Bayesian Decision Theory

Jeonghun Yoon

Terms

Random variable

Bayes rule

Classification

Decision Theory

Bayes classifier

Conditional independence

Naive Bayes Classifier

Laplacian smoothing

MLE / Likehood / Prior

Random Variable

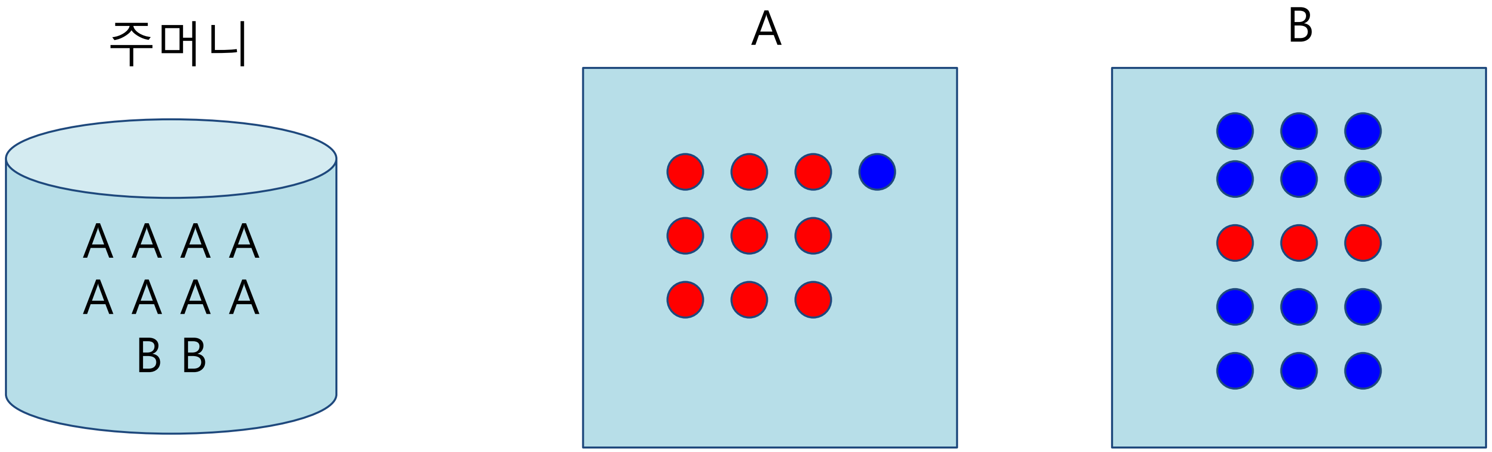
주사위를 던졌을 때 3이 나올 확률은 ⅙ 이다. 이것을 수학적으로 표현해보자.

주사위를 던졌을 때 3이 나오는 사건을 c라고 하자. 사건을 실수에 대응시키는 함수를 X라고 하자. 여기서 $X(c) \ = \ 3$ 이다. 만약 주사위를 던졌을 때 6이 나오는 사건을 d라고 하면 $X(d) = 6$ 이다.

여기서 X를 랜덤변수라고 한다. 즉 **사건을 실수에 대응시켜주는 함수**가 **랜덤변수**이다.

따라서 주사위를 던졌을 때 3일 나올 확률을 수학적으로 표현해보면 $\boldsymbol{P(X(c = \ 3) = \ }$ ⅙ 이다. 이것을 간단히 표현하면 $\boldsymbol{P(X = 3)}$ = ⅙ 또는 $\boldsymbol{P(3)}$ = ⅙ 이라고 표현할 수 있다.

Bayes rule



질문 : 파란 공을 뽑았다. 이 파란 공은 A에서 나왔을까? B에서 나왔을까?

