*COMP 1433: Introduction to Data Analytics &*
*COMP 1003: Statistical Tools and Applications*

# Lecture 2 – Probability Basics for Data Analytics

Dr. Jing Li
Department of Computing
The Hong Kong Polytechnic University

*Jan 18&20, 2022*

# Looking Back (Last Lecture)

- Course arrangement plan and structure
- Introduction to Data Analytics (An Introduction)
  - **Key concepts**, *data*, *data analytics*, etc.
  - Applications, e.g., *search engine, product recommendation, spam detection*, etc.
  - Advance Technology, e.g., *classification, regression, clustering, matching, time-series analysis*, etc.
- Some basic concepts of *data mining* and *big data*.
  - Good to have an impression about what they are.
  - You'll have chances to learn more about them later!

# Probability VS. Data Analytics

- In analytics process, we usually use random variables to describe the data. ***Why***?
  - Mathematical process to solve real-life problem
  - To make data computable.
  - To discover the patterns and trends behind data.
  - Allows the analysis of distribution and statistics.
- Probability helps predict the how likely that an event will happen.
  - E.g., *weather prediction*, *product recommendation*.

# What is probability?

- *Probability* is a numerical description of *how likely an event is to occur* and or *how likely that a proposition is true*. --- From Wikipedia.
- We run a random experiment $n$ times, during which an event $A$ occurs $m$ times, then we say the *frequency* of $A$'s occurrence is $f_A = \frac{m}{n}$.
- When $n$ is large enough, $f_A$ will be very close to a value $p$, which is defined as the probability of $A$ to occur, i.e., $\lim_{n \to +\infty} f_A \equiv P(A) = p$
  - When we toss a coin, the probability of "heads up" is 0.5

# Roadmap

- Basic Concepts of Probability
- Conditional Probability and Bayes' Formula
- Data Analytics Application: Naïve Bayes
  - Text Classification
  - Naïve Bayes Formulation

# Roadmap

- Basic Concepts of Probability
- Conditional Probability and Bayes' Formula
- Data Analytics Application: Naïve Bayes
  - Text Classification
  - Naïve Bayes Formulation

# What is *sample space*?

- The set of all possible outcomes of an experiment
- Example: *coin flipping*
- What are the possible outcomes?
  - The coin lands *Heads up* - $H$
  - The coin lands *Tails up* - $T$
- The sample space: $S = \{H, T\}$
- For fair coins
  - What are the chances of $H$ and $T$ happening?
  - Empirically, $H$ and $T$ have 50-50 chance to happen.
  - The *probability* for $H$ to happen is 50% (0.5), so does $T$

# What is an *event*?

- Subsets of the sample space
- Example: *rolling a die once*
  - The sample space $S = \{1,2,3,4,5,6\}$
  - An example event $E = \{1,3,4\}$
  - If we *roll the die once*, and the it lands with $1$ or $3$ or $4$ up, then we say the event $E$ *occurs.*

- The probability that $\mathrm{E}$ *occurs* is:
  - $P(E) = \frac{1+1+1}{6} = \frac{3}{6} = \frac{1}{2}$

# What is *probability*?

- A *probability function* $P$ that assigns a real number (the probability of $E$) to every event $E \subset S$
- $P$ *must* satisfy the following basic properties:
  - $0 \leq P(E) \leq 1$
  - $P(S) = 1$
- ***An important property to speed up computing***:
  - For any *disjoint events*, $E_i$ $(i = 1,2, \dots, n)$, we have $P(E_1 \cup E_2 \cup \cdots \cup E_n) = P(E_1) + P(E_2) + \cdots + P(E_n)$

# Some Notions

- $P(E \cup F)$: the probability that E or F occurs
- $P(E, F)$ or $P(EF)$: the probability that both E and F occurs.
- $P(E^c)$: the probability that $E$ does not occur.

# Revisit the Coin Flipping Example

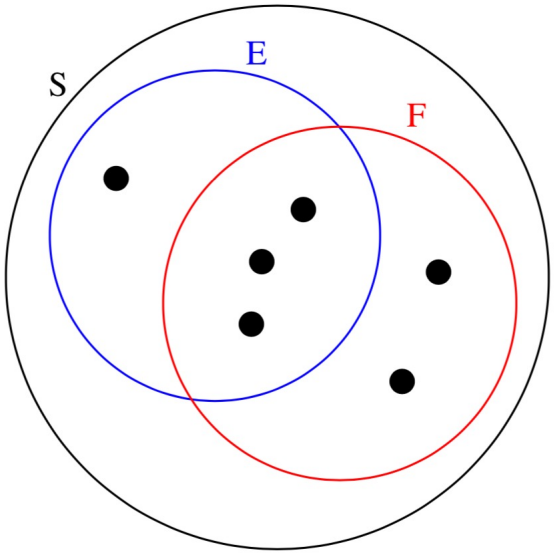
- **Scenario 1:** if the coin is flipped twice, what is the probability of two *heads*?
  - $P(H, H) = \frac{1}{4}$
- **Scenario 2:** if the coin is flipped twice, what is the probability of two *heads*, given that *we know the first toss gave a head*.
  - $P(H, H | H\ at\ the\ first\ toss) = \frac{1}{2}$

