Naïve Bayes Classifier

Il-Chul Moon
Dept. of Industrial and Systems Engineering
KAIST
icmoon@kaist.ac.kr

Weekly Objectives

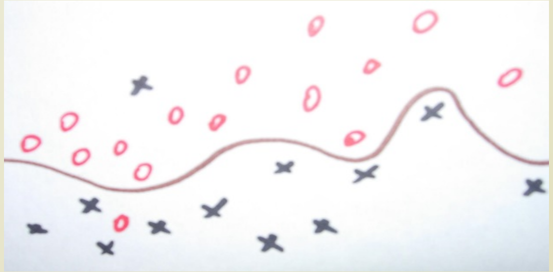
- Learn the optimal classification concept
  - Know the optimal predictor
  - Know the concept of Bayes risk
  - Know the concept of decision boundary
- Learn the naïve Bayes classifier
  - Understand the classifier
  - Understand the Bayesian version of linear classifier
  - Understand the conditional 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

OPTIMAL CLASSIFICATION AND DECISION BOUNDARY

Supervised Learning

Supervised Learning

You know the true answers of some of instances

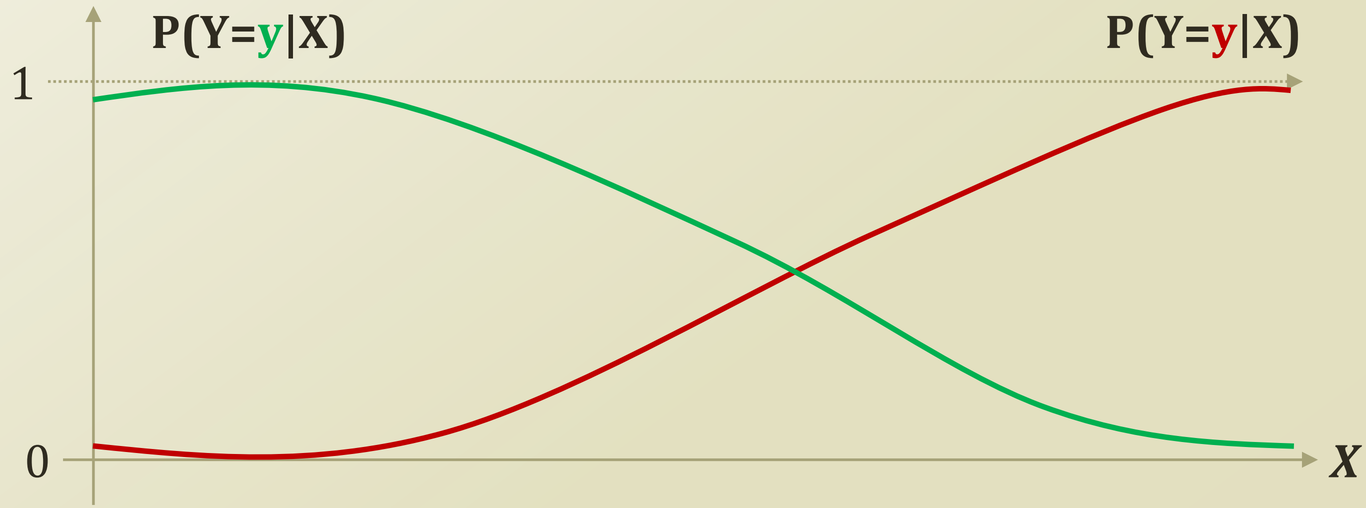


- **You know the true value, and you can provide examples of the true value.**
- Cases, such as
  - Spam filtering
  - Automatic grading
  - Automatic categorization
- Classification or Regression of
  - Hit or Miss: Something has **either disease or not.**
  - Ranking: Someone received **either A+, B, C, or F**.
  - Types: An article is **either positive or negative**.
  - Value prediction: The price of this artifact is **X**.
- Methodologies
  - Classification: estimating a discrete dependent value from observations
  - Regression: estimating a (continuous) dependent value from observations

