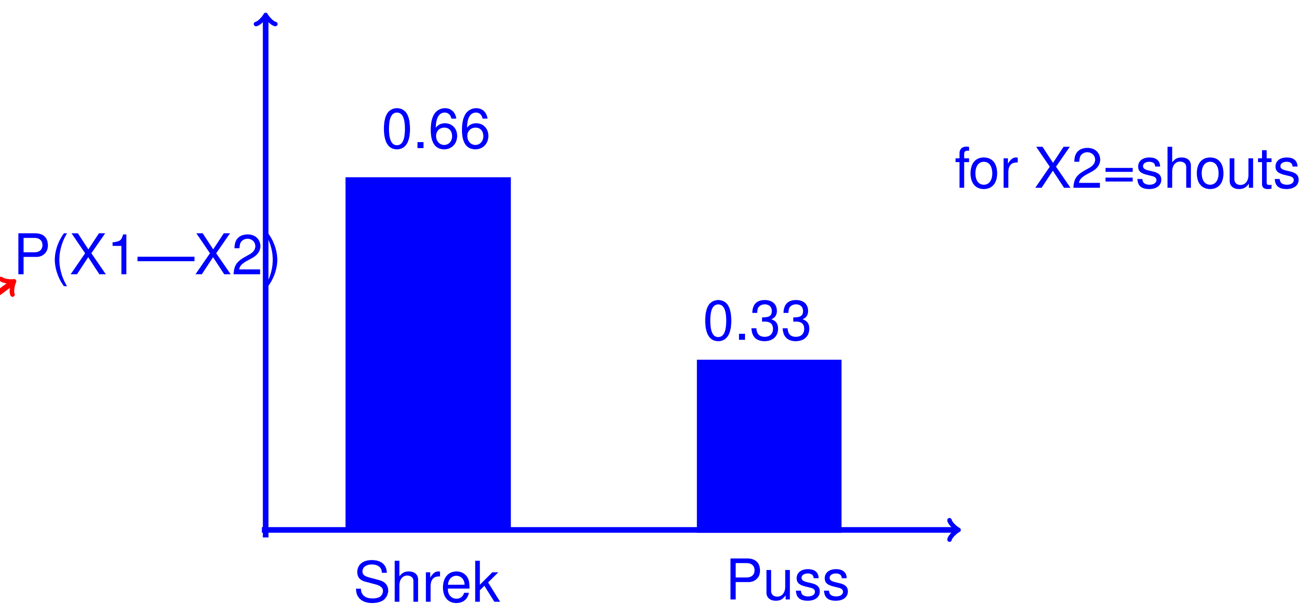
Syntax: Conditional Distribution

We define for random variables $X_1, ...X_n$:

$$P(X|Y_1, ..., Y_n) := \lambda < v_1, ..., v_n >: \lambda v : P(X = v|Y_1 = v_1, ..., Y_n = v_n)$$

$P(X|Y_1, ..., Y_n)$ is a function that takes values for $Y_1, ..., Y_n$ and returns a distribution $P(X)$.

“conditional distribution”



Theorem: Marginals

For any random variables X, Y, we have

$$P(X = x) = \sum_y P(X = x, Y = y)$$

$$= \sum_y P(Y = y) \times P(X = x | Y = y)$$

The probability of $X = x$
can be computed if we have
the conditional probability
$P(X = x|Y = y)$ and $P(Y = y)$.

$P(X)$ is called the marginal distribution
of the joint distribution $P(X, Y)$.

Overview

Introduction to Probabilities

	Chains	Complex dependencies and/or feature functions
only visible variables	Markov Chains	Markov Random Fields
visible and invisible variables	Hidden Markov Models	Conditional Random Fields

Def: Stochastic Process

A stochastic process is a sequence of random variables

$X_0, X_1, X_2, ..., X_i, ...,$ denoted as $\{X_n\}$.

“initial state”

“state at time i”

Def: Markov chain

A Markov chain is a sequence of random variables such that

$$P(X_i|X_1,...,X_{i-1}) = P(X_i|X_{i-1})$$

The probability of $X = v$
depends only on the value of
the predecessor of X.
“The future is independent of the past,
given the present state.”

Example: Markov chain

Let our named sets be the weather on consecutive days, with V={sun,rain}.

D1	D2	D3
sun	sun	sun
sun	sun	rain
...	...	...

This is a Markov chain, if
P(D3—D1,D2) = P(D3—D2)

The probability distribution of D3 given a value for D2
is the same as
the probability distribution of D3 given values for D2 and D1

Def: Homogeneous Markov Chain

A Markov chain $D_1, ..., D_n$ is homogeneous if

$$\forall i, j : P(D_i | D_{i-1}) = P(D_j | D_{j-1})$$

i.e., the conditional probability is the same for all days.

We write $P(v_i | v_j)$ for $P(D_k = v_i | D_{k-1} = v_j)$

e.g. we write P(sunny—rainy)

for P(D2=sunny — D1=rainy)

which is the same as P(D3=sunny — D2=rainy)

which is the same as P(D4=sunny — D3=rainy)

...

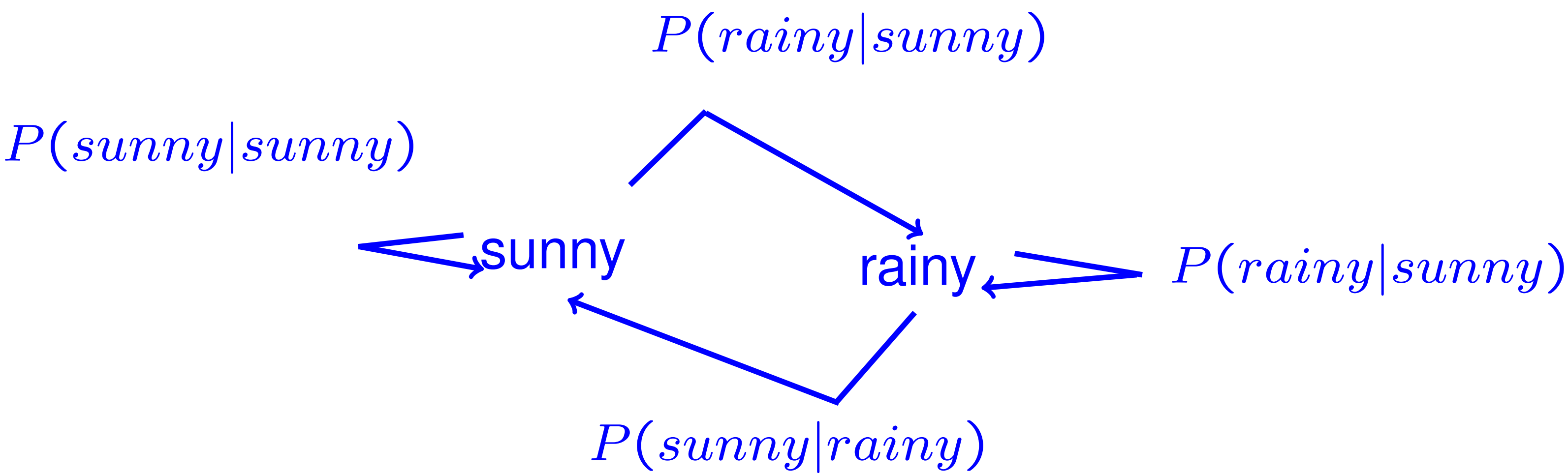
(We will consider only homogeneous Markov Chains for now)