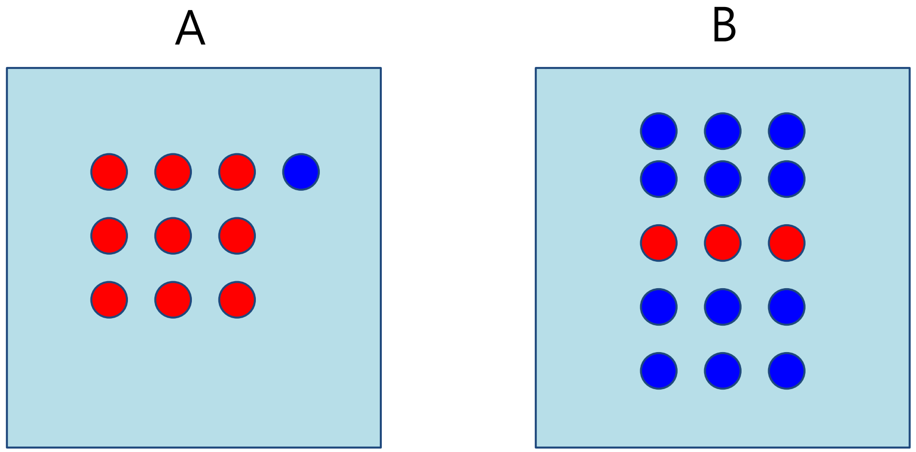
Bayes rule

대답 1)

B에서 나왔을 것 같다. B에 들어있는 파랑 공이 훨씬 더 많기 때문이다.

이것을 수학적으로 표현하면, 상자 A에서 파란 공을 뽑을 확률 $P(\text{파랑}|A) = \frac{1}{10}$ 보다 상자 B에서 파랑 공을 뽑을 확률 $P(\text{파랑}|B) = \frac{12}{15}$가 더 크다 라고 할 수 있다.

이 조건부 확률을 우도(Likelihood) 또는 우도 함수라고 한다.



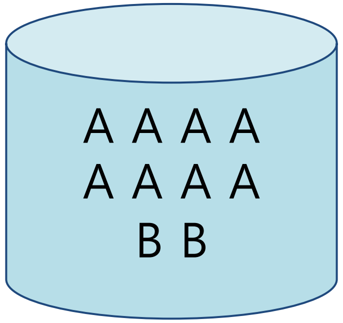
Bayes rule

대답 2)

주머니 속을 보면 상자 A 제비가 7개 상자 B 제비가 3개이다. 따라서 파랑 공은 A에서 나왔을 것이다. 이것을 수학적으로 표현하면, 상자 A를 선택할 확률 $P(A) = \frac{9}{10}$이 상자 B를 선택할 확률 $P(B) = \frac{1}{10}$ 보다 높다라고 할 수 있다.

이 확률은 사전 확률(prior probability)이라고 한다.

주머니



Bayes rule

다음 질문들은 동치다.

- 파란 공을 뽑았다. 이 파란 공은 A에서 나왔을까? B에서 나왔을까?
- 파랑 공이 관찰 된 조건하에, 파랑 공이 A에서 나왔을 확률이 높을까? B에서 나왔을 확률이 높을까?
- $P(A|\text{파랑}) > P(B|\text{파랑})$ 인가? 또는 $P(A|\text{파랑}) < P(B|\text{파랑})$ 인가?

이 조건부 확률을 사후 확률(Posterior probability)이라 부른다.

Bayes rule

Bayes rule를 수학적으로 유도해보자.

