Language Models Review: 1-28

- Why are language models (LMs) useful?
- Maximum Likelihood Estimation for Binomials
- Idea of Chain Rule, Markov assumptions
- Why is word sparsity an issue?
- Further interest: Leplace Smoothing, Good-Turing Smoothing, LMs in topic modeling.

Disjoint Sets vs. Independent Events

Independence: … iff P(A,B) = P(A)P(B)

Disjoint Sets: If two events, A and B, come from disjoint sets, then
P(A,B) = 0

Disjoint Sets vs. Independent Events

Independence: … iff P(A,B) = P(A)P(B)

Disjoint Sets: If two events, A and B, come from disjoint sets, then
P(A,B) = 0

Does independence imply disjoint?

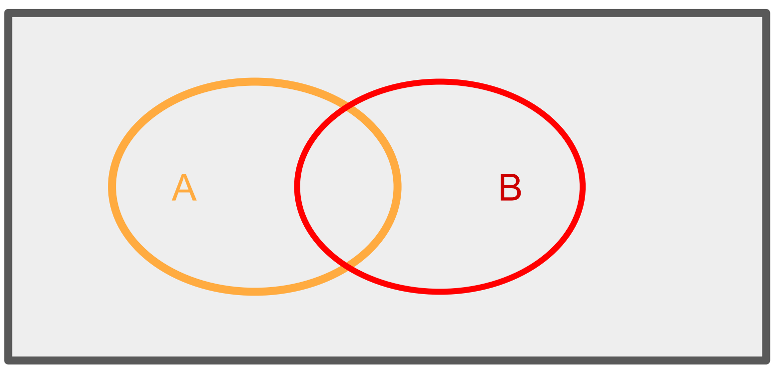
Disjoint Sets vs. Independent Events

Independence: … iff P(A,B) = P(A)P(B)

Disjoint Sets: If two events, A and B, come from disjoint sets, then
P(A,B) = 0

Does independence imply disjoint? No

Proof: A counterexample: A: first coin flip is heads, B: second coin flip is heads;
P(A)P(B) = P(A,B), but .25 = P(A, B) =/= 0



Disjoint Sets vs. Independent Events

Independence: … iff P(A,B) = P(A)P(B)

Disjoint Sets: If two events, A and B, come from disjoint sets, then

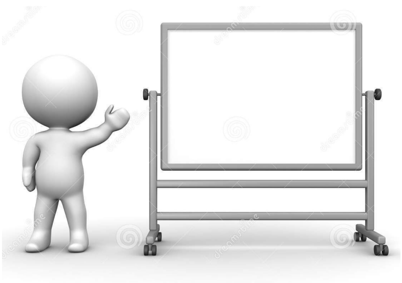
P(A,B) = 0

Does independence imply disjoint? No

Proof: A counterexample: A: first coin flip is heads, B: second coin flip is heads;

P(A)P(B) = P(A,B), but .25 = P(A, B) =/= 0

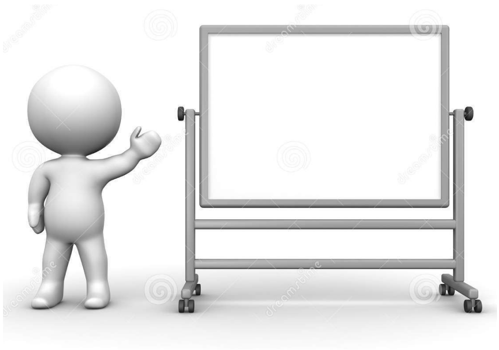
Does disjoint imply independence?



Tools for Decomposing Probabilities

Whiteboard Time!

- Table
- Tree



Examples:

- urn with 3 balls (with and without replacement)
- conversation lengths
- championship bracket

Probabilities over >2 events...

Independence:

$A_1, A_2, \ldots, A_n$ are independent iff $P(A_1, A_2, \ldots, A_n) = \prod P(A_i)$

Probabilities over >2 events...

Independence:

$A_1, A_2, \ldots, A_n$ are independent iff $P(A_1, A_2, \ldots, A_n) = \prod P(A_i)$

Conditional Probability:

$P(A_1, A_2, \ldots, A_{n-1} \mid A_n) = P(A_1, A_2, \ldots, A_{n-1}, A_n) / P(A_n)$

$$P(A_1, A_2, \ldots, A_{m-1} \mid A_m, A_{m+1}, \ldots, A_n) = P(A_1, A_2, \ldots, A_{m-1}, A_m, A_{m+1}, \ldots, A_n) / P(A_m, A_{m+1}, \ldots, A_n)$$

(just think of multiple events happening as a single event)

Conditional Independence

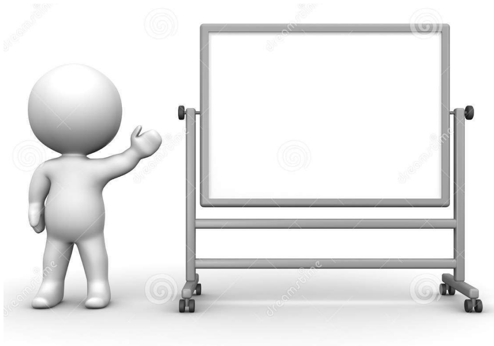
A and B are conditionally independent, given C, IFF

$P(A, B \mid C) = P(A|C)P(B|C)$

Equivalently, $P(A|B,C) = P(A|C)$

Interpretation: *Once we know C, B doesn't tell us anything useful about A.*

Example: Championship bracket



Bayes Theorem - Lite

GOAL: Relate $P(A|B)$ to $P(B|A)$

Let's try:

Bayes Theorem - Lite

GOAL: Relate $P(A|B)$ to $P(B|A)$

Let's try:

(1) $P(A|B) = P(A,B) / P(B)$, def. of conditional probability

(2) $P(B|A) = P(B,A) / P(A) = P(A,B) / P(A)$, def. of conf. prob; sym of set union

Bayes Theorem - Lite

GOAL: Relate $P(A|B)$ to $P(B|A)$

Let's try:

(1) $P(A|B) = P(A,B) / P(B)$, def. of conditional probability

(2) $P(B|A) = P(B,A) / P(A) = P(A,B) / P(A)$, def. of conf. prob; sym of set union

(3) $P(A,B) = P(B|A)P(A)$, algebra on (2) ← known as "Multiplication Rule"

Bayes Theorem - Lite

GOAL: Relate $P(A|B)$ to $P(B|A)$

Let's try:

(1) $P(A|B) = P(A,B) / P(B)$, def. of conditional probability

(2) $P(B|A) = P(B,A) / P(A) = P(A,B) / P(A)$, def. of conf. prob; sym of set union

(3) $P(A,B) = P(B|A)P(A)$, algebra on (2) ← known as "Multiplication Rule"

(4) $P(A|B) = P(B|A)P(A) / P(B)$, Substitute $P(A,B)$ from (3) into (1)

Bayes Theorem - Lite

GOAL: Relate $P(A|B)$ to $P(B|A)$

Let's try:

(1) $P(A|B) = P(A,B) / P(B)$, def. of conditional probability

(2) $P(B|A) = P(B,A) / P(A) = P(A,B) / P(A)$, def. of conf. prob; sym of set union

(3) $P(A,B) = P(B|A)P(A)$, algebra on (2) ← known as "Multiplication Rule"

(4) $\boxed{P(A|B) = P(B|A)P(A) / P(B)}$, Substitute P(A,B) from (3) into (1)

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,

for all i = 1 ... k, where $A_1 \ldots A_k$ **partition** Ω

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,

for all i = 1 ... k, where $A_1 \ldots A_k$ **partition** Ω

partition: $P(A_1 \cup A_2 \ldots \cup A_k) = \Omega$

$P(A_i, A_j) = 0$, for all $i \neq j$

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,
for all i = 1 ... k, where $A_1 \dots A_k$ **partition** Ω

partition: $P(A_1 \cup A_2 \dots \cup A_k) = \Omega$
$P(A_i, A_j) = 0$, for all $i \neq j$

law of total probability: If $A_1 \dots A_k$ **partition** Ω,
then for any event, B

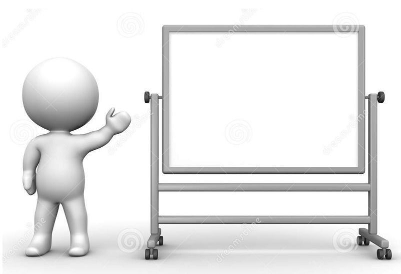
$$\mathrm{P}(B) = \sum_{i=1}^{k} \mathrm{P}(B|A_i)\mathrm{P}(A_i)$$

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,
for all i = 1 ... k, where $A_1 \ldots A_k$ **partition** Ω

partition: $P(A_1 \cup A_2 \ldots \cup A_k) = \Omega$
$P(A_i, A_j) = 0$, for all i ≠ j

law of total probability: If $A_1 \ldots A_k$ **partition** Ω,
then for any event, B



$$P(B) = \sum_{i=1}^{k} P(B|A_i)P(A_i)$$

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,

for all $i = 1 \ldots k$, where $A_1 \ldots A_k$ **partition** Ω

Let's try:

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,

for all i = 1 ... k, where A_1 ... A_k **partition Ω**

Let's try:

(1) $P(A_i|B) = P(A_i,B) / P(B)$

(2) $P(A_i,B) / P(B) = P(B|A_i)\, P(A_i) / P(B)$, by multiplication rule

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,

for all i = 1 ... k, where $A_1 \ldots A_k$ **partition** Ω

Let's try:

(1) $P(A_i|B) = P(A_i,B) / P(B)$

(2) $P(A_i,B) / P(B) = P(B|A_i)\, P(A_i) / P(B)$, by multiplication rule

but in practice, we might not know P(B)

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,

for all i = 1 ... k, where A_1 ... A_k **partition Ω**

Let's try:

(1) $P(A_i|B) = P(A_i,B) / P(B)$

(2) $P(A_i,B) / P(B) = P(B|A_i)\, P(A_i) / P(B)$, by multiplication rule
but in practice, we might not know P(B)

(3) $P(B|A_i)\, P(A_i) / P(B) = P(B|A_i)\, P(A_i) / \left(\sum_{i=1}^{k} P(B|A_i)P(A_i) \right)$, by law of total probability

Law of Total Probability and Bayes Theorem

GOAL: Relate $P(A_i|B)$ to $P(B|A_i)$,

for all i = 1 ... k, where A_1 ... A_k **partition Ω**

Let's try:

(1) $P(A_i|B) = P(A_i,B) / P(B)$

(2) $P(A_i,B) / P(B) = P(B|A_i)\ P(A_i) / P(B)$, by multiplication rule

but in practice, we might not know P(B)

(3) $P(B|A_i)\ P(A_i) / P(B) = P(B|A_i)\ P(A_i) / \left(\sum_{i=1}^{k} P(B|A_i)P(A_i) \right)$, by law of total probability

Thus, $\boxed{P(A_i|B) = P(B|A_i)\ P(A_i) / \left(\sum_{i=1}^{k} P(B|A_i)P(A_i) \right)}$

Probability Theory Review: 2-2

- Conditional Independence
- How to derive Bayes Theorem based
- Law of Total Probability
- Bayes Theorem in Practice