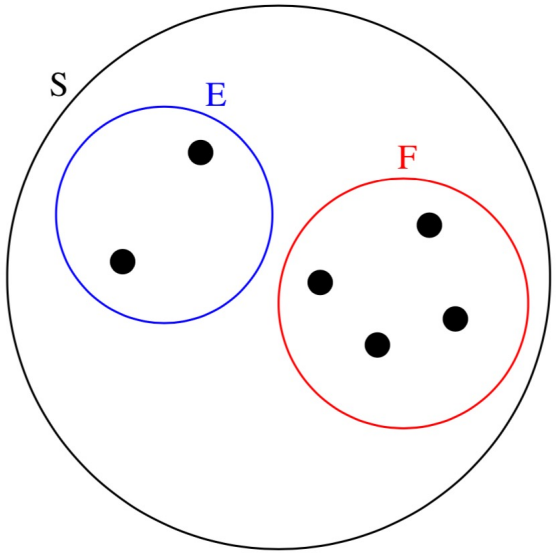
# Roadmap

- Basic Concepts of Probability
- Conditional Probability and Bayes' Formula
- Data Analytics Application: Naïve Bayes
  - Text Classification
  - Naïve Bayes Formulation

# Conditional Probability

- If $E$ and $F$ are events, then $P(E|F)$ is the *conditional probability* of $E$, given $F$.
- $P(E|F) \equiv \frac{P(E,F)}{P(F)}$, assuming that $P(F) \neq 0$




*What is $P(E|F)$ for the two cases?*

# Example: Conditional Probability

- Suppose we draw a card from a shuffled set of 52 playing cards.
- What is the probability of drawing a *Queen*, given that the card drawn is of suits *Hearts*.
  - $P(Q|H) = \frac{P(Q,H)}{P(H)} = \frac{1/52}{1/4} = \frac{1}{13}$
- What is the probability of drawing a *Queen*, given that the card drawn is a *Face* card?
  - $P(Q|F) = \frac{P(Q,F)}{P(F)} = \frac{P(Q)}{P(F)} = \frac{4/52}{12/52} = \frac{1}{3}$

# Discussion: Teddy and Charlie

- It is well known that *Uncle Bob* has two beautiful children, *Teddy* and *Charlie*. However, no one knows whether they are *sons* or *daughters*.
- One day, you met *Bob* in the park. He told you that he has at least a *daughter*. Can you estimate the probability both *Teddy* and *Charlie* are daughters?
- On the other day, you met *Uncle Bob* again. He was with his beautiful daughter and introduced that she was *Teddy*. Now, can you estimate again the probability both *Teddy* and *Charlie* are daughters?

# Law of Total Probability

- Sometimes, the computation of $P(E)$ will be easier if we condition $E$ on another event $F$, namely, from
- $P(E) = P\big(E(F \cup F^c)\big) = P(E, F) + P(E, F^c)$
- Also, $P(E, F) = P(E|F)P(F)$

  *and* $P(E, F^c) = P(E|F^c)P(F^c)$
- $P(E) = P(E|F)P(F) + P(E|F^c)P(F^c)$

# Example: Law of Total Probability

- An insurance company holds the following data concerning the probability of an *insurance claim*:
  - For people under age 30, the probability is 4%
  - For people over age 30, the probability is 2%
- And it is known that 30% of the targeted population is under age 30.
- What is the probability of an insurance claim for a randomly chosen person?

# Example: Law of Total Probability

- $S = \{all\ persons\ under\ consideration\}$
- $C = \{persons\ filing\ a\ claim\}$
- $U = \{persons\ under\ age\ 30\}$
- Thus, $P(C) = P(C|U)\ P(U) + P(C|U^c)\ P(U^c)$

$$= \frac{4}{100}\frac{3}{10} + \frac{2}{100}\frac{7}{10}$$

$$= \frac{26}{1000} = 2.6\% .$$

# Bayes' formula

- Sometimes, we need a formula that *inverts conditioning*, e.g., predicts an event conditioned on some observations.
- Since $P(EF) = P(E|F)P(F)$ *and* $P(EF) = P(F|E)P(E)$
- Then we have


$$P(F|E) = \frac{P(E,F)}{P(E)} = \frac{P(E|F)P(F)}{P(E|F)P(F) + P(E|F^c)P(F^c)}$$