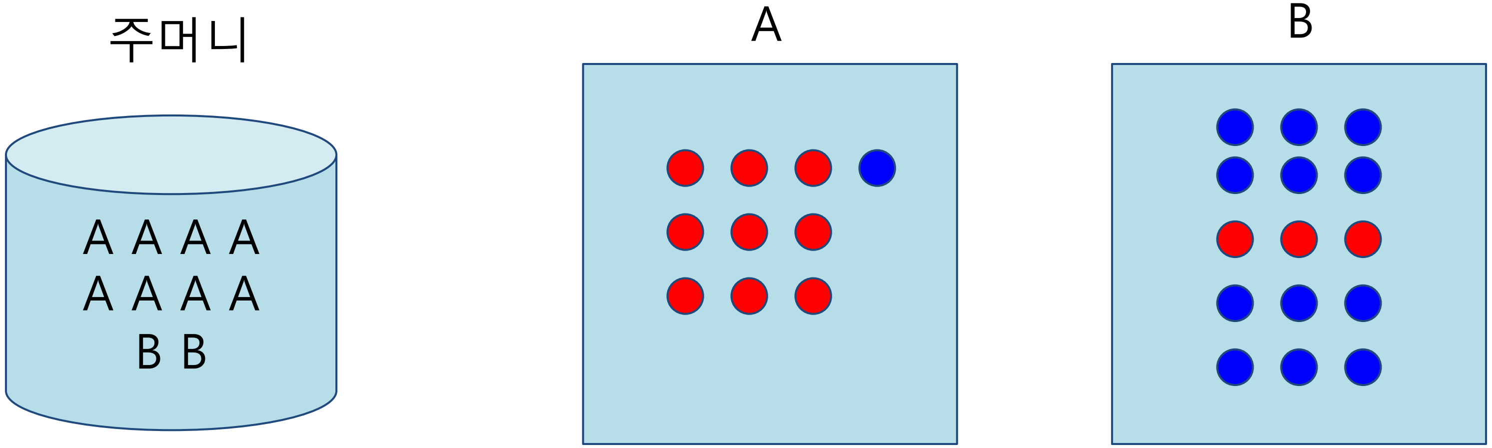
$P(X,Y) = P(Y,X)$

$\Leftrightarrow P(X)P(Y|X) = P(Y)P(X|Y)$

$\Leftrightarrow P(X|Y) = \frac{P(Y|X)P(X)}{P(Y)}$

($P(Y) = \sum_X P(Y|X)P(X)$)

Bayes rule



$$P(A|blue) = \frac{P(blue|A)P(A)}{P(blue)} = \frac{\left(\frac{1}{10}\right)\times(\frac{8}{10})}{\frac{24}{100}} = \frac{1}{3}$$

$$P(B|blue) = \frac{P(blue|B)P(B)}{P(blue)} = \frac{\left(\frac{12}{15}\right)\times(\frac{2}{10})}{\frac{24}{100}} = \frac{2}{3}$$

(사후확률) 신뢰도 0.66으로 파랑 공은 B에서 나왔다.

Bayes rule

likelihood
(우도 값)

prior
(사전 확률)

$$p(\theta|\mathbb{x}) = \frac{p(\mathbb{x}|\theta)p(\theta)}{\sum p(\mathbb{x}|\theta)p(\theta)}$$

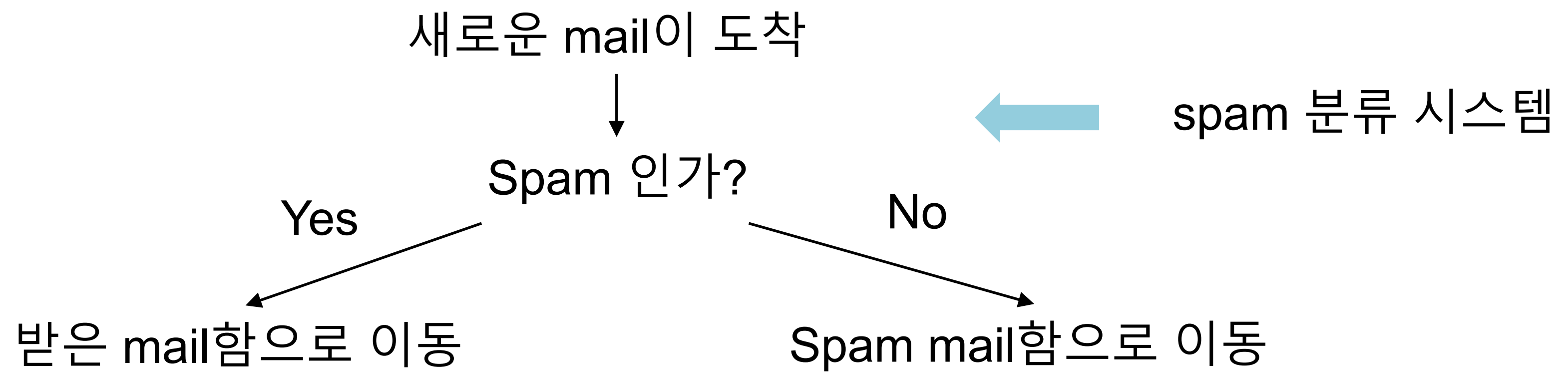
posteriori
(사후 확률)

