# Lecture 11
# Bayesian Inference

## CS 180 – Intelligent Systems

Dr. Victor Chen

# Review on Probability Theory

# Random variables

- We use ***random variables*** to describe uncertain state. ***Random variables*** take on values in a *domain.*

  - **R**: *Is it raining?*
  - **R** in {True, False}

  - **W**: *What's the weather?*
  - **W** in {Sunny, Cloudy, Rainy, Snow}

  - **D**: *What is the outcome of rolling two dice?*
  - **D** in {(1,1), (1,2), … (6,6)}

  - **S**: *What is the speed of my car (in MPH)?*
  - **S** in [0, 200]

# Events

***Event:*** a complete assignment of *values* to all random variables

E.g., if two Boolean variables *Cavity* and *Toothache*,

Then there are four distinct events:

*Cavity = false* ^*Toothache = false*
*Cavity = false* ^ *Toothache = true*
*Cavity = true* ^ *Toothache = false*
*Cavity = true* ^ *Toothache = true*

# Joint probability

A ***joint probability*** $P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n)$ refers to the probability of an event.

| **Atomic events** | **P** |
|---|---|
| *Cavity = false ^ Toothache = false* | 0.8 |
| *Cavity = false ^ Toothache = true* | 0.1 |
| *Cavity = true ^ Toothache = false* | 0.05 |
| *Cavity = true ^ Toothache = true* | 0.05 |

The joint probabilities of all the events **must sum to 1**

# Marginal (prior) probability

If you sum the joints of all events where X = x, you get the ***marginal (prior) probability*** P(X = x)

$$P(X = x) = P\big((X = x \wedge Y = y_1) \vee \ldots \vee (X = x \wedge Y = y_n)\big)$$

$$= P\big((x, y_1) \vee \ldots \vee (x, y_n)\big) = \sum_{i=1}^{n} P(x, y_i)$$

# Derive marginal from joint

| **P(Cavity, Toothache)** | |
|---|---|
| *Cavity = false ^Toothache = false* | 0.8 |
| *Cavity = false ^ Toothache = true* | 0.1 |
| *Cavity = true ^ Toothache = false* | 0.05 |
| *Cavity = true ^ Toothache = true* | 0.05 |

| **P(Cavity)** | |
|---|---|
| *Cavity = false* | ? |
| *Cavity = true* | ? |

| **P(Toothache)** | |
|---|---|
| *Toothache = false* | ? |
| *Toochache = true* | ? |

# Derive marginal from joint

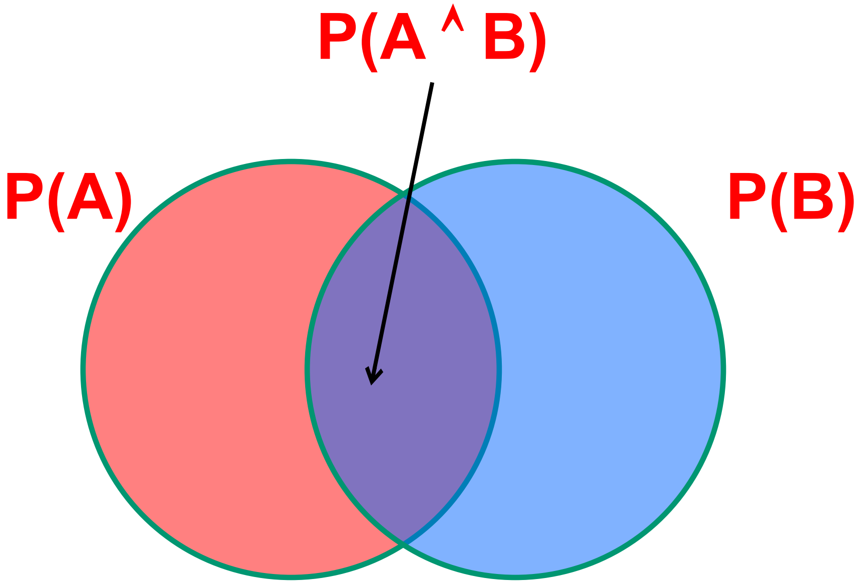
| **P(Cavity, Toothache)** | |
|---|---|
| *Cavity = false ^Toothache = false* | 0.8 |
| *Cavity = false ^ Toothache = true* | 0.1 |
| *Cavity = true ^ Toothache = false* | 0.05 |
| *Cavity = true ^ Toothache = true* | 0.05 |