Markov Chains as graphs

We can represent a Markov Chain as a graph,

where the nodes are the values v,

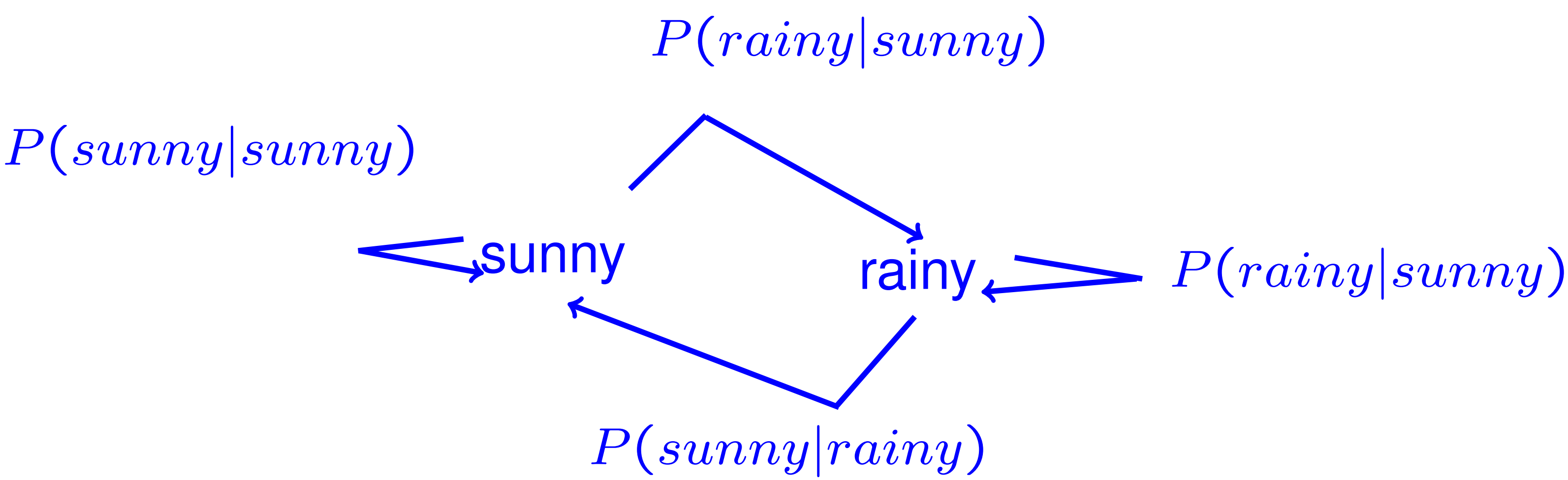
and the edge weights are $P(v_i|v_j)$:



Markov chains as matrices

Transition Matrix T

$$\begin{array}{ccc} & \text{sunny} & \text{rainy} \\ \text{sunny} & P(sunny|sunny) & P(rainy|sunny) \\ \text{rainy} & P(sunny|rainy) & P(rainy|rainy) \end{array}$$

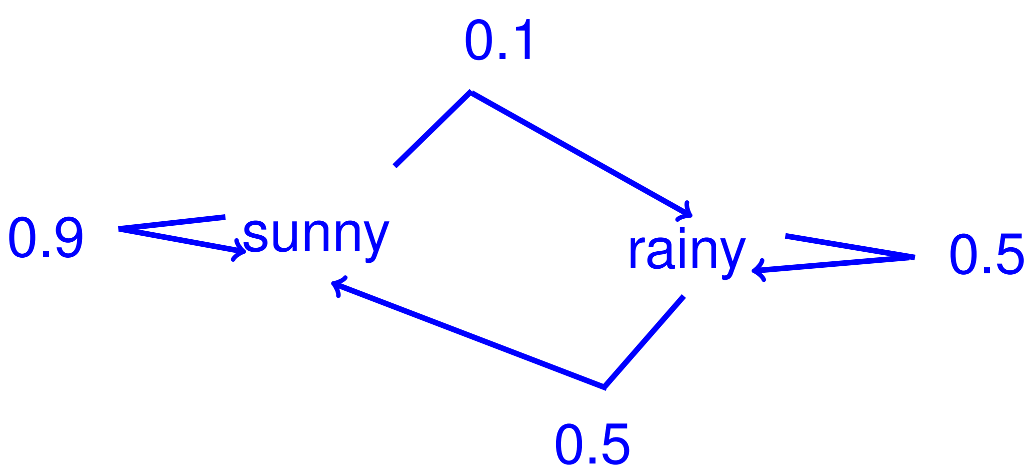


Wikipedia/Markov Chains

Markov chains as matrices

Transition Matrix T

	sunny	rainy
sunny	0.9	0.1
rainy	0.5	0.5



Wikipedia/Markov Chains

Caring about the Joint Probability

We are now interested in one particular sequence of events:

$$P(D_1 = v_1, ..., D_n = v_n)$$

Definition of conditional probability

$$= P(D_n = v_n | D_1 = v_1, ..., D_{n-1} = v_{n-1}) \times P(D_1 = v_1, ..., D_{n-1} = v_{n-1})$$

Markov property

$$= P(D_n = v_n | D_{n-1} = v_{n-1}) \times P(D_1 = v_1, ..., D_{n-1} = v_{n-1})$$