- 사후 확률 : 관찰 값들이 관찰 된 후에 모수(parameter)의 발생 확률을 구한다.
- 사전 확률 : 관찰 값들이 관찰 되기 전에 모수의 발생 확률을 구한다.
- 우도 값 : 모수의 값이 주어졌을 때 관찰 값들이 발생할 확률

Spam 메일 분류

Mailbox 에서 spam설정을 한 적이 없다.
(물론 수동으로 조건을 입력하면 더 잘 작동한다.)

그런데 어떻게 아래와 같은 알고리즘이 작동하는 걸까?



How?

Classification

$\mathbb{X} \xrightarrow{f} y$

Feature vector
특성 벡터

Classifier
분류기

Class
부류

E-mail (Words)
Web Site (Words)
꽃 (꽃의 생김새)

Spam or Not
Sports or Science or News
Line-flower or Mass-flower

For example,
(visit, money, buy, girl, Viagra) $\xrightarrow{f}$ Spam

Decision Theory

이미지 분류(처리)에서 decision thoery를 적용했을 때 processing 과정

images $\xrightarrow{\text{Image processing}}$ measurement $\mathbb{x}$ $\xrightarrow{\text{Decision rule}}$

C_1 Cancel

C_2 Cancel

가정 :
- 이미지를 measurements($\mathbb{x}$) 로 만들고, 그것을 이용하여 암인지 아닌지를 분류하는 것
- Measuremetns와 그것이 속하는 부류(class)의 결합확률분포를 이용 $p(\mathbb{x}, C_i) = p(\mathbb{x}|C_i)p(C_i)$

목표 :

Make the "Best"$_{\text{how to define?}}$ decision.

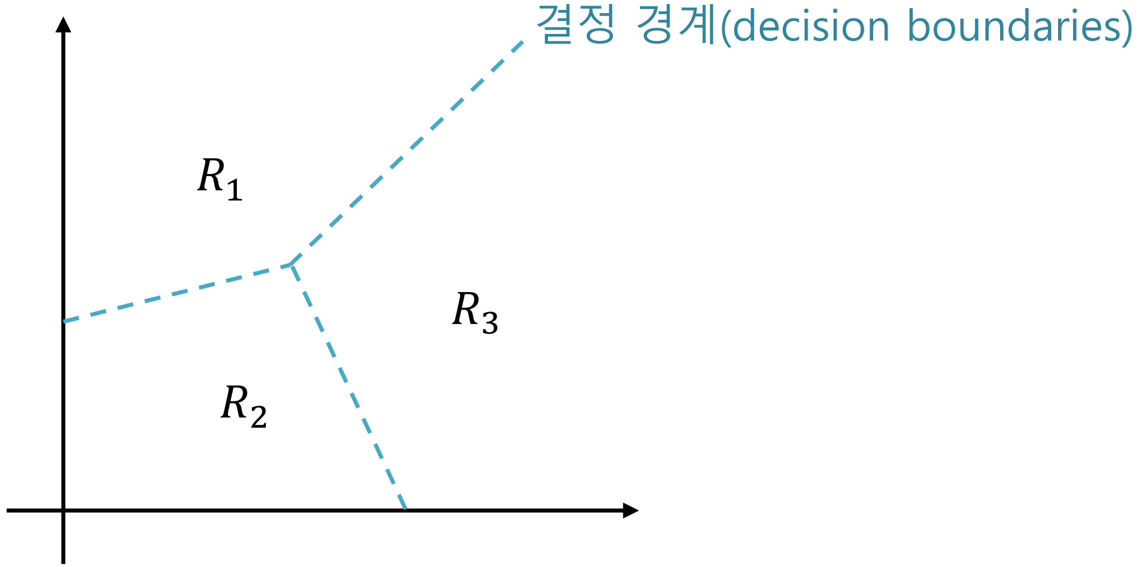
Classification using decision theory

Decision rule은 결정 경계$_{\text{decision boundaries}}$를 이용하여 입력 공간$_{\text{input space}}$을 결정 구역$_{\text{decision regions}}$으로 나눈다.