Optimal Classification

- Optimal predictor of Bayes classifier
  - $f^* = argmin_f P(f(X) \neq Y)$
  - Function approximation of error minimization
- Assuming only two classes of Y
  - $f^*(x) = argmax_{Y=y} P(Y = y|X = x)$

$$\sum_{y \in Y} P(Y = y|X = x) = ?$$



Detour: Thumbtack MLE and MAP

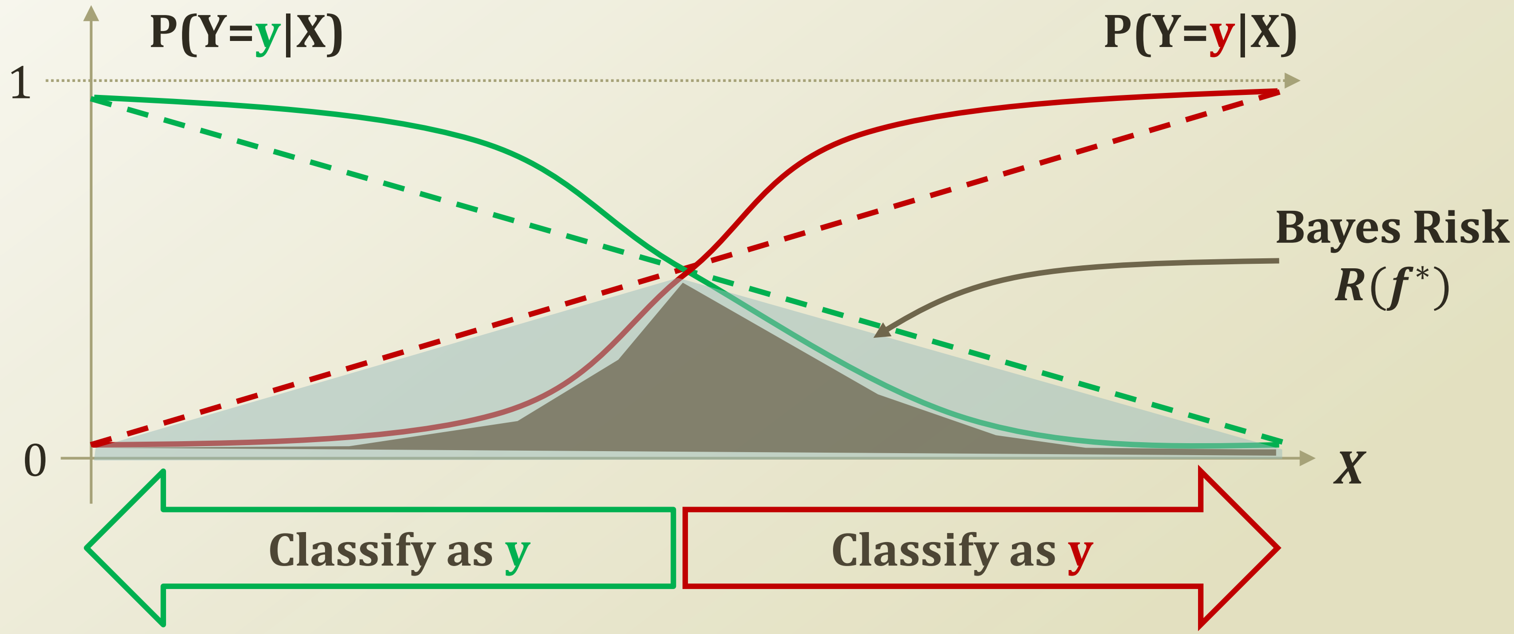
- Your response was
  - Previously in MLE, we found θ from $\hat{\boldsymbol{\theta}} = \boldsymbol{argmax_{\theta} P(D|\theta)}$
    - $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
    - $\hat{\theta} = \frac{a_H}{a_H + a_T}$
  - Now in MAP, we find θ from $\hat{\boldsymbol{\theta}} = \boldsymbol{argmax_{\theta} P(\theta|D)}$
    - $P(\theta|D) \propto \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$
    - $\hat{\theta} = \frac{a_H+\alpha-1}{a_H+\alpha+a_T+\beta-2}$
  - The calculation is same because anyhow it is the maximization
- Assume
  - Y={H,T}, then $\boldsymbol{\theta}$ is a probability value to see the head
  - X=D, previous trials, dataset
  - $\hat{\boldsymbol{\theta}} = \boldsymbol{argmax_{\theta} P(\theta|D)}$