*Law of total probability*

# Example: Bayes' formula

- Suppose 1 in 1,000 persons has a certain disease.
- For 99% of the diseased persons, a test will yield positive results.
- For 5% of the healthy persons, a test will also yield positive results (false alarm).
- What is the probability of a positive test diagnose the disease?
  - $D = \{Diseased\ persons\}$
  - $H = \{Healthy\ persons\}$
  - $+ = \{Persons\ with\ positive\ test\ results\}$

# Example: Bayes' formula

- What is the probability of a positive test diagnose the disease?
  - $D = \{Diseased\ persons\}$
  - $H = \{Healthy\ persons\}$
  - $+ = \{Persons\ with\ positive\ test\ results\}$
- By the given statistics, we have
  - $P(D) = 0.001, P(+|D) = 0.99, P(+|H) = 0.05$
- With Bayes' formula:

$$P(D|+) = \frac{P(+|D) \cdot P(D)}{P(+|D) \cdot P(D) + P(+|H) \cdot P(H)}$$

$$= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.05 \cdot 0.999} \cong 0.0194$$

# Independent Events

- Two events $E$ and $F$ are *independent* if
  - $P(E, F) = P(E)P(F)$
- In this case:
  - $P(E|F) = \frac{P(E,F)}{P(F)} = \frac{P(E)P(F)}{P(F)} = P(E)$, assuming $P(F) \neq 0$
- In other words,
  - *knowing $F$ occurred doesn't change the probability of E.*

# Example: Independent Events

- Two numbers are drawn at random from $\{1,2,3,4\}$
- If order is unimportant, then what is the sample space $S$?
- Define the following functions on $S$:
  - $X(\{i,j\}) = i + j$
  - $Y(\{i,j\}) = |i - j|$
- Which of the following pairs are independent?
  - $X = 5$ and $Y = 2$
  - $X = 5$ and $Y = 1$

# Roadmap

- Basic Concepts of Probability
- Conditional Probability and Bayes' Formula
- Data Analytics Application: Naïve Bayes
  - Text Classification
  - Naïve Bayes Formulation

# Roadmap

- Basic Concepts of Probability
- Conditional Probability and Bayes' Formula
- Data Analytics Application: Naïve Bayes
  - Text Classification
  - Naïve Bayes Formulation

# Is this spam?

**Subject:** **Important notice!**
**From:** Stanford University <newsforum@stanford.edu>
**Date:** October 28, 2011 12:34:16 PM PDT
**To:** undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

# What is the subject of this medical article?

MEDLINE Article

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

**MeSH Subject Category Hierarchy**

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

# Positive or negative movie review?

\+ *...zany characters and richly applied satire, and some great plot twists*

– *It was pathetic. The worst part about it was the boxing scenes...*

\+ *...awesome caramel sauce and sweet toasty almonds. I love this place!*

– *...awful pizza and ridiculously overpriced...*

# Positive or negative movie review?

\+ *...zany characters and **richly** applied satire, and some **great** plot twists*

− *It was **pathetic**. The **worst** part about it was the boxing scenes...*

\+ *...**awesome** caramel sauce and sweet toasty almonds. I **love** this place!*

− *...**awful** pizza and **ridiculously** overpriced...*

# Why sentiment analysis?

- *Movie*:  is this review positive or negative?
- *Products*: what do people think about the new iPhone?
- *Public sentiment*: how is consumer confidence?
- *Politics*: what do people think about this candidate or issue?
- *Prediction*: predict election outcomes or market trends from sentiment

# Basic Sentiment Classification

- Sentiment analysis is the detection of **attitudes**
- Simple task we focus on in this chapter
  - Is the attitude of this text positive or negative?
- We return to affect classification in later chapters

# Summary: Text Classification

- Sentiment analysis
- Spam detection
- Authorship identification
- Language Identification
- Assigning subject categories, topics, or genres
- …

# Roadmap

- Basic Concepts of Probability
- Conditional Probability and Bayes' Formula
- Data Analytics Application: Naïve Bayes
  - Text Classification
  - Naïve Bayes Formulation

# Text Classification: definition

- *Input*:
  - a document $d$
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

- *Output*: a predicted class $c \in C$

# Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “you have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

# Classification Methods: Supervised Machine Learning

- *Input:*
  - a document *d*
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of *m* hand-labeled documents $(d_1, c_1), \ldots., (d_m, c_m)$
- *Output:*
  - a learned classifier *γ:d* → *c*

