Naïve Bayes Classifier

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

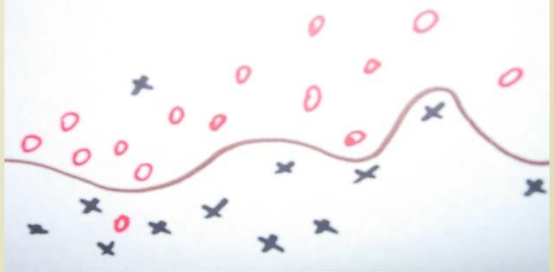
Weekly Objectives

- 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

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

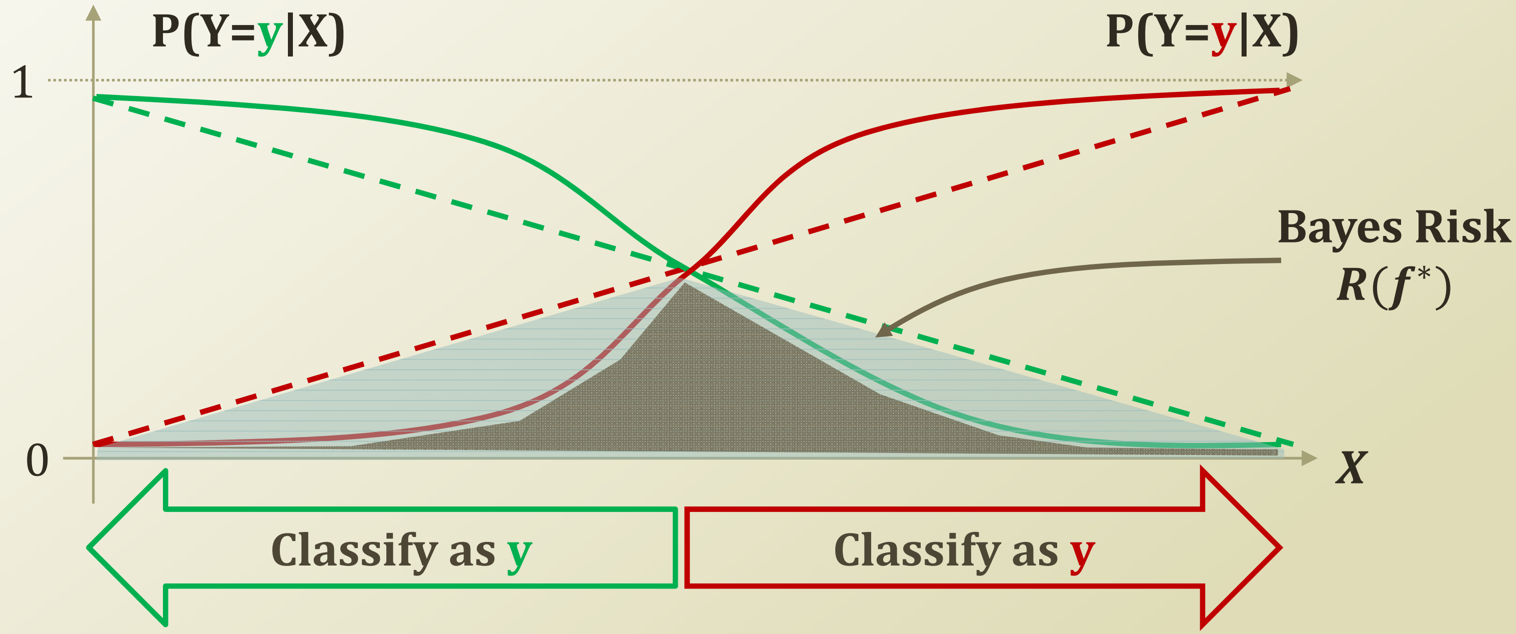
- Your response was
  - Previously in MLE, we found θ from $\widehat{\boldsymbol{\theta}} = \boldsymbol{argmax_{\theta} P(D|\theta)}$
    - $P(D|\theta) = \theta^{a_H}(1-\theta)^{a_T}$
    - $\hat{\theta} = \frac{a_H}{a_H + a_T}$
  - Now in MAP, we find θ from $\widehat{\boldsymbol{\theta}} = \boldsymbol{argmax_{\theta} P(\theta|D)}$
    - $P(\theta|D) \propto \theta^{a_H+\alpha-1}(1-\theta)^{a_T+\beta-1}$
    - $\hat{\theta} = \frac{a_H+\alpha-1}{a_H+\alpha+a_T+\beta-2}$
  - 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
  - $\widehat{\boldsymbol{\theta}} = \boldsymbol{argmax_{\theta} P(\theta|D)}$

