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

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

- $f^*(x) = argmax_{Y=y} P(X = x|Y = y)P(Y = y)$
  - $P(X=x|Y=y)$
    $=P(x_1=\text{sunny}, x_2=\text{warm}, x_3=\text{normal}, x_4=\text{strong}, x_5=\text{warm}, x_6=\text{same}|y=\text{Yes})$
  - $P(Y=y)=(y=\text{Yes})$
- 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

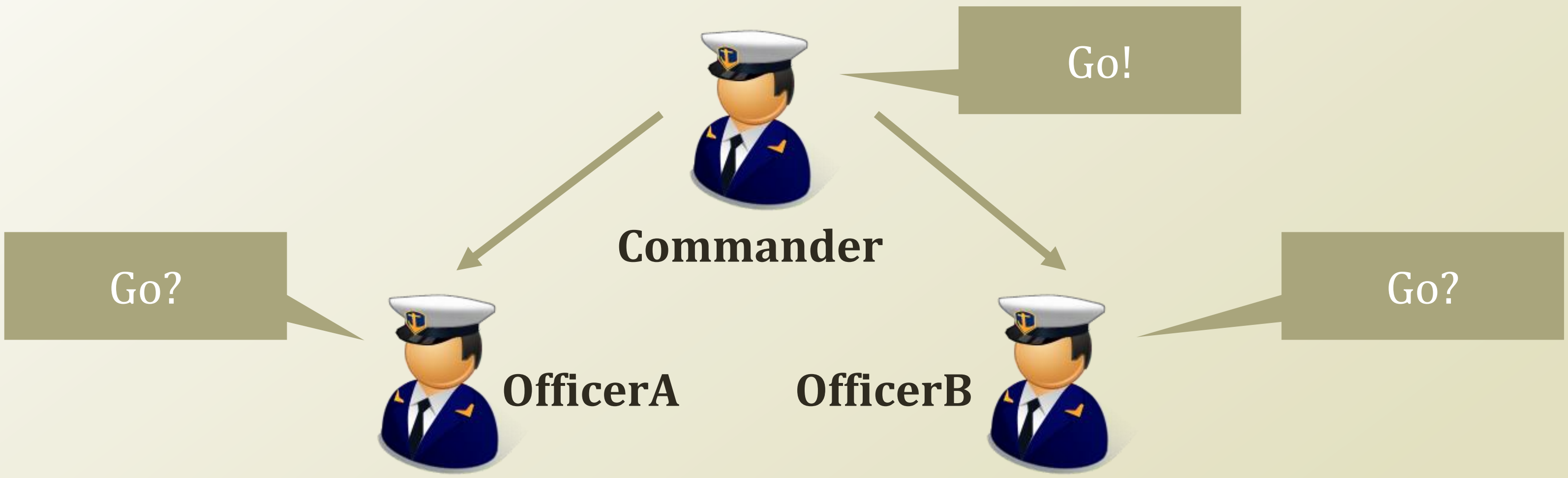
- What are the major sources of the dataset demand?
  - P(X=x|Y=y) for all x,y → $(2^d-1)k$
    - x is a vector value, and the length of the vector is d
    - d is the source of the demand
    - Then, reduce d?
    - Or, ????

Conditional Independence

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