  → $\boldsymbol{f^*(x) = argmax_{Y=y} P(Y=y|X)}$

User assumes
$\hat{\boldsymbol{\theta}} > 0.5$ then Y=H

Classifier tells
Y=H or not

Optimal Classification and Bayes Risk

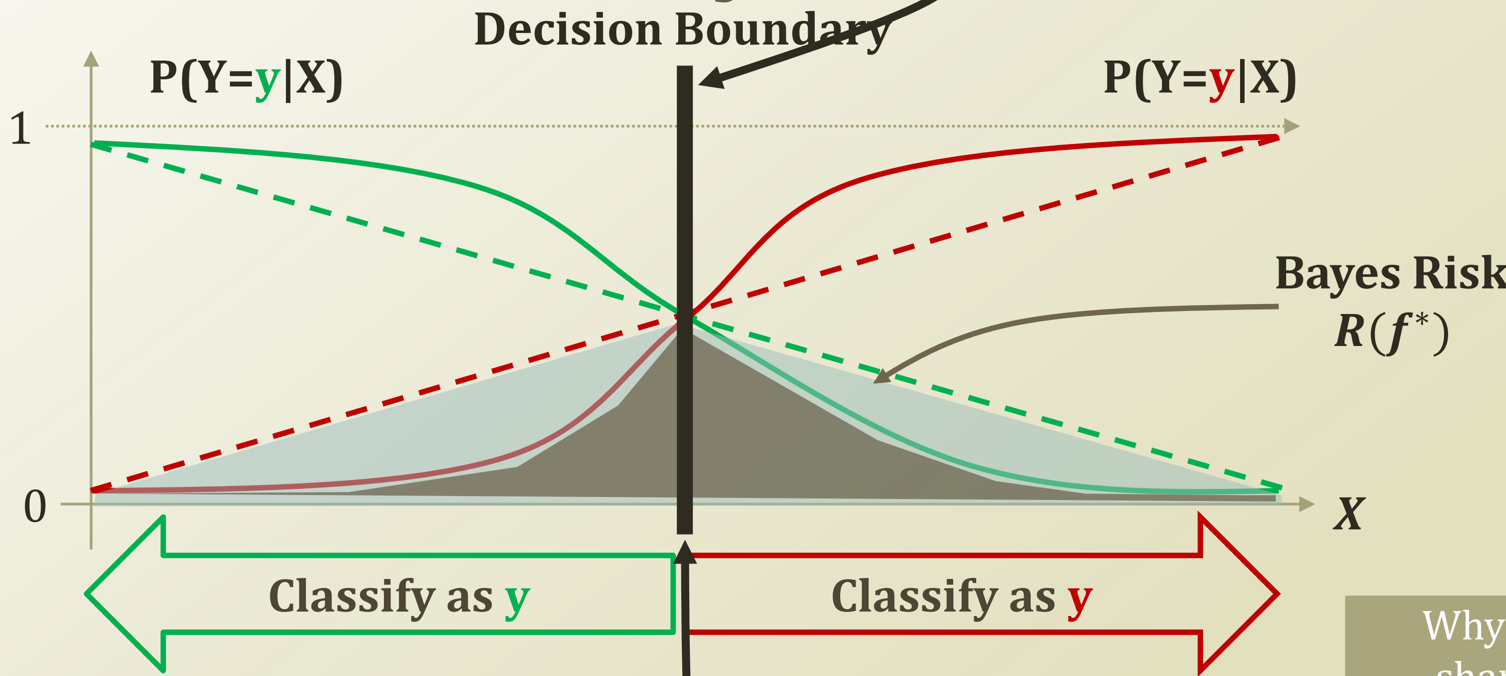


- Optimal classifier will make mistakes, $\boldsymbol{R(f^*) > 0}$
- Why?
  - Not enough information of the joint probability
- $P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y=y)}{P(X=x)}$
- $f^*(x) = argmax_{Y=y}P(Y = y|X = x) = argmax_{Y=y}P(X = x|Y = y)P(Y = y)$

