Naïve Bayes Classifier

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST

icmoon@kaist.ac.kr

Weekly Objectives

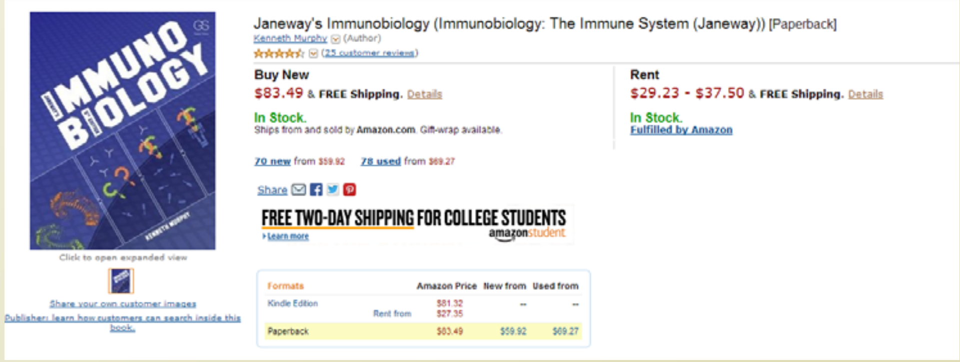
- Learn the optimal classification concept
  - Know the optimal predictor
  - Know the concept of Bayes risk
  - Know the concept of decision boundary
- Learn the naïve Bayes classifier
  - Understand the classifier
  - Understand the Bayesian version of linear classifier
  - Understand the conditional independence
  - Understand the naïve assumption
- Apply the naïve Bayes classifier to a case study of a text mining
  - Learn the bag-of-words concepts
  - How to apply the classifier to document classifications

TEXT MINING APPLICATION: SIMPLE SENTIMENT CLASSIFICATION

Product Review and Sentiment Analysis

- Amazon
  - Product information
  - Also, product review
- Product review
  - Some are positive
  - Some are negative
- What-if we have 10,000 reviews and want to find the negative ones?

Capture from Amazon



Janeway's Immunobiology (Immunobiology: The Immune System (Janeway)) [Paperback]
Kenneth Murphy (Author)
(25 customer reviews)

Buy New
$83.49 & FREE Shipping. Details
In Stock.
Ships from and sold by Amazon.com. Gift-wrap available.

Rent
$29.23 - $37.50 & FREE Shipping. Details
In Stock.
Fulfilled by Amazon

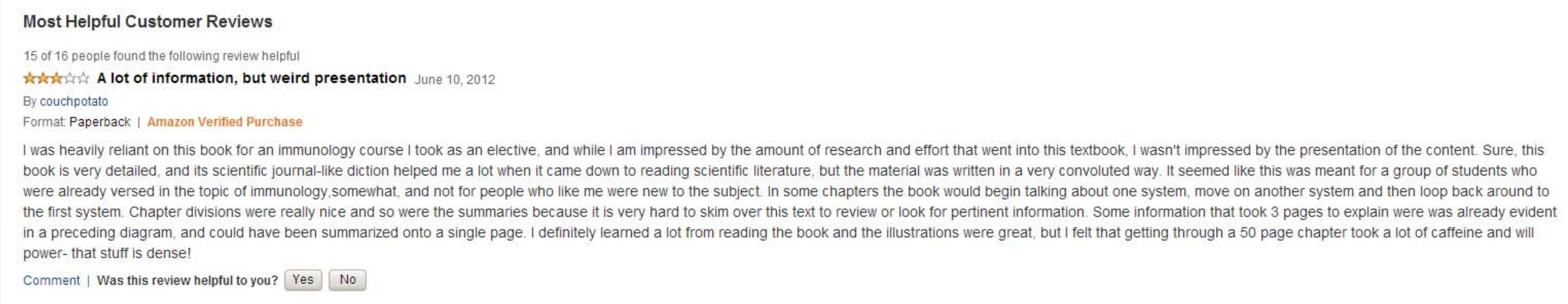
70 new from $59.92 78 used from $69.27

FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS
Learn more

Formats		Amazon Price	New from	Used from
Kindle Edition		$81.32	--	--
	Rent from	$27.35		
Paperback		$83.49	$59.92	$69.27

IMMUNO BIOLOGY
GS
KENNETH MURPHY
Click to open expanded view
Share your own customer images
Publisher: learn how customers can search inside this book.
Share
amazonstudent