Homogenity

$$= P(v_n | v_{n-1}) \times P(D_1 = v_1, ..., D_{n-1} = v_{n-1})$$

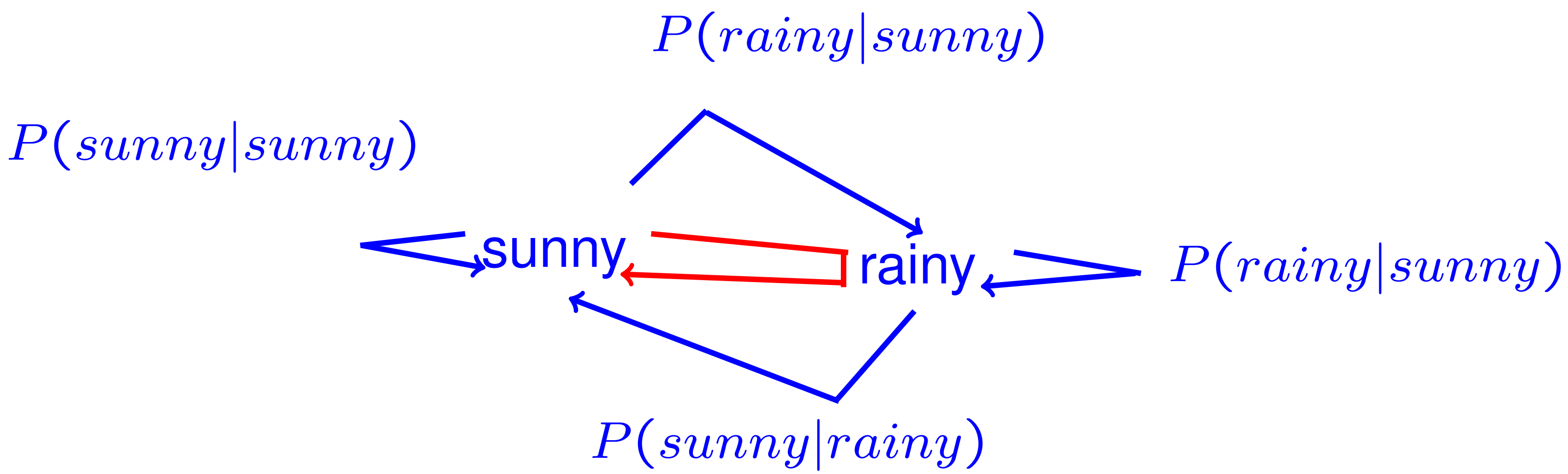
Recursion

$$= \prod_i P(v_i | v_{i-1}) \times P(D_1 = v_1)$$

Joint Probability in the Graph

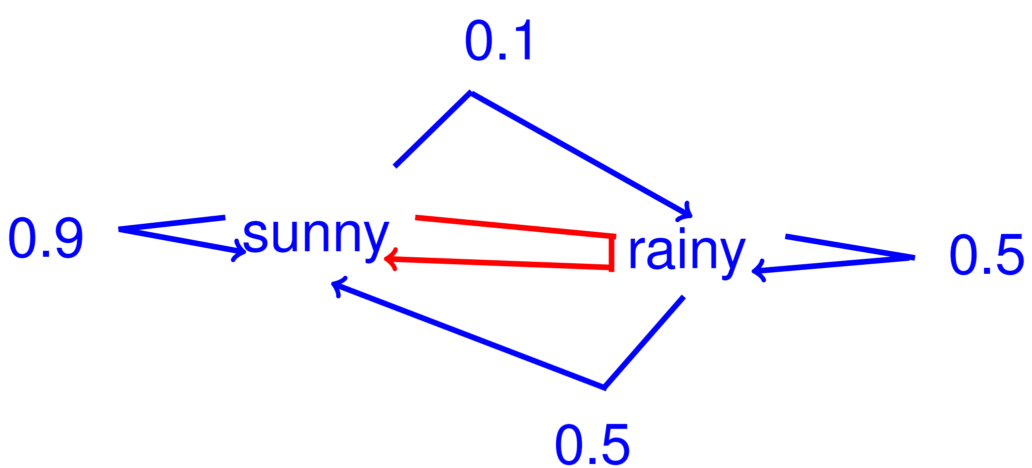
A joint probability corresponds to a path in the graph:

$$P(sunny, rainy, sunny) = P(D_1 = sunny)$$
$$\times P(rainy|sunny)$$
$$\times P(sunny|rainy)$$



Example: Joint Probability in Graph

$$
\begin{aligned}
P(sunny, rainy, sunny) = P(D_1 = sunny) \\
\times P(rainy|sunny) \\
\times P(sunny|rainy) \\
= P(D_1 = sunny) \times 0.1 \times 0.5
\end{aligned}
$$



Caring about the final state

We are now interested in the distribution of the final state:

$$P(D_n = v_n)$$

$$\Big\downarrow \text{ Definition of marginal probability}$$

$$= \sum_{v'} P(D_n = v_n, D_{n-1} = v')$$

$$\Big\downarrow \text{ Definition of conditional probability}$$

$$= \sum_{v'} P(D_n = v_n | D_{n-1} = v') \times P(D_{n-1} = v')$$

$$\Big\downarrow \text{ Homogenity}$$

$$= \sum_{v'} P(v_n | v') \times P(D_{n-1} = v')$$

The final state and the Matrix

$$P(D_n = v_n) = \sum_{v'} P(v_n|v') \times P(D_{n-1} = v')$$

$$P(D_n = sunny) = P(sunny|sunny) \times P(D_{n-1} = sunny)$$
$$+P(sunny|rainy) \times P(D_{n-1} = rainy)$$

Transition Matrix T

	sunny	rainy
sunny	$P(sunny\|sunny)$	$P(rainy\|sunny)$
rainy	$P(sunny\|rainy)$	$P(rainy\|rainy)$

$$P(D_n = sunny) =< P(D_{n-1} = sunny), P(D_{n-1} = rainy) > \times T_{column:sur}$$

The final state and the Matrix

Now let's look at the row vector $\vec{P}(D_n)$:

$$< P(D_n = sunny), P(D_n = rainy) >$$

=

$$< P(D_{n-1} = sunny), P(D_{n-1} = rainy) > \times \begin{pmatrix} P(sunny|sunny) & P(rainy|sunny) \\ P(sunny|rainy) & P(rainy|rainy) \end{pmatrix}$$

That is:

$$\vec{P}(D_n) = \vec{P}(D_{n-1}) \times T$$
$$\vec{P}(D_n) = \vec{P}(D_1) \times T \times ... \times T$$
$$\vec{P}(D_n) = \vec{P}(D_1) \times T^n$$

Def: Stationary Distribution

Now assume that $\vec{P}(D_n) = \vec{P}(D_{n-1})$

$$< P(D_n = sunny), P(D_n = rainy) >$$

=

$$< P(D_{\cancel{n}-1} = sunny), P\cancel{(}D_{n-1} = rainy) > \times \begin{pmatrix} P(sunny|sunny) & P(rainy|sunny) \\ P(sunny|rainy) & P(rainy|rainy) \end{pmatrix}$$

This means that $\vec{P}(D_n) = \vec{P}(D_n) \times T$

=¿ $\vec{P}(D_n)$ is an eigenvector of T !

Such a distribution is called a stationary distribution of T.

It implies that all future states will be equal to $\vec{P}(D_n)$.

Def: Irreducible Markov Chain

A Markov Chain given by a transition matrix T is irreducible,
if for any states i, j, there is an $n > 0$ such that $T^n_{i,j} > 0$.

In other words, the chain is able to move
from any state i to any state j
(in one or more steps).

Def: Period

A state i of a Markov Chain has period k

if any return to i occurs at step $k \times l$, for some $l > 0$

Formally:

$$k = gcd(\{n : P(X_n = i | X_0 = i) > 0\})$$

where gcd denotes the greatest common divisor.

If $k = 1$ then state i is said to be aperiodic.

Def: Ergodic Markov chains

For some Markov Chains, the stationary distribution exists and is unique:

$$q = lim_{n\to\infty} P(D_n)$$

This is the case if $\exists n > 0 : \forall i, j : (T^n)_{i,j} \neq 0$
i.e., already if $\forall i, j : (T)_{i,j} \neq 0$.
Then, T is ergodic.

q is the steady state or stationary distribution.

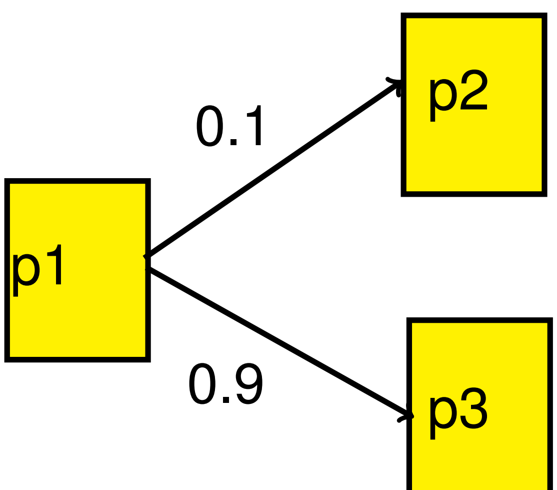
If $q = lim_{n\to\infty} P(D_n)$ exists, then this implies

- q is independent of $P(D_1)$
- q is an eigenvector of T

$$q = q \times T$$

Page Rank

Page Rank can be seen as a Markov Chain, where the values are the pages and $P(X_i = p)$ is the probability that a random surfer visits page p.



Random variables: $X_1, ...X_n$

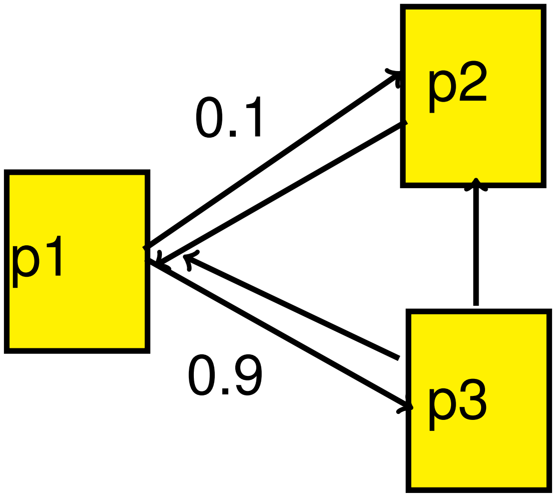
Values: $p_1, ..., p_m$ (=the pages)

$P(X_1 = p_j)$ is the probability that the surfer is at page p_j after i hops.

$P(p_i | p_{i-1})$ is given by the links.