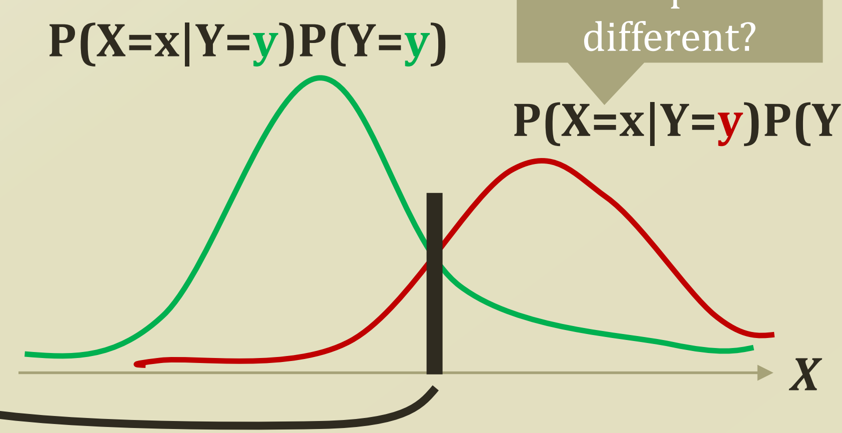
Class Conditional Density

Class Prior

Decision Boundary

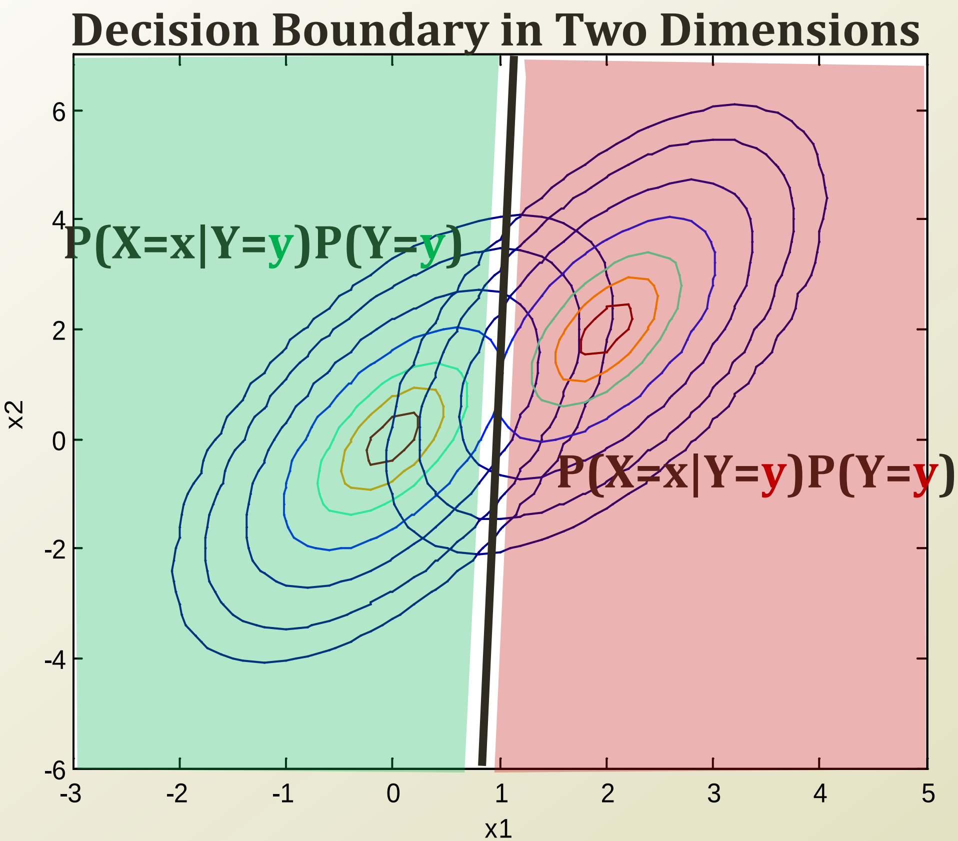


Why are shapes different?