- 입력 공간 : measurement가 존재하는 공간 (그림에서는 1사분면 전체)

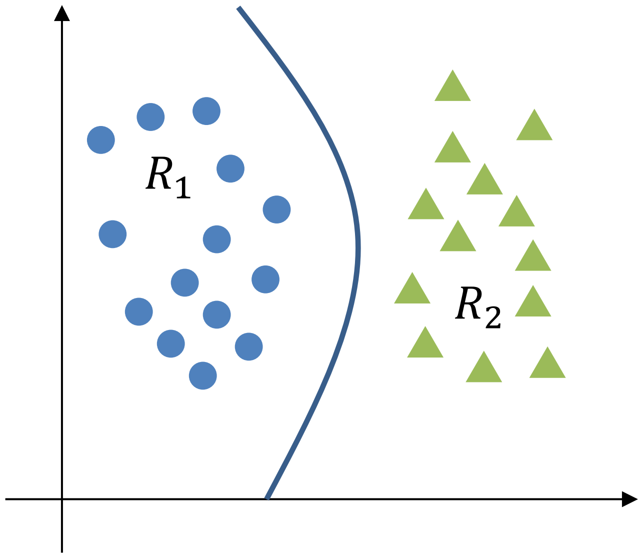
- 결정 구역 : 결정 경계를 기준으로 나눠진 공간. 결정 구역에 속한 measurement는 해당 구역으로 할당(classify)된다.

- 결정 경계 : 각 class를 구분짓는 경계



Classification using decision theory

예제) 2차원 vector measurement의 2 부류$_{class}$ 결정 경계



Decision boundary

결국 우리가 해결해야하는 것은 결정 경계를 찾는 것이다.

Decision theory의 가정에서 언급했던 것처럼 measurement(feature)와 class의 결합확률밀도 $P(\mathbb{x}, C)$를 이용 할 것이다.

그러면 결합확률밀도 함수를 이용하여 decision boundary를 어떻게 찾을까?

강의 처음에 언급했던 Bayes rule을 이용할 것이다.

Bayes classifier

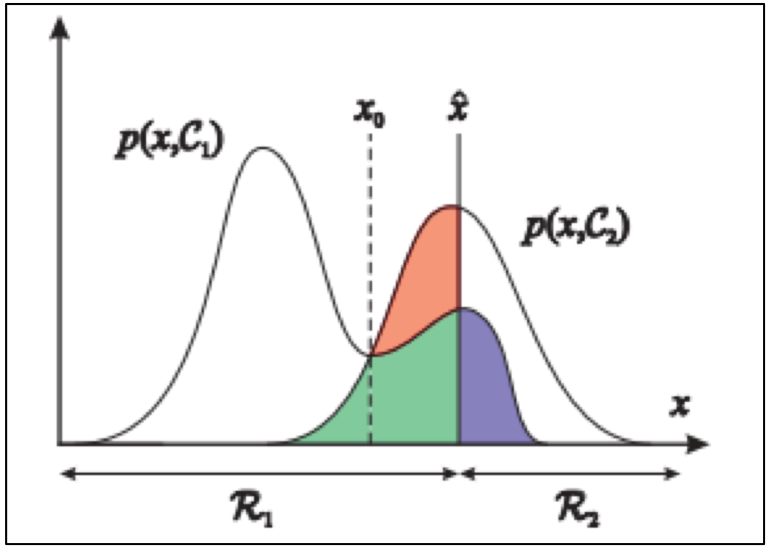
Bayes rule을 이용하여 결정 경계를 찾고 그것을 기준으로 measurement(feature)를 각 class로 분류하는 분류기를 Bayes classifier라고 한다.