Capture from Amazon



Most Helpful Customer Reviews

15 of 16 people found the following review helpful

★★★☆☆ **A lot of information, but weird presentation** June 10, 2012

By couchpotato

Format: Paperback | Amazon Verified Purchase

I was heavily reliant on this book for an immunology course I took as an elective, and while I am impressed by the amount of research and effort that went into this textbook, I wasn't impressed by the presentation of the content. Sure, this book is very detailed, and its scientific journal-like diction helped me a lot when it came down to reading scientific literature, but the material was written in a very convoluted way. It seemed like this was meant for a group of students who were already versed in the topic of immunology,somewhat, and not for people who like me were new to the subject. In some chapters the book would begin talking about one system, move on another system and then loop back around to the first system. Chapter divisions were really nice and so were the summaries because it is very hard to skim over this text to review or look for pertinent information. Some information that took 3 pages to explain were was already evident in a preceding diagram, and could have been summarized onto a single page. I definitely learned a lot from reading the book and the illustrations were great, but I felt that getting through a 50 page chapter took a lot of caffeine and will power- that stuff is dense!

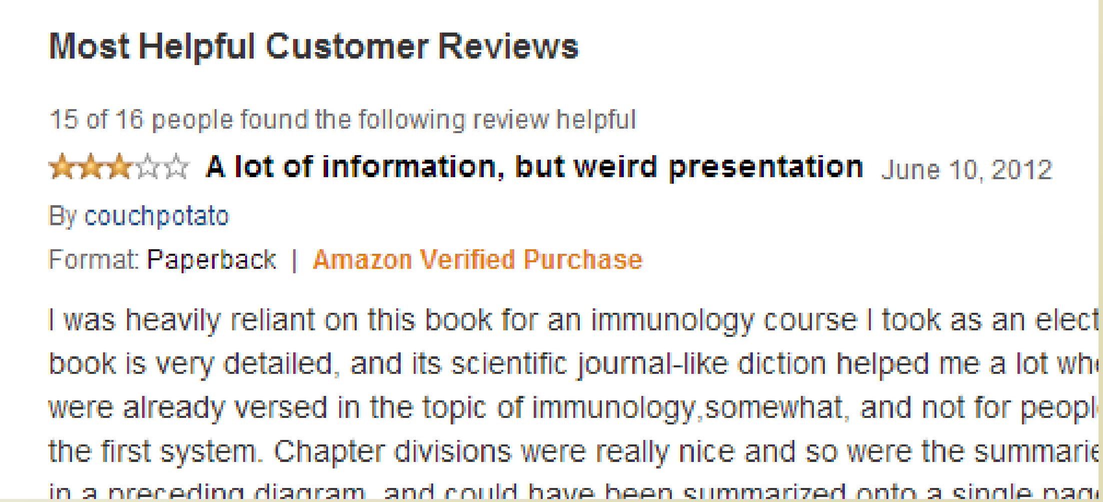
Why simple word searching doesn't work

- There are universal good and bad words
  - Excellent, good, super…
  - Horrible, worst, never…
- How about this?
  - Cool?
    - Cool Beer
  - Hot?
    - Hot Pizza
  - Big?
    - Big LCD
  - Small?
    - Small Size
- Searching and counting
  → Probabilistic approach

Bag Of Words

- For statistical analyses
  - We turned the review text into a vector

Capture from Amazon



Most Helpful Customer Reviews

15 of 16 people found the following review helpful

★★★☆☆ **A lot of information, but weird presentation** June 10, 2012

By couchpotato

Format: Paperback | Amazon Verified Purchase

I was heavily reliant on this book for an immunology course I took as an elect
book is very detailed, and its scientific journal-like diction helped me a lot whe
were already versed in the topic of immunology,somewhat, and not for peopl
the first system. Chapter divisions were really nice and so were the summarie
in a preceding diagram, and could have been summarized onto a single pag

- A vector <1,0,0,1>
- A word list <I, cool, lcd, reliant>
- Together,
  - The review contains words: “I” and “reliant”

Sample Dataset

- Bag of words
  - 198 documents
  - 29717 unique words
- Classes
  - Positive Sentiment
  - Negative Sentiment
- How to apply the Naïve Bayes Classifier?
  - $f_{NB}(x) = argmax_{Y=y} P(Y = y) \prod_{1 \le i \le d} P(X_i = x_i | Y = y)$
  - You need to calculate…
    - $P(Y = y)$
    - $P(X_i = x_i | Y = y)$