- $f^*(x) = argmax_{Y=y} P(Y = y|X = x)$
  $= argmax_{Y=y} P(X = x|Y = y)P(Y = y)$
- What-if Gaussian class conditional density?
- $P(X = x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Decision Boundary in Two Dimension



$$f^*(x) = argmax_{Y=y} P(Y = y|X = x)$$
$$= argmax_{Y=y} P(X = x|Y = y)P(Y = y)$$

- Two multivariate normal distribution for the class conditional densities
- Decision boundary
  - A linear line
- Linear decision boundary
- Any problem in the real world applications?
  - Observing the combination of x_1 and x_2

$$P(X = x; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

$$P(X = (x_1, x_2)|Y = y) = \frac{1}{\sqrt{2\pi|\Sigma_y|}} \exp(-\frac{(x-\mu_y)\Sigma_y^{-1}(x-\mu_y)'}{2})$$

Learning the Optimal Classifier

- Optimal classifier
  - $f^*(x) = argmax_{Y=y} P(Y = y|X = x)$

$$= argmax_{Y=y} \underbrace{P(X = x|Y = y)}_{\textbf{Class Conditional Density}} \underbrace{P(Y = y)}_{\textbf{Class Prior}}$$

- Need to know
  - Prior = Class Prior = $P(Y = y)$
  - Likelihood = Class Conditional Density = $P(X = x|Y = y)$
- How to know the values?
  - Through observations from the dataset, D
  - Then, does D has all X and Y?
    - Particularly, X in all combinations?

NAÏVE BAYES CLASSIFIER

Dataset for Optimal Classifier Learning

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

- $f^*(x) = argmax_{Y=y} P(X = x|Y = y)P(Y = y)$
  - P(X=x|Y=y)
    =P(x_1=sunny, x_2=warm, x_3=normal, x_4=strong, x_5=warm, x_6=same|y=Yes)
  - P(Y=y)=(y=Yes)
- How many parameters are needed? How many observations are needed?
  - P(X=x|Y=y) for all x,y — $(2^d-1)k$
  - P(Y=y) for all y — k-1

Often, what happens is
N >> $(2^d-1)k$ >> |D|

- Remember that we are not living in the perfect world!
  - Noise exists, so need to model it as a random variable with a distribution
  - Replications are needed!

Why need an additional assumption?

- $f^*(x) = argmax_{Y=y} P(X = x|Y = y)P(Y = y)$
  - To learn the above model, we need a very large dataset that is impossible to get
- The model has relaxed unrealistic assumptions, but now the model has become impossible to learn.
  - Time to add a different assumption
  - An assumption that is not so significant like the ones being relaxed

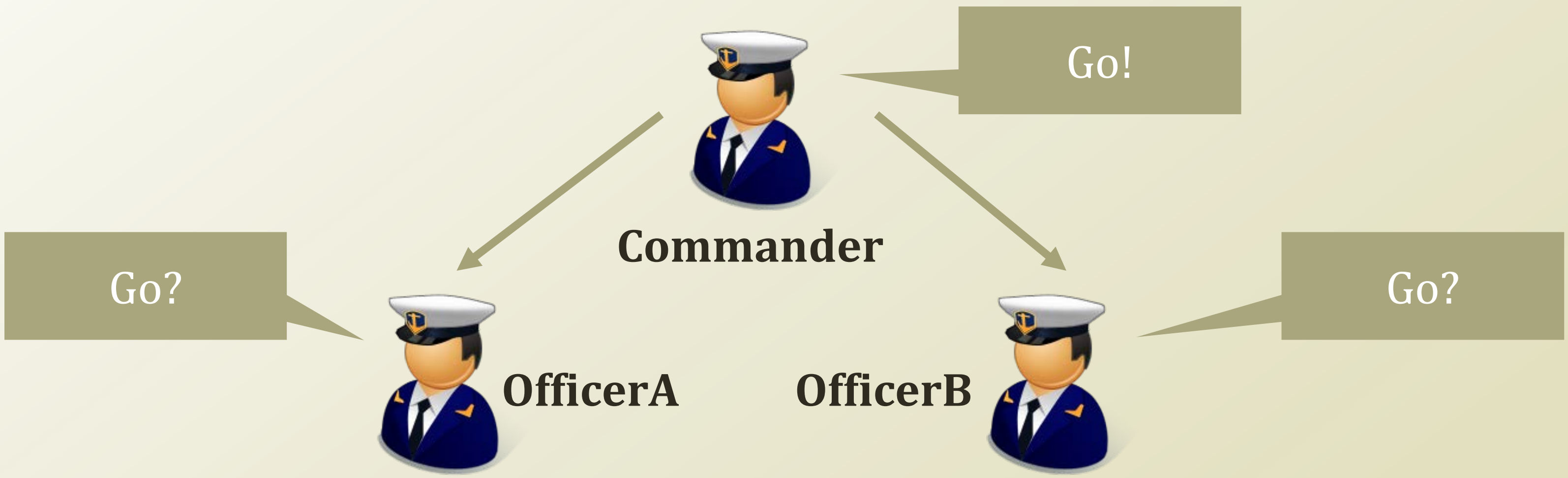
- What are the major sources of the dataset demand?
  - P(X=x|Y=y) for all x,y → $(2^d-1)k$
    - x is a vector value, and the length of the vector is d
    - d is the source of the demand
    - Then, reduce d?
    - Or, ????