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

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

Classifier tells
Y=H or not

Optimal Classification and Bayes Risk

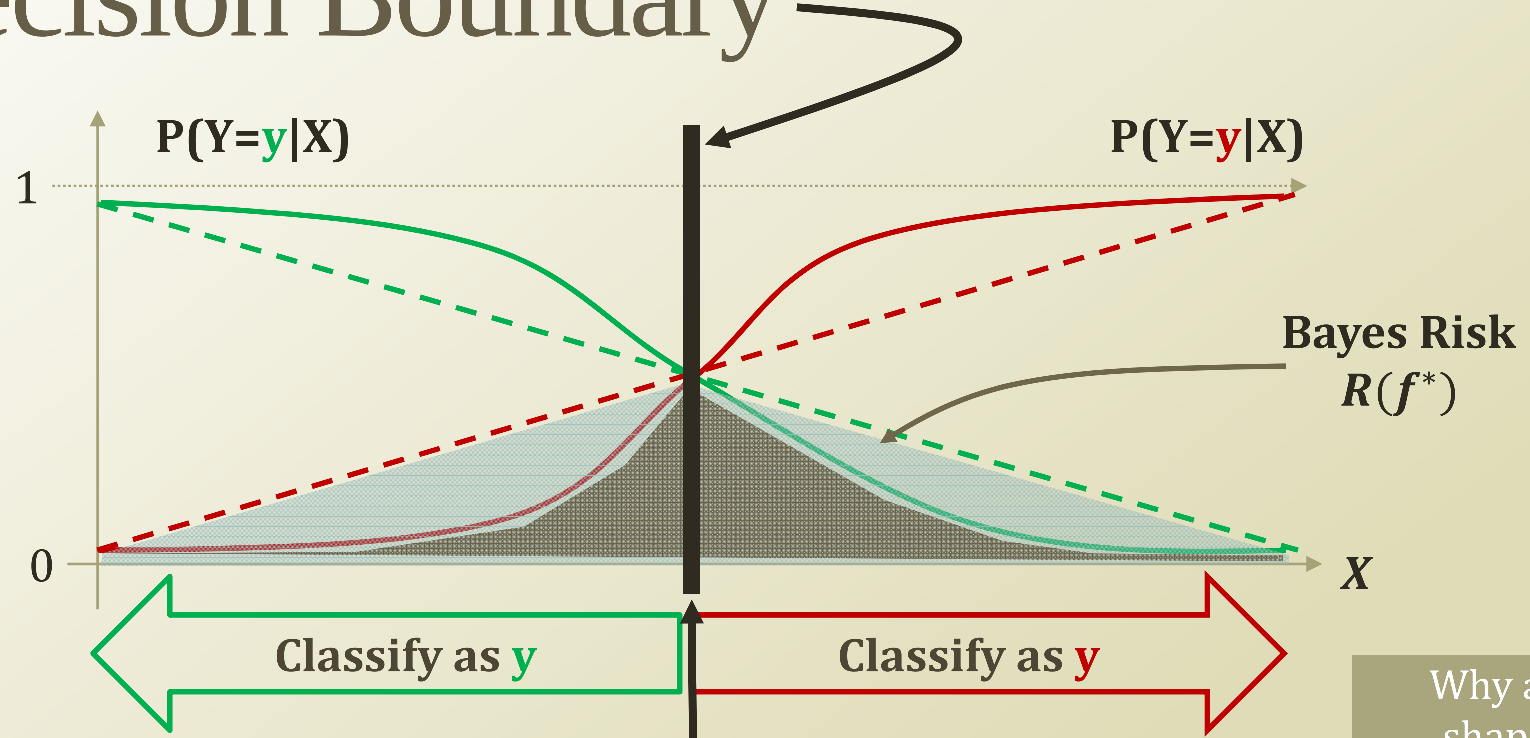


- Optimal classifier will make mistakes, $\boldsymbol{R(f^*) > 0}$
- Why?
  - Not enough information of the joint probability
- $P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y=y)}{P(X=x)}$
- $f^*(x) = argmax_{Y=y}P(Y = y|X = x) = argmax_{Y=y}P(X = x|Y = y)P(Y = y)$