Bayes classifier의 2가지 기준

- 오분류 비율의 최소화
- 기대 손실의 최소화

Decision Boundary for average error

2 class decision (x는 1차원)



그림출처 : PRML 책

- $P(\mathbb{x}, C_1)$:
$\mathbb{x}$와 C_i의 결합확률. $\mathbb{x}$가 C_i 와 함께 존재하는 확률
또는 $\mathbb{x}$가 C_i에 실제로 존재하는 확률로 해석하면 된다.

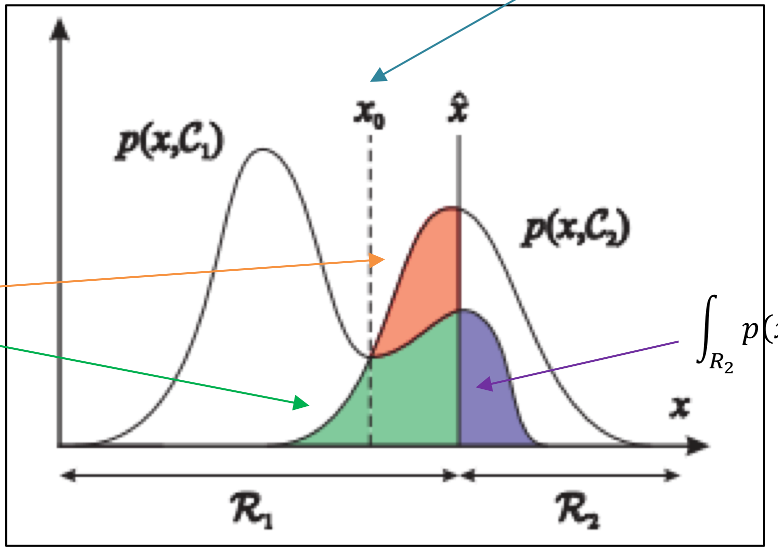
- 결정경계가 정해지고 난 이후에,
  - R_1에 속하면 C_1으로 분류
  - R_2에 속하면 C_2으로 분류

Decision Boundary for average error

2 class decision (x는 1차원)

x_0는 misclassifications(오분류)이 최소가 되는 지점.
이 영역에서 주황색영역이 사라지고 녹색만 남는다.

$\int_{R_1} p(x, C_2)dx$



$\int_{R_2} p(x, C_1)dx$

$\hat{x}$을 기준으로 분류할 경우, 오분류 영역을 살펴보자.

- C_1에 속하는데 R_2으로 분류하는 경우 (보라색 영역)
- C_2에 속하는데 R_1으로 분류하는 경우 (연두색 + 주황색 영역)

$$p(error) = \int_{-\infty}^{\infty} p(error, x)dx = \int_{R_1} p(x, C_2)dx \int_{R_2} p(x, C_1)dx$$

Bayes Decision Rule (error)

결합 확률 분포 $p(x, C_i)$가 큰 C_i에 x를 할당

- $p(x, C_1) > p(x, C_2)$이면, x를 C_1에 할당
- $p(x, C_1) < p(x, C_2)$이면, x를 C_2에 할당

$\updownarrow$ $p(x, C_i) = p(C_i|x)p(x)$ 조건부 확률 (아직 bayes rule은 나오지 않았다.)