Conditional Independence

- A passing-by statistician tells us
  - Hey, what if?
    - $P(X =< x_1, \dots, x_i > | Y = y) \rightarrow \prod_i P(X_i = x_i | Y = y)$
  - Your response: Is it possible?
    - Statistician: Yes! If $x_1,\dots,x_i$ are conditionally independence given y

- Conditional Independence
  - x_1 is conditionally independent of x_2 given y
  - $(\forall x_1, x_2, y) \quad P(x_1|x_2, y) = P(x_1|y)$
  - Consequently, the above asserts
    - $P(x_1, x_2|y) = P(x_1|y)P(x_2|y)$
  - Example,
    - P(Thunder|Rain, Lightning)=P(Thunder|Lightening)
    - If there is a ***lightening***, there will be a ***thunder*** with a prob. ***p*** regardless of ***raining***

Conditional vs. Marginal Independence



- Marginal independence
  - P(OfficerA=Go|OfficerB=Go) > P(OfficerA=Go)
  - **This is not marginally independent!**
    - X and Y are independent if and only if P(X)=P(X|Y)
    - Consequently, P(X,Y)=P(X)P(Y)
- Conditional independence
  - P(OfficerA=Go|OfficerB=Go,Commander=Go) =P(OfficerA=Go|Commander=Go)
  - **This is conditionally independent!**

Dataset for Optimal Classifier Learning with Conditional Independent Assumption

Sky	Temp	Humid	Wind	Water	Forecst	EnjoySpt
Sunny	Warm	Normal	Strong	Warm	Same	Yes
Sunny	Warm	High	Strong	Warm	Same	Yes
Rainy	Cold	High	Strong	Warm	Change	No
Sunny	Warm	High	Strong	Cool	Change	Yes

- Previously, $f^*(x) = argmax_{Y=y} P(X = x|Y = y)P(Y = y)$
  - ***P(X=x|Y=y) has (2^d-1)k cases***
- Let's apply the conditional independent assumption to the all features of X (=all variables in the vector of *x*)
- Now, $f^*(x) = argmax_{Y=y} P(X = x|Y = y)P(Y = y)$

$$\approx argmax_{Y=y} P(Y = y) \prod_{1 \le i \le d} P(X_i = x_i|Y = y)$$

  - How many parameters after adopting the assumption?
  - $\boldsymbol{P(X_i = x_i|Y = y)}$ ***has (2-1)dk cases***
- *You: Wait! The passing-by statistician! Is that right????!!!!*

Naïve Bayes Classifier

- Statistician: Yeah. I know that the assumption is naïve. Why don't you call it as naïve Bayes classifier?