Page Rank

By adding random jumps from all pages to all pages the chain becomes ergodic, and the steady state exists.



$$q = lim_{n \to \infty} P(X_n)$$

Overview

Introduction to Probabilities

	Chains	Complex dependencies and/or feature functions
only visible variables	Markov Chains	Markov Random Fields
visible and invisible variables	Hidden Markov Models	Conditional Random Fields

Hidden Markov Models

A HMM universe has "visible" and "hidden" random variables, with a set of "hidden" and "visible" values:

Usually, we observe the visible variables (here: the sentence) over time and we want to determine the hidden variables (here: the POS tags) over time.

	(visible)		(hidden)		
World	W1	W2	T1	T2	Probability
ω_1	Elvis	sings	PN	Verb	$P(\omega_1) = 0.2$
ω_2	Elvis	sings	Adj	Verb	$P(\omega_2) = 0.1$
ω_3	Elvis	runs	Prep	PN	$P(\omega_3) = 0.1$
	...		...		

Probabilistic POS-Tagging

Given a sentence $w_1, ..., w_n$

we want to find $argmax_{t_1,...,t_n} P(w_1, ..., w_n, t_1, ..., t_n)$.

	(visible)		(hidden)		
World	W1	W2	T1	T2	Probability
ω_1	Elvis	sings	PN	Verb	$P(\omega_1) = 0.2$
ω_2	Elvis	sings	Adj	Verb	$P(\omega_2) = 0.1$
ω_3	Elvis	runs	Prep	PN	$P(\omega_3) = 0.1$
	...		...		

Markov Assumption 1

Every tag depends just on its predecessor

$$P(T_i|T_1, ..., T_{i-1}) = P(T_i|T_{i-1})$$

Markov Assumption 1

Every tag depends just on its predecessor

$$P(T_i|T_1, ..., T_{i-1}) = P(T_i|T_{i-1})$$

The probability that PN, V, D is followed by a noun is the same as the probability that D is followed by a noun:

$$P(N|PN, V, D) = P(N|D)$$

Markov Assumption 1

Every tag depends just on its predecessor

$$P(T_i|T_1, ..., T_{i-1}) = P(T_i|T_{i-1})$$

The probability that PN, V, D is followed by a noun is the same as the probability that D is followed by a noun:

$$P(N|PN, V, D) = P(N|D)$$

Elvis	sings	a	song
PN	Verb	Det	?

Markov Assumption 1

Every tag depends just on its predecessor

$$P(T_i|T_1, ..., T_{i-1}) = P(T_i|T_{i-1})$$

The probability that PN, V, D is followed by a noun is the same as the probability that D is followed by a noun:

$$P(N|PN, V, D) = P(N|D)$$

Elvis	sings	a	song
PN	Verb	Det	?

Markov Assumption 2

Every word depends just on its tag:

$$P(W_i | W_1, ..., W_{i-1}, T_1, ..., T_i) = P(W_i | T_i)$$

Markov Assumption 2

Every word depends just on its tag:

$$P(W_i|W_1, ..., W_{i-1}, T_1, ..., T_i) = P(W_i|T_i)$$

The probability that the 4th word is “song” depends just on the tag of that word:

$$P(song|Elvis, sings, a, PN, V, D, N) = P(song|N)$$

Markov Assumption 2

Every word depends just on its tag:

$$P(W_i|W_1, ..., W_{i-1}, T_1, ..., T_i) = P(W_i|T_i)$$

The probability that the 4th word is “song” depends just on the tag of that word:

$$P(song|Elvis, sings, a, PN, V, D, N) = P(song|N)$$

Elvis	sings	a	?
PN	Verb	Det	Noun

Markov Assumption 2

Every word depends just on its tag:

$$P(W_i|W_1, ..., W_{i-1}, T_1, ..., T_i) = P(W_i|T_i)$$

The probability that the 4th word is "song" depends just on the tag of that word:

$$P(song|Elvis, sings, a, PN, V, D, N) = P(song|N)$$

Elvis	sings	a	?
PN	Verb	Det	Noun

Homogeneity Assumption 1

The tag probabilities are the same at all positions

$$P(T_i|T_{i-1}) = P(T_k|T_{k-1}) \forall i, k$$

Homogeneity Assumption 1

The tag probabilities are the same at all positions

$$P(T_i|T_{i-1}) = P(T_k|T_{k-1}) \forall i, k$$

The probability that a Det is followed by a Noun
is the same at position 7 and 2:

$$P(T_7 = Noun|T_6 = Det) = P(T_2 = Noun|T_1 = Det)$$

Homogeneity Assumption 1

The tag probabilities are the same at all positions

$$P(T_i|T_{i-1}) = P(T_k|T_{k-1}) \forall i, k$$

The probability that a Det is followed by a Noun is the same at position 7 and 2:

$$P(T_7 = Noun|T_6 = Det) = P(T_2 = Noun|T_1 = Det)$$

Let's denote this probability by

$$P(Noun|Det)$$ "Transition probability"

$$P(s|t) := P(T_i = s|T_{i-1} = t)(for any i)$$

Homogeneity Assumption 2

The word probabilities are the same at all positions

$$P(W_i|T_i) = P(W_k|T_k) \forall i, k$$

Homogeneity Assumption 2

The word probabilities are the same at all positions

$$P(W_i|T_i) = P(W_k|T_k) \forall i, k$$

The probability that a PN is “Elvis”

is the same at position 7 and 2:

$$P(W_7 = Elvis|T_7 = PN) = P(W_2 = Elvis|T_2 = PN) = 80\%$$

Homogeneity Assumption 2

The word probabilities are the same at all positions

$$P(W_i|T_i) = P(W_k|T_k) \forall i, k$$

The probability that a PN is "Elvis"

is the same at position 7 and 2:

$$P(W_7 = Elvis|T_7 = PN) = P(W_2 = Elvis|T_2 = PN) = 80\%$$

Let's denote this probability by

$P(Elvis|PN)$ ← "Emission probability"

$$P(w|t) := P(W_i = w|T_i = t)(foranyi)$$

Def: HMM

A (homogeneous) Hidden Markov Model (also: HMM) is a sequence of random variables, such that

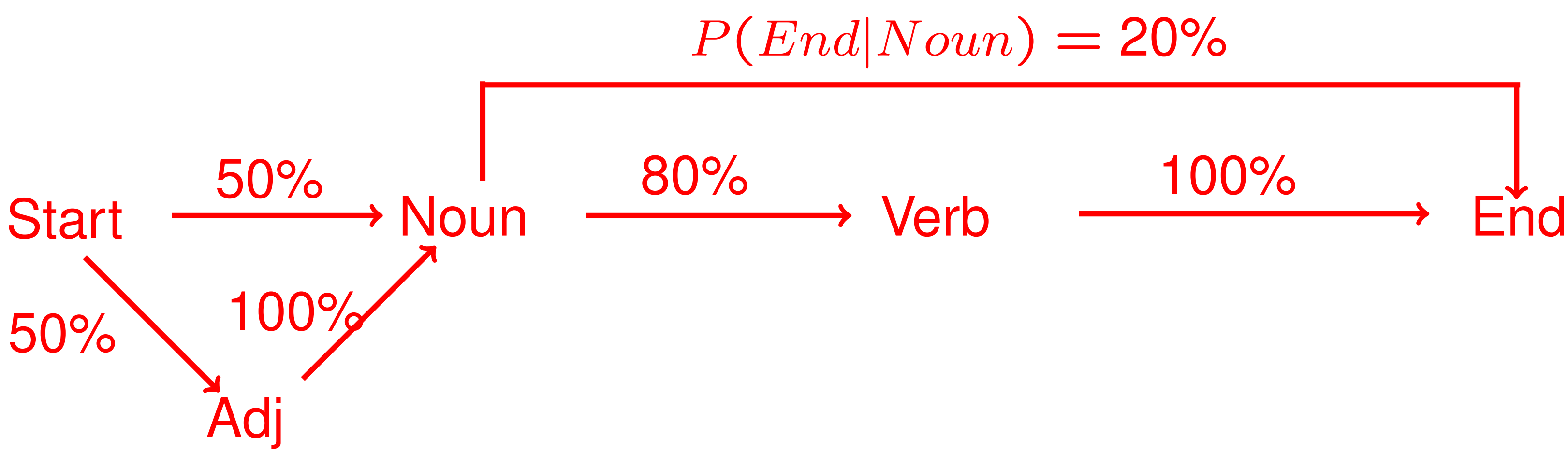
$$P(\underbrace{w_1, ..., w_n}_{\text{Words of the sentence}}, \underbrace{t_1, ..., t_n}_{\text{POS-tags}}) = \prod_i \underbrace{P(w_i|t_i)}_{\text{Emission probabilities}} \times \underbrace{P(t_i|t_{i-1})}_{\text{Transition probabilities}}$$