| **P(Cavity)** | |
|---|---|
| *Cavity = false* | 0.9 |
| *Cavity = true* | 0.1 |

| **P(Toothache)** | |
|---|---|
| *Toothache = false* | 0.85 |
| *Toochache = true* | 0.15 |

# Conditional probability (likelihood)

For any two events A and B, $P(A \mid B) = \dfrac{P(A \wedge B)}{P(B)} = \dfrac{P(A,B)}{P(B)}$



# Derive conditional from joint

| **P(Cavity, Toothache)** | |
|---|---|
| *Cavity = false* ∧ *Toothache = false* | 0.8 |
| *Cavity = false* ∧ *Toothache = true* | 0.1 |
| *Cavity = true* ∧ *Toothache = false* | 0.05 |
| *Cavity = true* ∧ *Toothache = true* | 0.05 |

| **P(Cavity)** | |
|---|---|
| *Cavity = false* | 0.9 |
| *Cavity = true* | 0.1 |

| **P(Toothache)** | |
|---|---|
| *Toothache = false* | 0.85 |
| *Toothache = true* | 0.15 |

What is P(*Cavity = true* | *Toothache = false*)?

*P(Cavity = true* ∧ *Toothache = false)/P(Toothache = false)* = 0.05 / 0.85 = 0.059

What is P(*Cavity = false* | *Toothache = true*)?

*P(Cavity = false* ∧ *Toothache = true)/P(Toothache = true)* = 0.1 / 0.15 = 0.667

# Derive conditional from joint

| **P(Cavity, Toothache)** | |
|---|---|
| *Cavity = false ^Toothache = false* | 0.8 |
| *Cavity = false ^ Toothache = true* | 0.1 |
| *Cavity = true ^ Toothache = false* | 0.05 |
| *Cavity = true ^ Toothache = true* | 0.05 |

| **P(Cavity \| Toothache = true)** | |
|---|---|
| *Cavity = false* | 0.667 |
| *Cavity = true* | 0.333 |

| **P(Cavity\| Toothache = false)** | |
|---|---|
| *Cavity = false* | 0.941 |
| *Cavity = true* | 0.059 |

| **P(Toothache \| Cavity = true)** | |
|---|---|
| *Toothache= false* | 0.5 |
| *Toothache = true* | 0.5 |

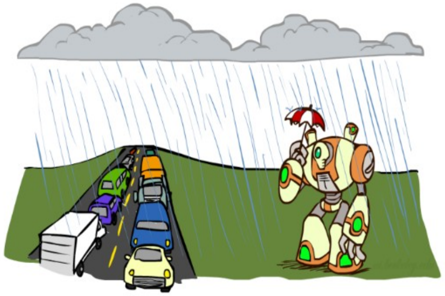
| **P(Toothache \| Cavity = false)** | |
|---|---|
| *Toothache= false* | 0.889 |
| *Toothache = true* | 0.111 |

# Derive joint from conditional

***Chain rule***:

$$P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$$

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{Traffic})$$



# Independence

Two events A and B are *independent* iff

$P(A, B) = P(A)\,P(B)$

- In other words, $P(A \mid B) = P(A)$ or $P(B \mid A) = P(B)$

**Conditional independence**: A and B are *conditionally independent* given C iff

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

- Equivalently:
  $P(A \mid B, C) = P(A \mid C)$ or $P(B \mid A, C) = P(B \mid C)$

# Conditional independence: Example

*Toothache*: if the patient has a toothache

*Cavity*: if the patient has a cavity

*Catch*: if the dentist's probe catches in the cavity

If the patient has a cavity, the probability that the probe catches in it doesn't depend on whether she has a toothache

$$P(\mathit{Catch} \mid \mathit{Toothache}, \mathit{Cavity}) = P(\mathit{Catch} \mid \mathit{Cavity})$$

