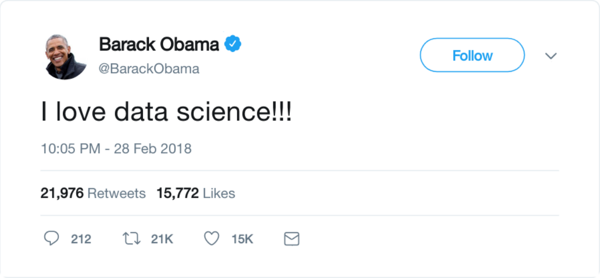


DSC 40A

Lecture 16

Naïve Bayes, pt I

Last Time

- $P(A \mid B)$ = probability of A given that we know B has occurred.

- **Bayes' Theorem**:

$$P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)}$$

- **Independence**: $P(A|B) = P(A)$, or $P(A \cap B) = P(A)\,P(B)$.

- **Conditional Independence**: $P(A \cap B|C) = P(A|C)\,P(B|C)$

Computing Probabilities

Two ways:

1. With math (combinatorics).

2. With data.

Example

You draw a card from a deck of 52 cards. What is the probability that it is red?

$$\frac{|E|}{|\Omega|} = \frac{26}{52} = \frac{1}{2}$$

Example

A person is chosen at random from DSC 40A. What is the probability that they can play piano?

$$\frac{|E|}{|\Omega|} = \frac{\text{\# in class who can play piano}}{\text{\# in class}}$$

Example

A person in the United States is chosen at random. What is the probability that they can play piano?

$$\frac{|E|}{|\Omega|} = \frac{\text{\# in US who can play piano}}{\text{\# in US}} = \frac{?}{\text{\# in US}}$$

Estimating Probabilities

- We can estimate probabilities by randomly **sampling**.
- Example: ask 1000 people if they can play piano.
- If the sample is **representative**:

$$\frac{\text{\# in US who can play piano}}{\text{\# in US}} \approx \frac{\text{\# in sample who can play piano}}{\text{\# in sample}}$$

Example: Survey

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Example: Survey

- What is the probability that someone is married?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

$6/10 = 60\%$

Law of Large Numbers

- As the sample size grows, the estimate becomes more accurate.
- Example:
  - Flip coin $n = 10$ times, 5 repetitions. Proportions of heads:
    $$0.4, 0.3, 0.8, 0.5, 0.5$$
  - Flip coin $n = 1000$ times, 5 repetitions. Proportions of heads:
    $$0.52, 0.50, 0.51, 0.49, 0.48$$

Estimating Joint Probabilities

- To estimate $P(A \cap B)$, count number satisfying both A and B.
- Divide by size of sample.

Example

What is the probability that someone is married and their favorite car brand is Daimler-Benz?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

$2/10 = 20\%$

Estimating Conditional Probabilities

- To estimate $P(A \mid B)$, count number satisfying both A and B.
- Divide by number satisfying B.

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

Example

$$\frac{\# A \cap B}{\# B} = \frac{2}{3}$$

What is the probability that someone is married given that their favorite car brand is Daimler-Benz?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Estimating Conditional Probabilities

- To estimate $P(A \mid B \cap C)$, count number satisfying $A \cap B \cap C$
- Divide by number satisfying $B \cap C$.

$$P(A|B) \approx \frac{\#\, A \text{ and } B}{\#\, B}$$

$$P(A|B \cap C) = \frac{\#\, A \text{ and } B \text{ and } C}{\#\, B \text{ and } C}$$

Example

What is the probability that someone's favorite car brand is Toyota given that they are married and their favorite food is Thai?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

$$\frac{\# A \cap B \cap C}{\# B \cap C} = \frac{1}{3}$$

Example

$$\frac{\# A \cap B \cap C \cap D}{\# B \cap C \cap D} = \frac{0}{0}$$

What is the probability that someone's favorite car brand is Toyota given that they are married and their favorite food is Thai and their favorite subject is physics?

Relationship Status	Favorite Subject	Favorite Car Brand	Favorite Cuisine
Married	English	Nissan	Mexican
Divorced	Physics	Daimler-Benz	Lebanese
Married	Chemistry	Daimler-Benz	Italian
Unmarried	History	Toyota	Chinese
Married	Biology	Ford	Thai
Unmarried	Chemistry	Audi	Lebanese
Unmarried	Biology	Honda	Chinese
Married	Commerce	Ford	Thai
Married	Maths	Daimler-Benz	Mediterranean
Married	Chemistry	Toyota	Thai

Estimating Conditional Probabilities

- We might not have enough data to estimate conditional probabilities with very specific conditions.
- Does **assuming** conditional independence help?
- Example: $P(A \mid C_1 \cap C_2 \cap C_3)$.
- Assume C_1, C_2, C_3 conditionally independent given A. Then:

$$P(A \mid C_1 \cap C_2 \cap C_3) = \frac{P(C_1 \cap C_2 \cap C_3 \mid A) P(A)}{P(C_1 \cap C_2 \cap C_3)}$$

Under conditional indep: $P(C_1 \cap C_2 \cap C_3 \mid A) = P(C_1 \mid A) P(C_2 \mid A) P(C_3 \mid A)$

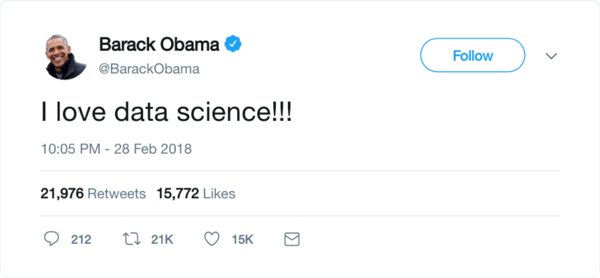
$$= \frac{P(C_1 \mid A) P(C_2 \mid A) P(C_3 \mid A) P(A)}{P(C_1 \cap C_2 \cap C_3)}$$

Sentiment Analysis

- Goal: given a tweet, determine if it is **positive**, **negative**, or **neutral**.