... with $t_0 = Start$

HMMs as graphs

$$P(w_1, ..., w_n, t_1, ..., t_n) = \prod_i P(w_i|t_i) \times P(t_i|t_{i-1})$$

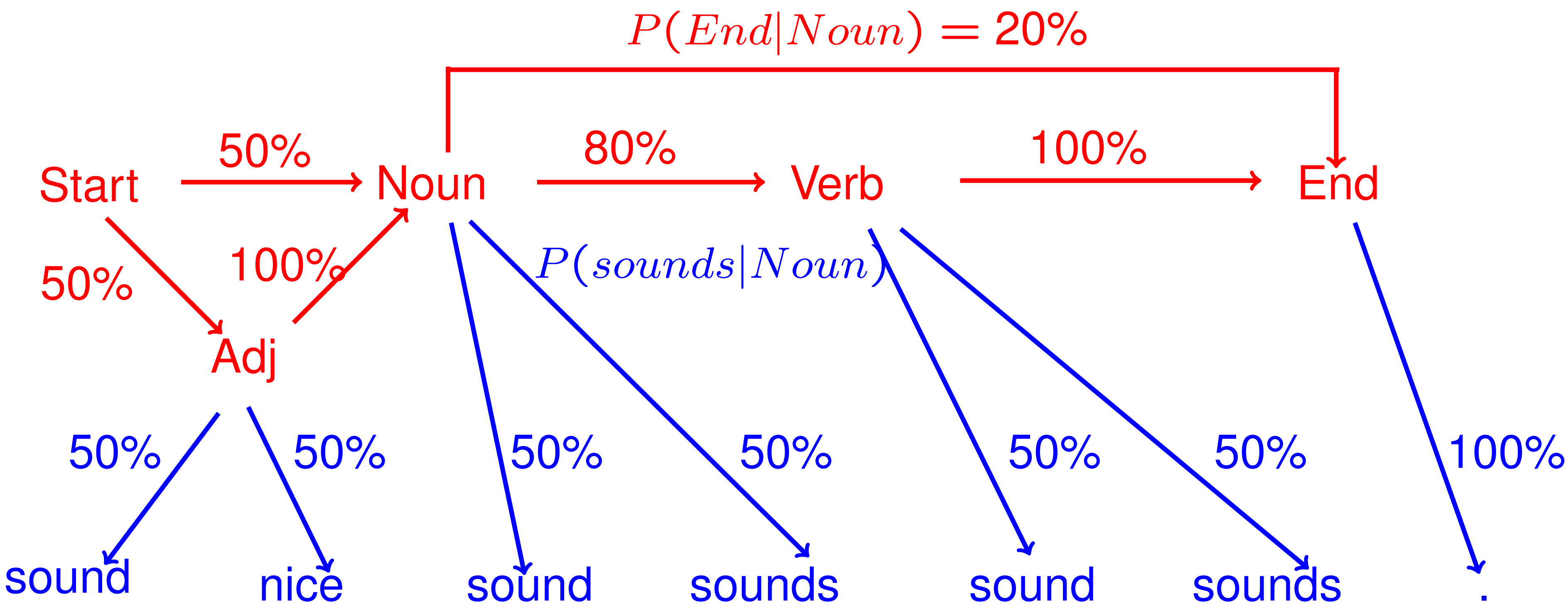
Transition probabilities



HMMs as graphs

$$P(w_1, ..., w_n, t_1, ..., t_n) = \prod_i P(w_i|t_i) \times P(t_i|t_{i-1})$$

Emission probabilities

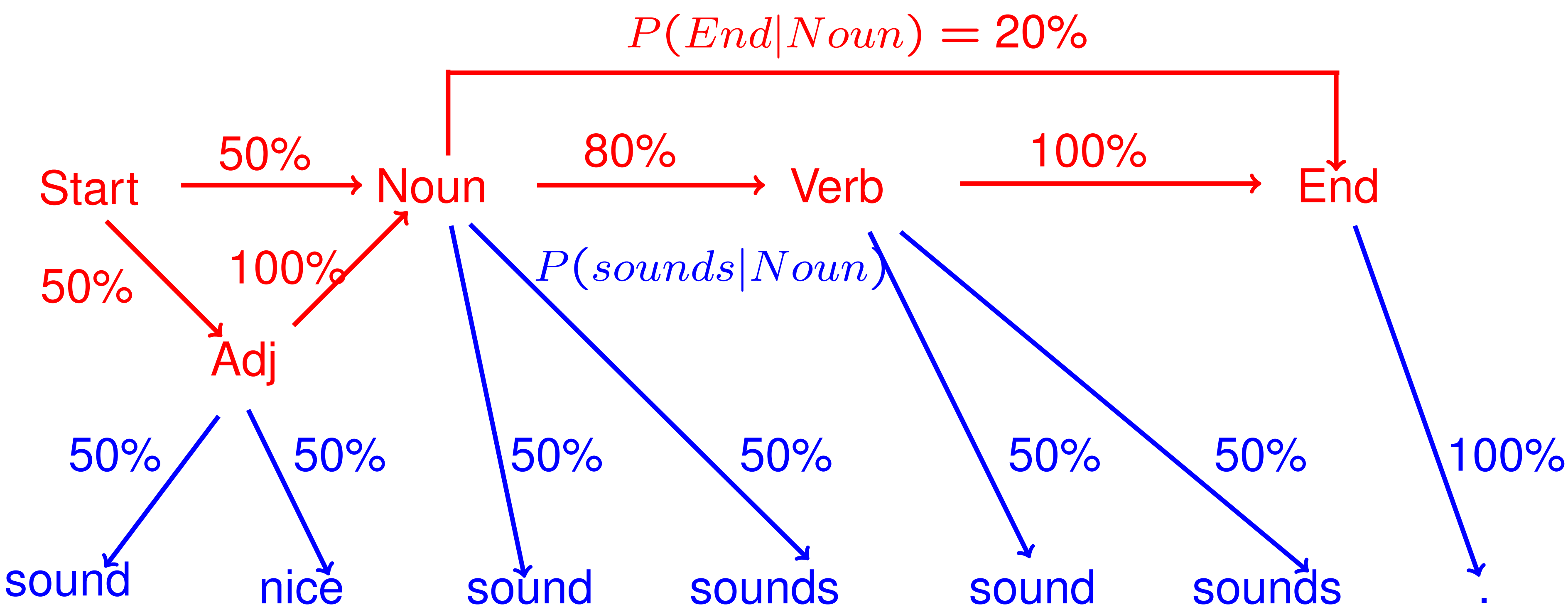


HMMs as graphs

$$P(w_1, ..., w_n, t_1, ..., t_n) = \prod_i P(w_i|t_i) \times P(t_i|t_{i-1})$$

P(nice, sounds, ., Adj, Noun, End) =

50% * 50% * 100% * 50% * 20% * 100% = 2.5%



HMM questions

$$P(w_1, ..., w_n, t_1, ..., t_n) = \prod_i P(w_i|t_i) \times P(t_i|t_{i-1})$$

Main questions:

- Given $P(t_i|t_j)$ and $P(w_i|t_j)$,
  what is the probability of a sentence with tags

- Given $P(t_i|t_j)$ and $P(w_i|t_j)$,
  what is the most likely sequence of T_i that generated a sentence

- What are the $P(t_i|t_j)$, $P(w_i|t_j)$

Overview

Introduction to Probabilities

	Chains	Complex dependencies and/or feature functions
only visible variables	Markov Chains	$\boxed{\text{Markov Random Fields}}$
visible and invisible variables	Hidden Markov Models	Conditional Random Fields

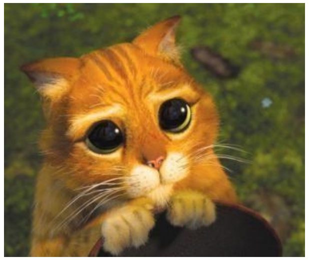
Def: Markov Random Field

A Markov Random Field (MRF) is a set of random variables that satisfies:

$$P(X_i|X_1, .., X_{i-1}, X_{i+1}, ..., X_n) = P(X_i|N(X_i))$$

where $N(X_i)$ is the set of random variables that are neighbors of X_i.