Therefore, *Catch and Toothache* are conditionally independent given *Cavity*

*Question: are Catch and Toothache independent?*

No since $P(\mathit{Catch}, \mathit{Toothache}) \neq P(\mathit{Catch}\,)\, P(\mathit{Toothache})$

# Use conditional independence to simplify joint calculation

According to the chain rule:

$$P(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})$$

$$= P(\mathit{Cavity})\, P(\mathit{Catch} \mid \mathit{Cavity})\, P(\mathit{Toothache} \mid \mathit{Catch}, \mathit{Cavity})$$

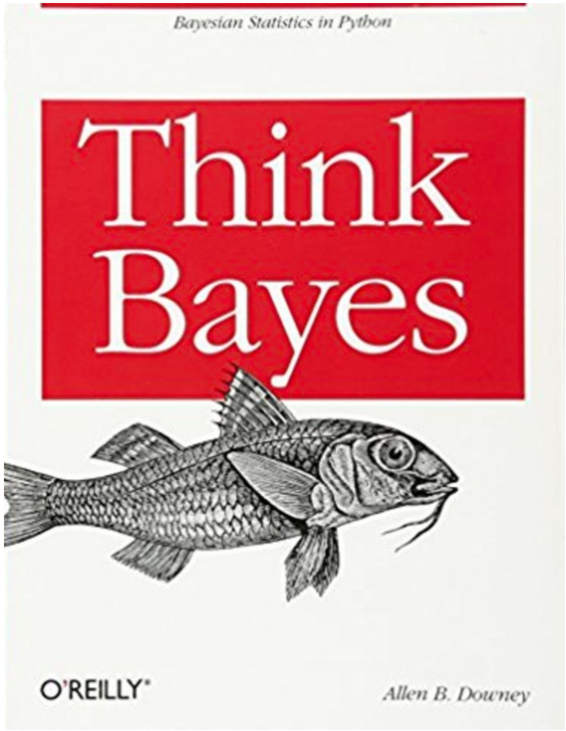
if conditional independence:

$$\boxed{= P(\mathit{Cavity})\, P(\mathit{Catch} \mid \mathit{Cavity})\, P(\mathit{Toothache} \mid \mathit{Cavity})}$$

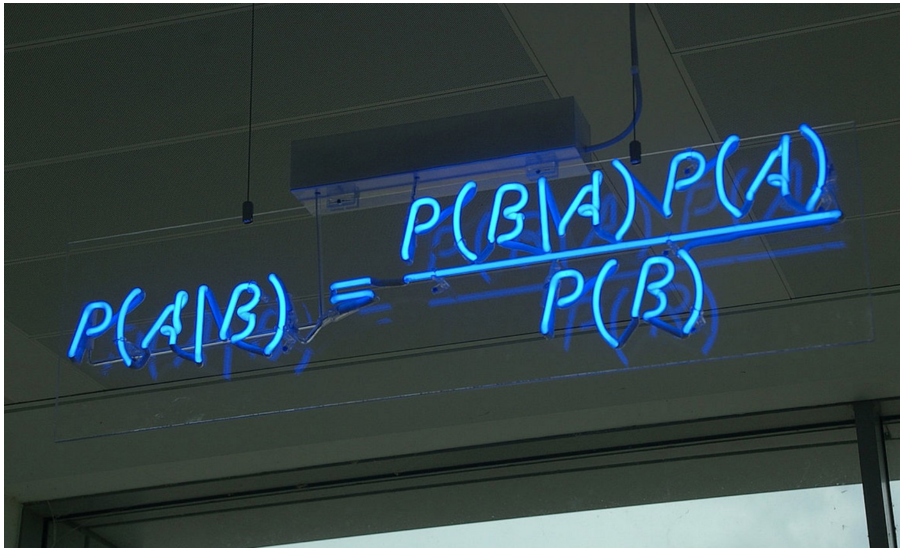
use conditional independence to
estimate the joint in an easy way

# Bayesian inference

A inference tool that uses conditional independence to simplify the calculation of joint probabilities.