- Given:
  - Class Prior $\mathbf{P(Y)}$
  - $\boldsymbol{d}$ conditionally independent features $\boldsymbol{X}$ given the class $\boldsymbol{Y}$
  - For each $\boldsymbol{X_i}$, we have the likelihood of $\boldsymbol{P(X_i|Y)}$
- **Naïve Bayes Classifier Function**
  - $\boldsymbol{f_{NB}(x) = argmax_{Y=y} P(Y=y) \prod_{1 \le i \le d} P(X_i = x_i | Y = y)}$

- Naïve Bayes classifier is the optimal classifier
  - If the conditional independent assumptions on X hold
  - If the prior is right
- Any problems????

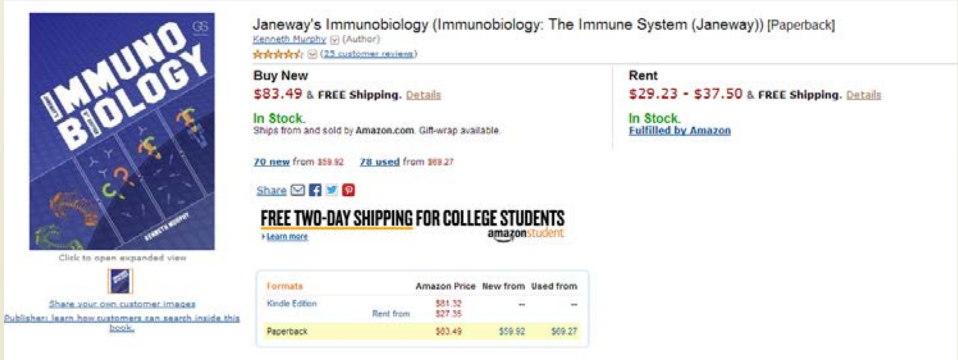
Problem of Naïve Bayes Classifier

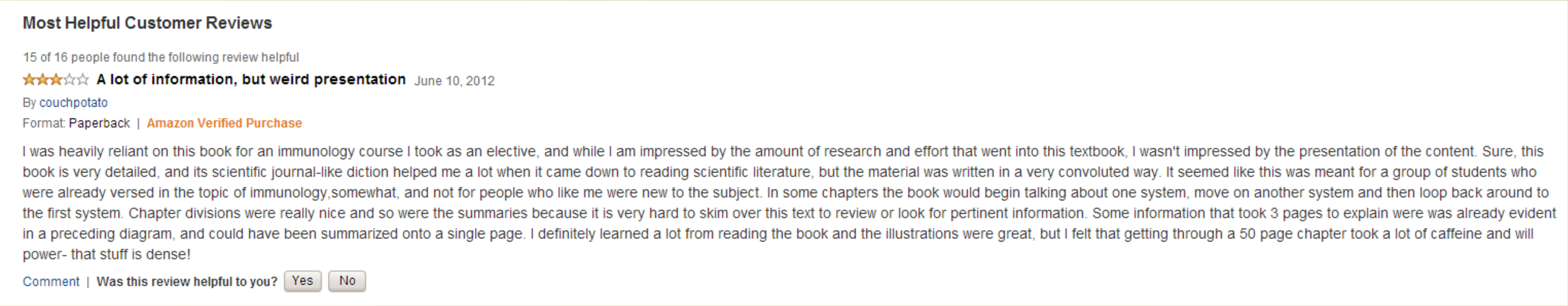
- Problem 1: Naïve assumption
  - Many, many, many cases, the variables of X are correlated
  - Why?
  - Multi-collinearity
- Problem 2: Incorrect Probability Estimations
  - Billionaire
    - Head, Head, Head…
  - MLE with insufficient data
    - There is no chance of Tail!
    - P(Y=tail) = 0
  - MAP with stupid prior
    - Is either our dataset or knowledge good enough to estimate the prior?
- Problem 2 is always there!
- Problem 1 is introduced by our assumption!

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?





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
  - 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…
    - $P(Y = y)$
    - $P(X_i = x_i | Y = y)$

Matlab Exercise!

- Let's do some coding…

Acknowledgement

- This slideset is greatly influenced
  - By Prof. Eric P. Xing at CMU

Further Readings

- Bishop Chapter 1, 8.2