That is: The probability of X_i taking a certain value given the values of the other variables depends only on the values of the variables in the neighborhood of X_i.



Example: Markov Random Field

world	X1	X2	X3	X4	X5	probability
w1:	Shrek	shouts	ferociously	with	pleasure	P(w1)=0.1
w2:	Shrek	shouts	ferociously	ϵ	ϵ	P(w2)=0.1
w3:	Shrek	purrs	ϵ	with	pleasure	P(w3)=0.02
w4:	Shrek	purrs	ϵ	ϵ	ϵ	P(w4)=0.02
w5:	Puss	shouts	ferociously	with	pleasure	P(w5)=0.01
...	...					...

X1, X2, and X3 depend on each other
(shouting is always ferociously, purring is never ferociously,
shouting is more likely to be by Shrek)

X4 and X5 depend on each other
(either both are the empty string, or X4=with and X5=pleasure)

(X1,X2,X3) and (X4,X5) are independent

Example: Markov Random Field

world	X1	X2	X3	X4	X5	probability
w1:	Shrek	shouts	ferociously	with	pleasure	P(w1)=0.1
w2:	Shrek	shouts	ferociously	ϵ	ϵ	P(w2)=0.1
w3:	Shrek	purrs	ϵ	with	pleasure	P(w3)=0.02
w4:	Shrek	purrs	ϵ	ϵ	ϵ	P(w4)=0.02
w5:	Puss	shouts	ferociously	with	pleasure	P(w5)=0.01
...	...					...

Neighbor sets:

$$N(X_1) = \{X_2, X_3\}$$

$$N(X_2) = \{X_1, X_3\}$$

$$N(X_3) = \{X_1, X_2\}$$

$$N(X_4) = \{X_5\}$$

$$N(X_5) = \{X_4\}$$

Def: MRF graph

$$P(X_i|X_1, .., X_{i-1}, X_{i+1}, ..., X_n) = P(X_i|N(X_i))$$

We depict the neighborhood as an undirected graph,
where the nodes are variables, and $X_i - X_j$ is an edge if $X_i \in N(X_j)$.

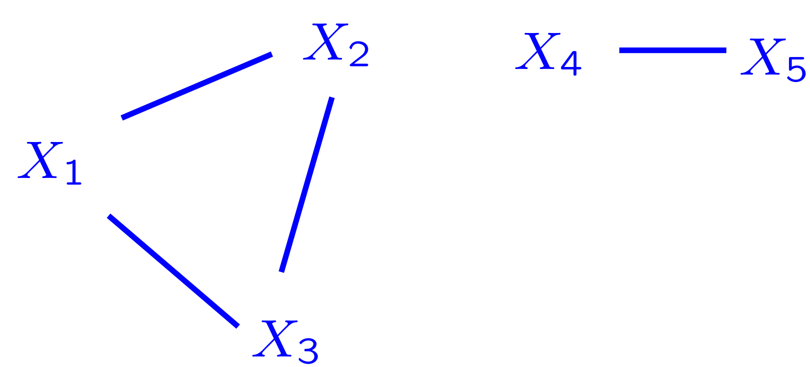
$$N(X_1) = \{X_2, X_3\}$$

$$N(X_2) = \{X_1, X_3\}$$

$$N(X_3) = \{X_1, X_2\}$$

$$N(X_4) = \{X_5\}$$

$$N(X_5) = \{X_4\}$$



MRFs and Markov Chains

Markov Chains:

$$P(X_i|X_1, ..., X_{i-1}) = P(X_i|X_{i-1})$$

The probability of $X = v$ depends only on the value of the predecessor of X.

Markov Random Fields:

$$P(X_i|X_1, .., X_{i-1}, X_{i+1}, ..., X_n) = P(X_i|N(X_i))$$

The probability of $X = v$ depends only on the value of the neighbors of X, but neighborhood is symmetric, and the probability is also conditioned on the “future” X_i.

Syntax: Projection

We define

$$\pi_{i1,...,im}(\vec{x}) := < x_{i1}, ..., x_{im} >$$

Example:

$$\pi_{\{2,5\}}(< a, b, c, d, e, f, g >) = < b, e >$$

Special case: Factorizable MRFs

If all probabilities in an MRF are > 0, then $\exists \phi_i$ such that

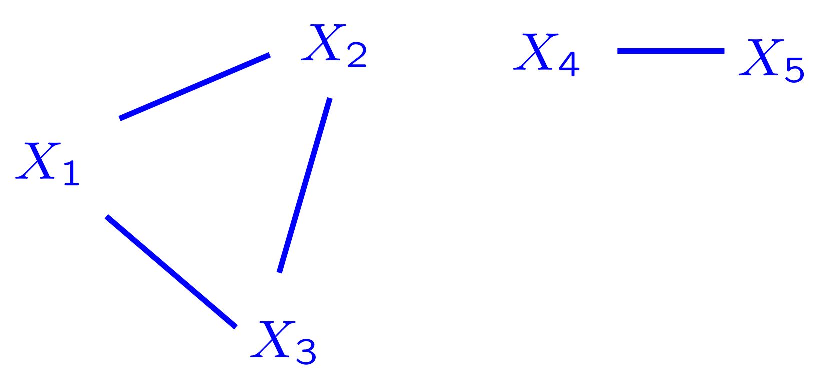
$$P(\vec{X} = \vec{x}) = \prod_i \phi_i(\pi_{C_i}(\vec{x}))$$

Every ϕ_i is a function that takes as input only the variables of the ith clique.

where the C_i are the maximal cliques in the MRF graph.

$C_1 = \{X_1, X_2\}$

$C_2 = \{X_2, X_3, X_4\}$



(We consider only factorizable MRFs)

Wikipedia/Hammersley-Clifford-Theorem

Example: Factorizable MRF

world	X1	X2	X3	X4	X5	probability
w1:	Shrek	shouts	ferociously	with	pleasure	P(w1)=0.1
w2:	Shrek	shouts	ferociously	ϵ	ϵ	P(w2)=0.1
w3:	Shrek	shouts	ferociously	with	ϵ	P(w3)=0.002
w3:	Shrek	shouts	ferociously	ϵ	pleasure	P(w4)=0.002
w3:	Shrek	shouts	ϵ	with	pleasure	P(w4)=0.01
...	...					...

$$C_1 = \{X_1, X_2\} \qquad C_2 = \{X_2, X_3, X_4\}$$

$$\phi_1(x_1, x_2, x_3) = (x_1 = Shrek \wedge x_2 = shouts \wedge x_3 = ferociously) \vee$$
$$(x_1 = Puss \wedge x_2 = purrs \wedge x_3 = \epsilon) ? 0.1 : 0.01$$

$$\phi_2(x_4, x_5) = (x_4 = with \wedge x_5 = pleasure) \vee (x_4 = x_5 = \epsilon) ? 1.0 : 0.02$$

$$P(\vec{X} = \vec{x}) = \phi_1(x_1, x_2, x_3) \times \phi_2(x_4, x_5)$$

(might not sum to 1)

Normalization

To obtain a value in [0,1], we normalize by Z

$$P(\vec{X} = \vec{x}) = \frac{1}{Z} \prod_i \phi_i(\pi_{C_i}(\vec{x}))$$

where Z is simply the sum over the products for all possible sequences of values $\vec{x'}$:

$$Z = \sum_{\vec{x'}} \prod_i \phi_i(\pi_{C_i}(\vec{x'}))$$

Yes, run over all possible sequences $\vec{x'} =< x'_1, ..., x'_n >$ and sum up the products!