# Bayes Rule



$$Posterior = \frac{Likelihood * Prior}{Normalization}$$

Why is this useful?

- *Posterior* is proportional to *likelihood × prior*
- P(A) is the *prior* and P(A|B) is the *posterior*
- P(B|A) is the *likelihood*
- P(B) is the *marginal*
- Theoretical foundation of Bayesian inference

# Bayes Rule example

Marie is getting married tomorrow, at an outdoor ceremony in the desert. In recent years, it has rained only 5 days each year (5/365 = 0.014). Unfortunately, the weatherman has predicted rain for tomorrow. When it actually rains, the weatherman correctly predicts rain 90% of the time. When it doesn't rain, he incorrectly predicts rain 10% of the time. What is the probability that it will rain on Marie's wedding?

$$P(\text{rain} \mid \text{predict}) = \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict})}$$

$$= \frac{P(\text{predict} \mid \text{rain})P(\text{rain})}{P(\text{predict} \mid \text{rain})P(\text{rain}) + P(\text{predict} \mid \neg\text{rain})P(\neg\text{rain})}$$

$$= \frac{0.9 \times 0.014}{0.9 \times 0.014 + 0.1 \times 0.986} = \frac{0.0126}{0.0126 + 0.0986} = 0.111$$

# Bayes rule: Example

1% of women at age 40 who take routine screening test have breast cancer. 80% of women with breast cancer will get positive test result. 9.6% of women without breast cancer will also get positive test result. Suppose a woman in this age group had a positive test result in a routine screening. What is the probability that she actually has breast cancer?

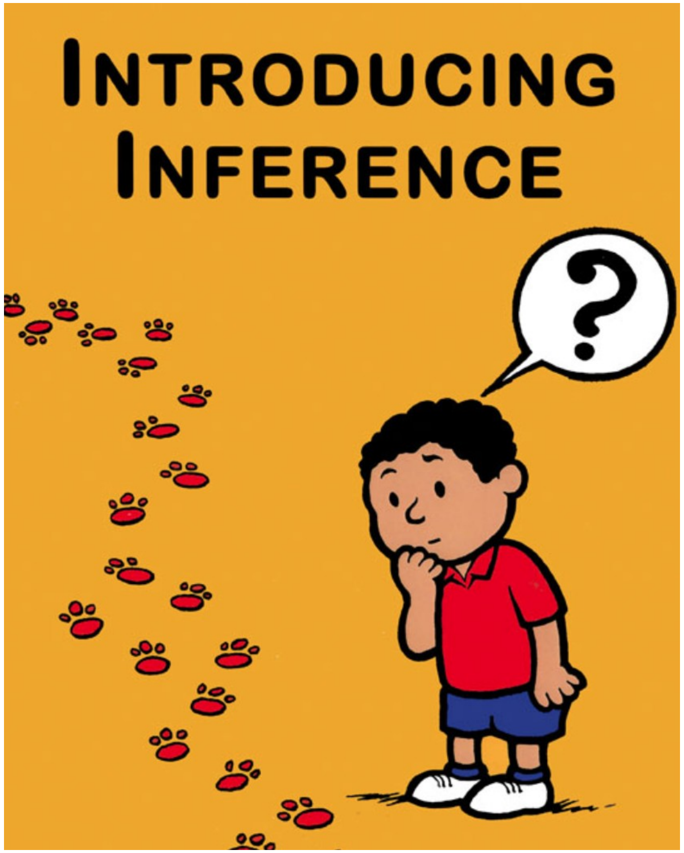
$$P(\text{cancer} \mid \text{positive}) = \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive})}$$

$$= \frac{P(\text{positive} \mid \text{cancer})P(\text{cancer})}{P(\text{positive} \mid \text{cancer})P(\text{cancer}) + P(\text{positive} \mid \neg\text{cancer})P(\neg\text{cancer})}$$

$$= \frac{0.8 \times 0.01}{0.8 \times 0.01 + 0.096 \times 0.99} = \frac{0.008}{0.008 + 0.095} = 0.0776$$

# Bayesian inference

Before we move on to Bayesian inference, let's take on a look at what inference is.