Sentiment Analysis

- Goal: given a tweet, determine if it is **positive**, **negative**, or **neutral**.



Informative Words

- Some words are very informative in sentiment analysis:
  - “love”, “fantastic”, “enjoy”, etc., are **positive**.
  - “hate”, “terrible”, “angry”, etc., are **negative**.

- But “love” doesn’t automatically make tweet positive:

Informative Words

- Some words are very informative in sentiment analysis:
  - “love”, “fantastic”, “enjoy”, etc., are **positive**.
  - “hate”, “terrible”, “angry”, etc., are **negative**.

- But “love” doesn’t automatically make tweet positive:

Sentiment Analysis and Probability

- How likely is it that a tweet containing "love" is **positive**?
- In other words, what is:

$$P(\text{tweet is } \textbf{positive} \mid \text{it contains "love"})?$$

- From the definition:

$$P(\textbf{positive} \mid \text{contains "love"})$$

$$= \frac{\text{\# tweets which are } \textbf{positive} \text{ and contain "love"}}{\text{\# tweets containing "love"}}$$

Estimating Probabilities

- Gathering all tweets ever tweeted is not feasible.

- Instead, we gather a sample and *approximate* these probabilities.

  $P(\textbf{positive} \mid \text{contains “love”})$

  $$= \frac{\text{\# tweets which are } \textbf{positive} \text{ and contain “love”}}{\text{\# tweets containing “love”}}$$

  $$\approx \frac{\text{\# tweets in sample which are } \textbf{positive} \text{ and contain “love”}}{\text{\# tweets in sample containing “love”}}$$

- Law of Large Numbers says: bigger the sample, better the approximation.

Estimating Probabilities

- We sample n tweets at random, label each as **positive**, **negative**, or **neutral** by hand.
- Mark whether each tweet contains “love”.
- The result is a table:

Sentiment	Contains “Love”
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

Estimating Probabilities

Sentiment	Contains "Love"
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

$P(\text{positive} \mid \text{contains "love"})$

$$\approx \frac{\text{\# tweets in sample which are positive and contain "love"}}{\text{\# tweets in sample containing "love"}}$$

$= 2/3$

Using Bayes' Theorem

- We could've used Bayes Theorem, too:

$$P(\textbf{positive} \mid \text{contains "love"})$$

$$= \frac{\overset{2/3}{P(\text{contains "love"} \mid \textbf{positive})} \cdot \overset{3/7}{P(\textbf{positive})}}{\underset{3/7}{P(\text{contains "love"})}} = \frac{2}{3}$$

$$P(\textbf{positive}) \approx 3/7$$

$$P(\text{contains "love"}) \approx 3/7$$

Classification

- Now we are given a tweet we have never seen before; it does not contain “love”.
- We wish to classify its sentiment **automatically**.
- We will compute:

$$P(\textbf{positive} \mid \text{does not contain “love”})$$
$$P(\textbf{negative} \mid \text{does not contain “love”})$$
$$P(\textbf{neutral} \mid \text{does not contain “love”})$$

- The classification is determined by which probability is highest.

Estimating Probabilities

Sentiment	Contains “Love”
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

$P($**positive** $\mid$ does not contain “love”$)$

$\approx$ 1/4

Estimating Probabilities

Sentiment	Contains "Love"
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

$P(\text{negative} \mid \text{does not contain "love"})$

$\approx 2/4$

Estimating Probabilities

Sentiment	Contains "Love"
positive	yes
positive	no
negative	no
negative	no
neutral	yes
neutral	no
positive	yes

P(**neutral** | does not contain "love")

≈ 1/4

Choosing the Most Likely Sentiment

$P($**positive** $\mid$ does not contain "love"$) \approx 1/4$

$P($**negative** $\mid$ does not contain "love"$) \approx 2/4$

$P($**neutral** $\mid$ does not contain "love"$) \approx 1/4$

- Since $P($~~**neutral**~~ negative $\mid$ does not contain "love"$)$ is largest, we assign tweet the sentiment of ~~**neutral**~~ negative.

Better Classifications

- We are making classification based on the presence/absence of one word.

- We'll get better results if we use more words:
  - w_1 = "love";
  - w_2 = "terrible";
  - w_3 = "angry";

- Suppose a tweet contains w_1, doesn't contain w_2, contains w_3. We want to compute:

$$P(\textbf{positive} \mid w_1 = \text{yes} \;\&\; w_2 = \text{no} \;\&\; w_3 = \text{yes})$$
$$P(\textbf{negative} \mid w_1 = \text{yes} \;\&\; w_2 = \text{no} \;\&\; w_3 = \text{yes})$$
$$P(\textbf{neutral} \mid w_1 = \text{yes} \;\&\; w_2 = \text{no} \;\&\; w_3 = \text{yes})$$

A Practical Problem

- Suppose we use a lot of words, $w_1, \ldots, w_k$.
- We approximate

$$P(\textbf{positive} \mid w_1 = \text{yes} \;\&\; w_2 = \text{yes} \;\&\; \ldots \;\&\; w_k = \text{no})$$

- There may not be enough data to satisfy such a specific condition.
- Next time: how do we use conditional independence assumption to help?