사후 확률 분포 $p(C_i|x)$가 큰 C_i에 x를 할당

- $p(C_1|x) > p(C_2|x)$이면, x를 C_1에 할당
- $p(C_1|x) < p(C_2|x)$이면, x를 C_2에 할당

Bayes error

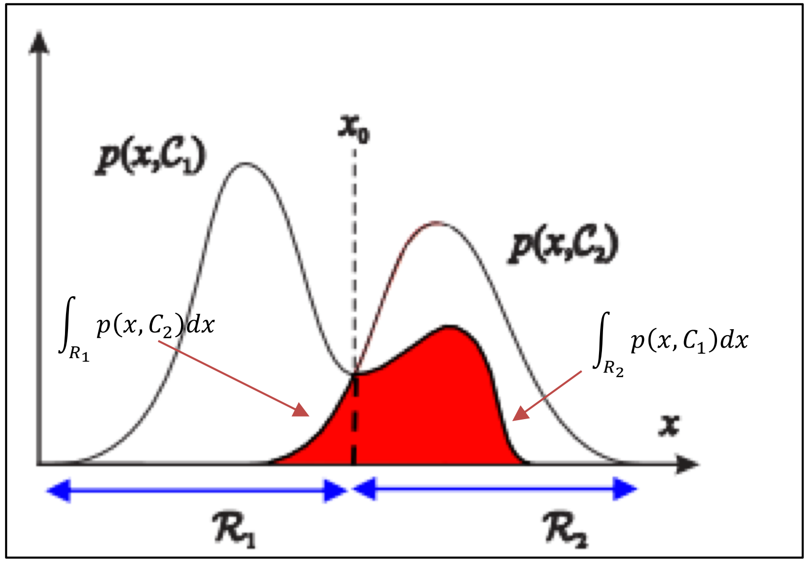
Bayes error는 오분류 확률을 나타낸다.

$p(error)$

$= \int_{-\infty}^{\infty} p(error, x)dx$

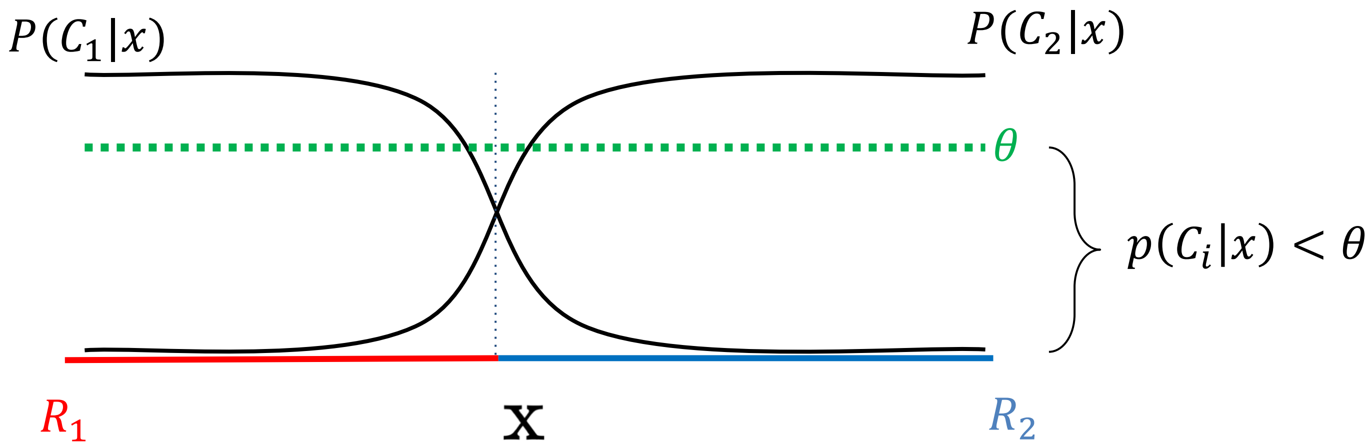
$= \int_{R_1} p(x, C_2)dx \int_{R_2} p(x, C_1)dx$

$= \int_{R_1} p(C_2|x)p(x)dx \int_{R_2} p(C_1|x)p(x)dx$



Reject option

확실하지 않은 영역에 대해서는 분류하지 않는다.



Bayes classifier (오분류 비율의 최소화 ver.)