# Inference problem

Given the value of some evidence variable E = e, find the value x of query variable X that **maximize the posterior** probability:

**posterior**

$$\hat{x} = \arg\max_x P(X = x \mid E = e)$$

- Examples:

  e = image features, X = {tiger, monkey, zebra},

# Bayesian inference (apply Bayes rule )

**posterior** **likelihood** **prior**

$$\hat{x} = \arg\max_x P(X = x \mid E = e) = \frac{P(E = e \mid X = x)P(X = x)}{P(E = e)}$$

**marginal**

**Here we reduce the inference problem to the problem of estimating likelihood and priors**

P(E=e) is $= x)P(X = x)$

If we have multiple evidences (features) $E_1, \ldots, E_n$ :

$$\hat{x} = \arg\max_x \; P(E_1 = e_1, \ldots, E_n = e_n \mid X = x) \;\; P(X = x)$$

If we assume that $E_1, \ldots, E_n$ are conditionally independent :

$$\boxed{\hat{x} = \arg\max_x \prod_{i=1}^{n} P(E_i = e_i \mid X = x) \; P(X = x)}$$

# Case study: Text document classification

**Inference:** assign a document to the class with the highest posterior P(class | document)

Question:  What are evidence variable and query variable?

Goal: estimate likelihoods P(document | class)  and priors P(class)

Likelihood: ***bag of words*** model

- The document is a sequence of words $(w_1, \ldots, w_n)$
- The order of the words in the document is not important
- Each word is independent of the others given document class
- Can be computes as:

$$P(document \mid class) = P(w_1, \; \ldots, w_n \mid class) = \prod_{i=1}^{n} P(w_i \mid class)$$

**Product of the likelihoods of individual words**

# Parameter estimation

How do we obtain likelihoods of individual words?

- We need *training data* of labeled documents
- Naïve approach:

$$\text{P(word | class)} = \frac{\text{\# of occurrences of this word in this class}}{\text{total \# of words in this class}}$$

- Any problem?

# Parameter estimation

If training data is not very large, the above method may not work very well.

We need to make sure likelihoods are not zero or too small.

**Solution: Smoothing**:

- Used to make sure likelihoods are not zero or too small.
- **Laplace smoothing:** mixing true likelihood in training data with uninform distribution

$$P(\text{word} \mid \text{class}) = \frac{\#\text{ of occurrences of this word in this class} + 1}{\text{total }\#\text{ of words in class} + V}$$

**(V: total number of unique words)**

# Example (Politics or Sports?)

new doc: **X = “Obama likes basketball”**

**Training set**

**Politics**

“Obama meets Merkel”
“Obama elected again”
“Merkel visits Greece again”

**Sports**

“OSFP European basketball champion”
“Miami NBA basketball champion”
“Greece basketball coach?”