# Classification Methods: Supervised Machine Learning

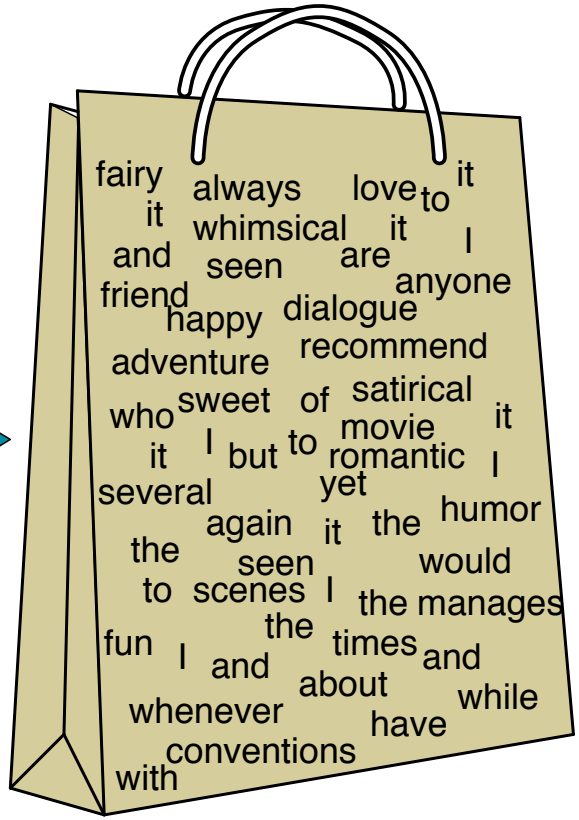
- Any kind of classifier
  - Naïve Bayes
  - Logistic regression
  - Neural networks
  - k-Nearest Neighbors
  - …

# Naive Bayes Intuition

- Simple ("naive") classification method based on Bayes rule (Bayes' formula)
- Usually written as Naïve Bayes (in French).
- Relies on very simple representation of document
  - **Bag of words**

# The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!




| | |
|---|---|
| it | 6 |
| I | 5 |
| the | 4 |
| to | 3 |
| and | 3 |
| seen | 2 |
| yet | 1 |
| would | 1 |
| whimsical | 1 |
| times | 1 |
| sweet | 1 |
| satirical | 1 |
| adventure | 1 |
| genre | 1 |
| fairy | 1 |
| humor | 1 |
| have | 1 |
| great | 1 |
| … | … |

# The bag of words representation

$\gamma($

| | |
|---|---|
| seen | 2 |
| sweet | 1 |
| whimsical | 1 |
| recommend | 1 |
| happy | 1 |
| ... | ... |

$)=c$

# Bayes' Rule Applied to Documents and Classes

- For a document *d* and a class *c*

$$P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$$

Our goal is to maximize $P(c \mid d)$ with a $c \in C$

# Naive Bayes Classifier (I)

**argmax**: the arguments of the maximum (e.g., find the optimal value of $c$ which maximizes $P(c|d)$)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(c \mid d)$$

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes' Formula

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

# Naive Bayes Classifier (II)

"Likelihood" "Prior"

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(d \mid c) P(c)$$

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \cdots, x_n \mid c) P(c)$$

Document d represented as features $x_1, x_2, \ldots, x_n$

# Naive Bayes Classifier (III)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \cdots, x_n \mid c) P(c)$$

O(|$X$|$^n$•|$C$|) parameters

How often does this class occur?

Could only be estimated if a very, very large number of training examples was available.

We can just count the relative frequencies in a corpus

# Multinomial Naive Bayes Independence Assumptions

$$P(x_1, x_2, \cdots, x_n \mid c)$$

- **Bag of Words assumption**: Assume position doesn't matter
- **Conditional Independence**: Assume the feature probabilities $P(x_i \mid c_j)$ are independent given the class *c.*

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

# Multinomial Naive Bayes Classifier

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

$$X = \{x_1, x_2, \ldots, x_n\}$$

# Applying Multinomial Naive Bayes Classifiers to Text Classification

positions ← all word positions in the test document

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

# Problems with multiplying lots of probabilities

- There's a problem with this:

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

- Multiplying lots of probabilities can result in floating-point underflow!
- Luckily, $\log(ab) = \log(a) + \log(b)$
- Let's sum logs of probabilities instead of multiplying probabilities!

# Put them in in the log space

Instead of this:

$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(x_i \mid c_j)$$

This:

$$c_{\mathrm{NB}} = \underset{c_j \in C}{\operatorname{argmax}} \left[ \log P(c_j) + \sum_{i \in \mathrm{positions}} \log P(x_i | c_j) \right]$$

This is ok since log doesn't change the ranking of the classes (class with highest prob still has highest log prob)

Model is now just max of sum of weights: a *linear* function of the inputs

So naive bayes is a *linear classifier*

# One Slide to Takeaway

- What is a **sample space**, **event**, **probability**?
- What is **conditional probability**?
- What is **Bayes' formula**?
- What are **independent events**?
- What is **text classification**?
- How to formulate a **Naïve Bayes classifier**?