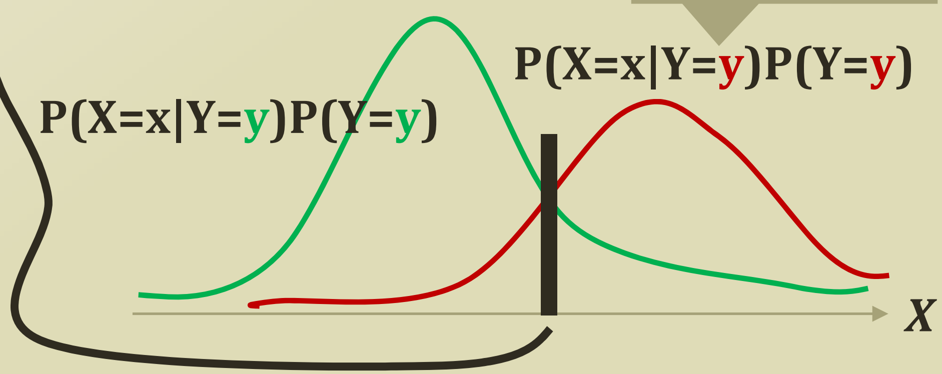
Class Conditional Density

Class Prior

Decision Boundary

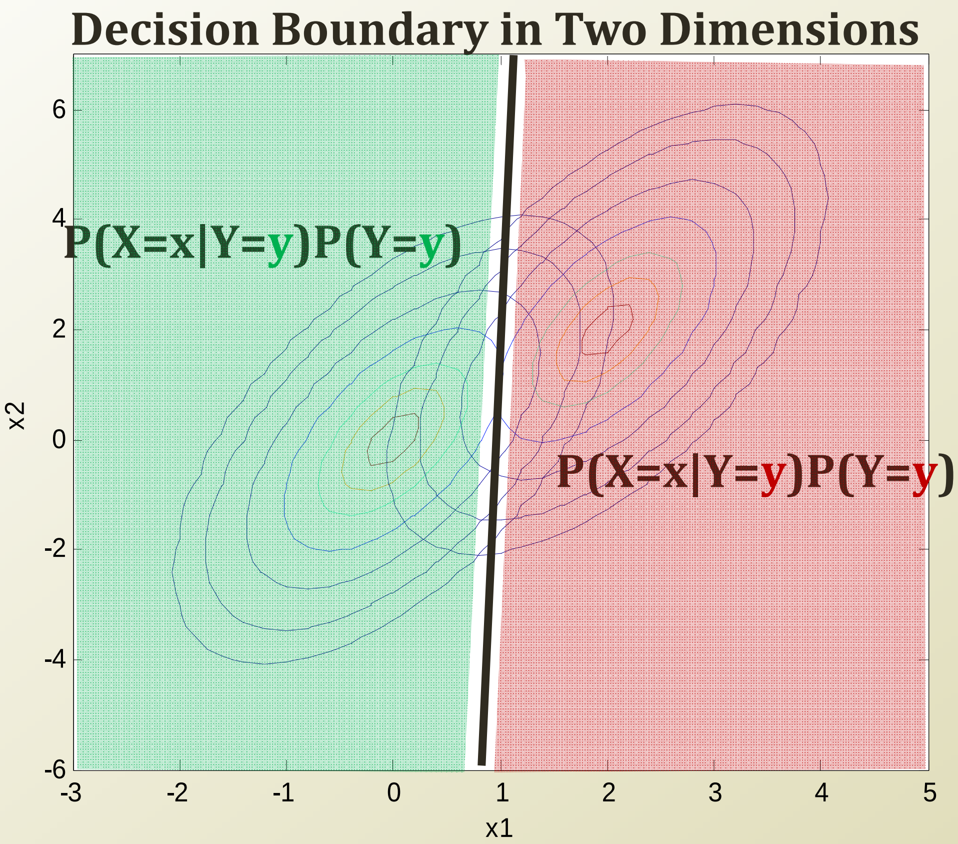


Why are shapes different?



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