P(p) = 0.5

P(s) = 0.5

**terms**

obama:2, meets:1, merkel:2, elected:1, again:2, visits:1, greece:1

OSFP:1, european:1, basketball:3, champion:2, miami:1, nba:1, greece:1, coach:1

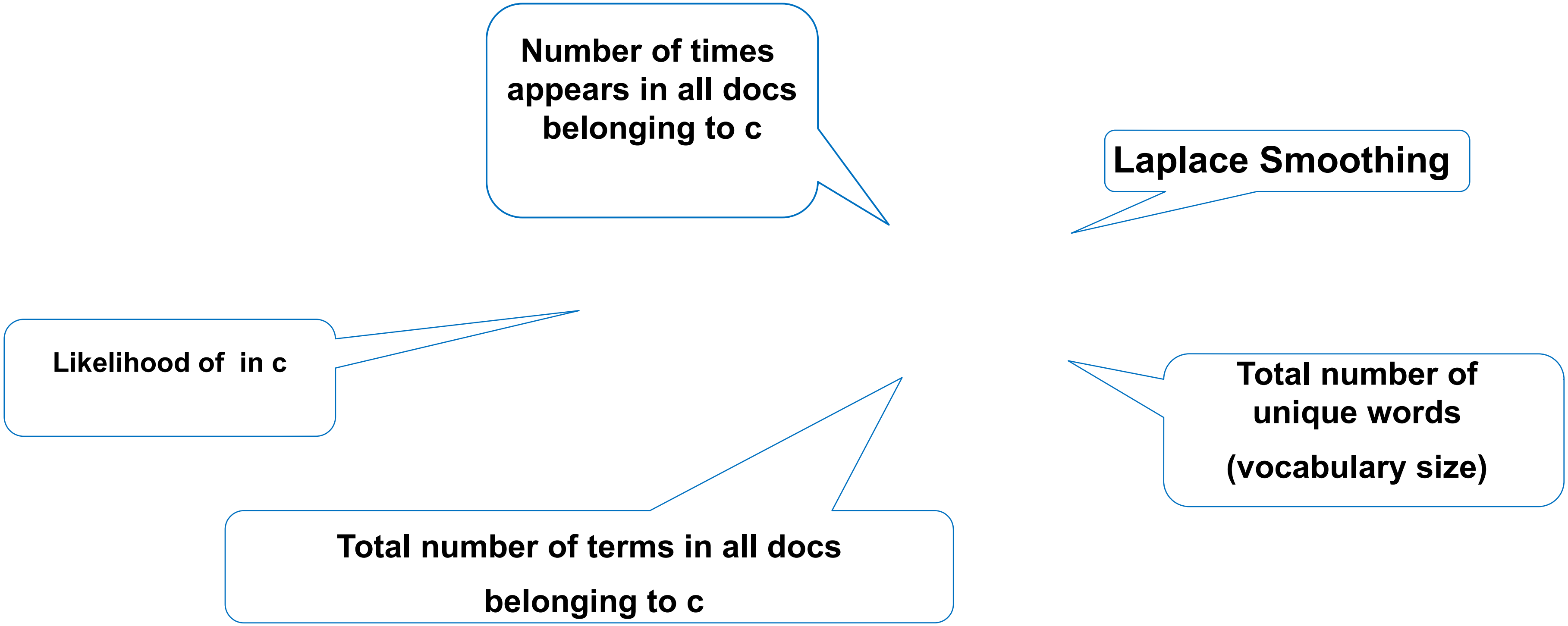
Total # of terms: 10

Total # of terms: 11

Vocabulary
(distinct terms)
size: 14

# Example  (Politics or Sports?)

For a document with k terms ,  the posterior of class  is:



Number of times appears in all docs belonging to c

Laplace Smoothing

Likelihood of  in c

Total number of unique words

(vocabulary size)

Total number of terms in all docs

belonging to c

# Example (Politics or Sports?)

| | Politics | Sports |
|---|---|---|
| **Documents in training** | “Obama meets Merkel”<br>“Obama elected again”<br>“Merkel visits Greece again” | “OSFP European basketball champion”<br>“Miami NBA basketball champion”<br>“Greece basketball coach?” |
| | P(p) = 0.5 | P(s) = 0.5 |
| **terms** | obama:2, meets:1, merkel:2, elected:1, again:2, visits:1, greece:1 | OSFP:1, european:1, basketball:3, champion:2, miami:1, nba:1, greece:1, coach:1 |
| **Vocabulary size: 14** | Total # of terms: : 10 | Total # of terms: 11 |

**new doc:** X = “Obama likes basketball”

$$P(\text{Politics}|X) = P(p)*P(\text{obama}|p)*P(\text{likes}|p)*P(\text{basketball}|p)$$

$$= 0.5 * (2+1)/(10+14) *(0+1)/(10+14) * (0+1)/(10+14) = 0.000108$$

$$P(\text{Sports}|X) = P(s)*P(\text{obama}|s)*P(\text{likes}|s)*P(\text{basketball}|s)$$

$$= 0.5 * (0+1)/(11+14) *(0+1)/(11+14) * (3+1)/(11+14) = 0.000128$$