This ensures $P(\vec{X} = \vec{x}) \in [0, 1]$.

Special case: MRFs w/ Features

For each clique C_i, we define feature functions

$$f_{i,1}(\pi_{C_i}(\vec{x})) \in R \qquad \text{...} \qquad f_{i,m}(\pi_{C_i}(\vec{x})) \in R$$

These form a vector $\vec{F_i} =< f_{i,1}, ...f_{i,m} >$

$\vec{F_i} =< f_{i,1}, ...f_{i,m} >$

and we define $\vec{F_i}(x) =< f_{i,1}(x), ...f_{i,m}(x) >$.

We define weights for the features:

$$\vec{w_i} \in R^m$$

Then we define the potentials as:

$$\phi_i(\pi_{C_i}(\vec{x})) = exp(\vec{w_i} \times \vec{F_i}(\pi_{C_i}(x)))$$

MRFs with features

With $\phi_i(\pi_{C_i}(\vec{x})) = exp(\vec{w}_i \times F_i(\pi_{\vec{C_i}}(x)))$
we have

$$P(\vec{X} = \vec{x}) = \frac{1}{Z} \prod_i \phi_i(\pi_{C_i}(\vec{x}))$$

$$P(\vec{X} = \vec{x}) = \frac{1}{Z} \prod_i exp(\vec{w}_i \times \vec{F}_i(\pi_{C_i}(\vec{x})))$$

$$P(\vec{X} = \vec{x}) = \frac{1}{Z} exp(\sum_i \vec{w}_i \times \vec{F}_i(\pi_{C_i}(\vec{x})))$$

Def: Log likelihood

The log-likelihood of a MRF is

$$log(P(\vec{X} = \vec{x}))$$

$$= log(\frac{1}{Z}exp(\sum_i \vec{w}_i \times \vec{F}_i(\pi_{C_i}(\vec{x}))))$$

$$= \sum_i \vec{w}_i \times \vec{F}_i(\pi_{C_i}(\vec{x})) - log(Z)$$

$$= -log(Z) + \sum_i \vec{w}_i \times \vec{F}_i(\pi_{C_i}(\vec{x}))$$

↑ "log" ↑ "linear"

"log-linear model"

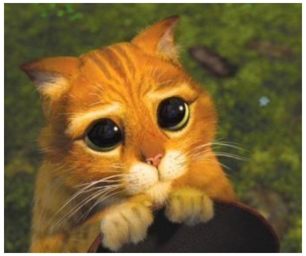
Overview

Introduction to Probabilities

	Chains	Complex dependencies and/or feature functions
only visible variables	Markov Chains	Markov Random Fields
visible and invisible variables	Hidden Markov Models	$\boxed{\text{Conditional Random Fields}}$

Conditional Random Fields

As in Hidden Markov Models, we have visible and hidden random variables:



Visible variables

Hidden variables

	X1	X2	X3	Y1	Y2	Y3
P(	Shrek	eats	Puss	PER	OTH	OTH)= 0.1
P(	Shrek	eats	Puss	PER	LOC	OTH)= 0.01
P(	Shrek	eats	Puss	PER	OTH	LOC)= 0.1
P(	Shrek	eats	Puss	PER	LOC	LOC)= 0.01
P(	Puss	eats	food	OTH	PER	PER)= 0.2
P(	Puss	eats	Puss	LOC	PER	OTH)= 0.012

Conditional Random Fields

Let us look at the conditional probability

$$P(Y1|X1, X2, X3, Y2, Y3)$$

	X1	X2	X3	Y1	Y2	Y3	
P(	Shrek	eats	Puss	PER	OTH	OTH)	= 0.1
P(	Shrek	eats	Puss	PER	LOC	OTH)	= 0.01
P(	Shrek	eats	Puss	PER	OTH	LOC)	= 0.1
P(	Shrek	eats	Puss	PER	LOC	LOC)	= 0.01

Conditional Random Fields

Let us look at the conditional probability

$$P(Y1|X1, X2, X3, Y2, Y3)$$

$$= P(Y1|X1, X2, X3, Y2)$$

$Y3$ does not have an influence!

	X1	X2	X3	Y1	Y2	Y3	
P(	Shrek	eats	Puss	PER	OTH	OTH)	= 0.1
P(	Shrek	eats	Puss	PER	LOC	OTH)	= 0.01
P(	Shrek	eats	Puss	PER	OTH	LOC)	= 0.1
P(	Shrek	eats	Puss	PER	LOC	LOC)	= 0.01

Def: Conditional Random Fields

A set of random variables is a conditional random field (CRF), if

$$P(Y_i|X_1,...,X_n,Y_1,...,Y_{i-1},Y_{i+1},...,Y_n) = P(Y_i|X_1,...,X_n,N(Y_i))$$

where $N(Y_i)$ are the neighbors of Y_i.

$$N_1 = \{Y2\}$$

	X1	X2	X3	Y1	Y2	Y3	
P(	Shrek	eats	Puss	PER	OTH	OTH)	= 0.1
P(	Shrek	eats	Puss	PER	LOC	OTH)	= 0.01
P(	Shrek	eats	Puss	PER	OTH	LOC)	= 0.1
P(	Shrek	eats	Puss	PER	LOC	LOC)	= 0.01

Neighbors

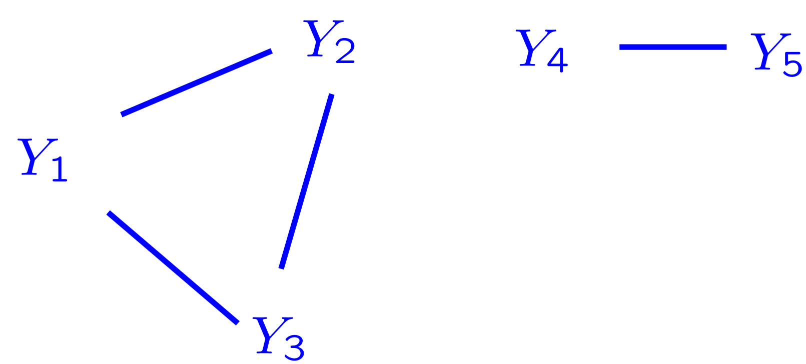
A set of random variables is a conditional random field (CRF), if

$$P(Y_i|X_1, ..., X_n, Y_1, ..., Y_{i-1}, Y_{i+1}, ..., Y_n) = P(Y_i|X_1, ..., X_n, N(Y_i))$$

where $N(Y_i)$ are the neighbors of Y_i. We arrange the neighbors in an undirected graph, where two variables are connected if they are neighbors. C_i are the maximal cliques in this graph.

$C_1 = \{Y_1, Y_2\}$

$C_2 = \{Y_2, Y_3, Y_4\}$



We define $\pi_{i1,...,im}(\vec{x}) :=< x_{i1}, ..., x_{im} >$

Example: $\pi_{C_1}(< a, b, c, d, e >) =< a, b, c >$

Special case: Factorizable CRFs

Strictly positive CRFs can be "factorized", i.e., there exist functions ϕ_i such that

$$P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}) = \prod_i \phi_i(\pi_{C_i}(\vec{y}), \vec{x})$$

ϕ_i are the potentials. They take as input:

- the entire vector $\vec{x}$
- the values of the Y_j that are in C_i

>relation to MRF

(We consider only factorizable CRFs)

CRFs and MRFs

$$P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}) = \prod_i \phi_i(\pi_{C_i}(\vec{y}), \vec{x})$$

A conditional random field is basically
a Markov Random Field that has
$\vec{x}$ as additional, fixed, inputs.

Special case: Chain CRFs

A Chain CRF is a CRF where the neighborhood graph is a chain.

$$Y_1 - Y_2 - Y_3 - Y_4$$

Then, the cliques have only 2 elements:

$$C_i = \{Y_i, Y_{i-1}\}$$

$$for 1 < i \leq n$$

The i-th clique is just Y_i with the preceding Y

(We consider only chain CRFs)

Syntax: Chain CRFs

In a chain CRF, we have:

$$P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}) = \frac{1}{Z} \prod_i \phi_i(\pi_{C_i}(\vec{y}), \vec{x})$$

$$\Big\downarrow \; C_i = \{Y_i, Y_{i-1}\}$$

$$P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}) = \frac{1}{Z} \prod_i \phi_i(y_i, y_{i-1}, \vec{x})$$

We already know the projection of the i-th clique, it's y_i, y_{i-1}.

Special case: Identical potentials

In a CRF with identical potentials, all ϕ_i are the same, but have i as input:

$$P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}) = \frac{1}{Z} \prod_i \phi_i(y_i, y_{i-1}, \vec{x})$$

ϕ has to know i to know the position in $\vec{x}$.

$$P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}) = \frac{1}{Z} \prod_i \phi(y_i, y_{i-1}, \vec{x}, i)$$

(We consider only CRFs with identical potentials)

Special case: CRFs with Features

We define feature functions

$$f_1(y_i, y_{i-1}, \vec{x}, i) \in R \qquad \text{...} \qquad f_m(y_i, y_{i-1}, \vec{x}, i) \in R$$

These form a vector $\vec{F} =< f_1, ...f_m >$,
and we define $\vec{F}(x) =< f_1(x), ...f_m(x) >$.

We define weights for the features:

$$\vec{w} \in R^m$$

Then we define the potential as:

$$\phi(y_i, y_{i-1}, \vec{x}, i) = exp(\vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i))$$

(We consider only CRFs with features)

CRFs with Features

We have

$$P(\vec{Y} = \vec{y}|\vec{X} = \vec{x}) = \frac{1}{Z}\prod_i \phi(y_i, y_{i-1}, \vec{x}, i)$$

and

$$\phi(y_i, y_{i-1}, \vec{x}, i) = exp(\vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i))$$

which yields

$$P(\vec{Y} = \vec{y}|\vec{X} = \vec{x}) = \frac{1}{Z}\prod_i exp(\vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i))$$

$$= \frac{1}{Z}exp(\sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i))$$

Def: CRF log likelihood

The log-likelihood of a CRF is

$$log(P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}))$$

$$= log(\frac{1}{Z} exp(\sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i)))$$

$$= \sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i) - log(Z)$$

Maximizing the CRF likelihood

We want to compute the best Y for X:

$$Y^* = argmax_Y P(\vec{Y} = \vec{y} | \vec{X} = \vec{x})$$

log is monotonic

$$= argmax_Y log(P(\vec{Y} = \vec{y} | \vec{X} = \vec{x}))$$

use log-likelihood from previous slide

$$= argmax_Y \sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i) - log(Z)$$

Z does not depend on Y

$$= argmax_Y \sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i)$$

Easy!

Reminder: Statistical NEA

Statistical NEA uses the following notations:

- a corpus $X =< x_1, ..., x_m >$
- class labels $Y =< y_1, ..., y_m >$
- features $F =< f_1, ..., f_n >$
- weights $W =< w_1, ..., w_n >$

Statistical NEA learns the weights W on a manually annotated training corpus (X, Y), as follows:

$$W = argmax_{W'} log(Pr(Y|X, W'))$$

Given a new corpus X', it computes the annotations Y' as

$$Y' = argmax_Y \sum_i W \times F(X', i, y_i)$$

This is the CRF formula, just that
we considered no dependencies, i.e., $C_i = \{Y_i\}$

Maximizing the CRF likelihood

We want to compute the best Y for X:

$$Y^* = argmax_Y P(\vec{Y} = \vec{y}|\vec{X} = \vec{x})$$

P is factorized

$$= argmax_Y \frac{1}{Z} \prod_i exp(\vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i))$$

$$e^a \times e^b = e^{a+b}$$

$$= argmax_Y \frac{1}{Z} exp(\sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i))$$

log is monotonic

$$= argmax_Y log(\frac{1}{Z} exp(\sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i)))$$

Maximizing the CRF likelihood

$$Y^* = argmax_Y log(\frac{1}{Z} exp(\sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i)))$$

$$log(a \times b) = log(a) + log(b)$$

$$= argmax_Y \sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i) - log(Z)$$

Z does not depend on Y

$$= argmax_Y \sum_i \vec{w} \times \vec{F}(y_i, y_{i-1}, \vec{x}, i)$$

If we consider only
singleton cliques, $C_i = \{Y_i\}$.

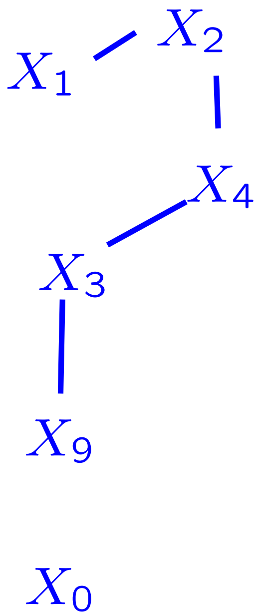
$$= argmax_Y \sum_i \vec{w} \times \vec{F}(y_i, \vec{x}, i)$$

Summary: Prob's, MRFs and CRFs

$$P(w_1) + P(w_2)$$

$$''X_1 = Shrek'' :=$$

$$P(\vec{X} = \vec{x}) = \frac{1}{Z} \prod_i \phi_i(\pi_{C_i}(\vec{x}))$$



$$P(X_1 = Shrek) = P(\{w_1, w_2\})$$

$$P(Y_1 | \vec{X}, Y_2, ..., Y_n) = P(Y_1 | \vec{X}, N(Y_1))$$

$$P(X_1 | X_2, ..., X_n) = P(X_1 | N(X_1))$$

$$\{w | X_1(w) = Shrek\}$$

(just kidding)

Summary: Prob's, MRFs and CRFs

- An event is a set of possible worlds

$$X_1 = Shrek := \{w|X_1(w) = Shrek\} = \{w_1, w_2\}$$

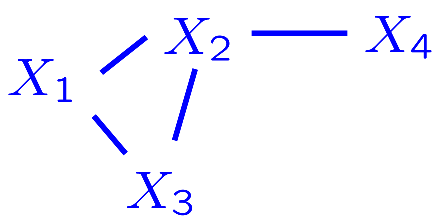
- Probabilities are defined on events

$$P(X_1 = Shrek) = P(\{w_1, w_2\}) = P(w_1) + P(w_2)$$

- MRFs model limited dependencies

$$P(X_1|X_2, ..., X_n) = P(X_1|N(X_1))$$

$$P(\vec{X} = \vec{x}) = \frac{1}{Z} \prod_i \phi_i(\pi_{C_i}(\vec{x}))$$



- CRFs are MRFs with hidden variables

$$P(Y_1|\vec{X}, Y_2, ..., Y_n) = P(Y_1|\vec{X}, N(Y_1))$$

References

Elkan: Log-linear models and conditional random fields

Collins: Log-Linear Models

Lafferty et al: Conditional Random Fields

Sunita Sarawagi: Information Extraction