사후 확률 분포 $p(C_i|x)$가 큰 C_i에 x를 할당

- $p(C_1|x) > p(C_2|x)$이면, x를 C_1에 할당
- $p(C_1|x) < p(C_2|x)$이면, x를 C_2에 할당

$$f^*(x)$$
$$= \arg\max_Y P(Y = y|X = x)$$
$$= \arg\max_Y P(X = x|Y = y)P(Y = y)$$

- Class conditional density
- Likelihood

Gaussian class conditional densities (1-dim)

$$P(X = x|Y = y) = \frac{1}{\sqrt{s\pi\sigma_y^2}}\exp(-\frac{(x - \mu_y^2)}{s\sigma_y^2})$$

Loss matrix

비용 함수(cost function), 손실 함수(loss function)을 도입

Loss matrix : 오분류에 따른 손실을 행렬로 표현한 것

	암 (actual value)	정상 (actual value)
암 (prediction value)	0	**1**
정상 (prediction value)	**1000**	0

Bayes classifier (기대 손실의 최소화 ver.)

Bayes risk $R(a_i|x)$: True status는 j, 선택된 action은 i일때 발생하는 loss

$$R(a_i|x) = \sum_j \overbrace{L(\underbrace{a_i}_{\text{action}}|C_j)}^{L_{ij}}\,p(\underbrace{C_j}_{\text{measurement}}|x)$$

Loss matrix

		H_0 is rejected.	H_0 is not rejected.
		암 (actual value)	정상 (actual value)
Prediction : H_1	암 (prediction value)	0	**1 (type 1 error)**
Prediction : H_0	정상 (prediction value)	**1000 (type 2 error)**	0

Bayes classifier (기대 손실의 최소화 ver.)

Input : x, output : $y(C_1, C_2)$ 인 경우 (2 class)

- C_1으로 분류했을 때의 risk
  - $R(a_1|x) = L_{11}p(C_1|x) + L_{12}p(C_2|x) = L_{11}p(X = x|Y = C_1)p(Y = C_1) + L_{12}p(X = x|Y = C_2)P(Y = C_2)$
- C_2로 분류했을 때의 risk
  - $R(a_2|x) = L_{21}p(C_1|x) + L_{22}p(C_2|x) = L_{21}p(X = x|Y = C_1)p(Y = C_1) + L_{22}p(X = x|Y = C_2)P(Y = C_2)$
- Decision rule
  - $R(a_1|x) > R(a_2|x)$ 이면 C_2로 분류
  - $R(a_1|x) < R(a_2|x)$ 이면 C_1으로 분류

Bayes decision rule : **Bayes risk를 최소화**하는 action을 선택

$$\hat{a}_i = \arg\min_{a_i} R(a_i|x)$$

Conditional independence

X is conditionally independent of Y given Z :

Probability distribution governing X is independent of the value of Y, given the value of Z

$$P(X = x|Y = y, Z = z) = P(X = x|Z = z) \Leftrightarrow P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Naïve Bayes

Conditional independence 가정에 의하여,

$P(You\ love\ to\ buy\ viagra|Spam) =$
$P(You|Spam) \times P(love|Spam) \times P(to|Spam) \times P(buy|Spam) \times P(viagra|Spam)$

이 성립한다. 특정한 조건이 주어지면 문장에서 각각의 단어가 나올 확률은 서로 독립이라는 것이다. 문장 구조 안에서 각 단어는 의존성을 띄는데, 이것을 무시하는 것이다.

그러면 이 가정을 왜 사용하려고 하는가?
그 이유 중 하나는 Bayesian probability를 계산하는 데 필요한 parameters 수를 줄이기 위해서이다.

Parameters in Naïve Bayes

번개가 치는 것을 예측해보자 (L). 아래의 두 조건은 conditionally independent 하다.

- 천둥(T)
- 비(R)

Conditionally independent 조건이 없을 때

- L이 주어졌을 때의 likelihood $P(T,R|L)$ 를 계산하기 위한 parameter의 개수 : $(2^2-1)\times 2 \ = \ 6$

$P(T=0,R=0|L=0)$
$P(T=0,R=1|L=0)$
$P(T=1,R=0|L=0)$

$P(T=1,R=0|L=1)$
$P(T=0,R=1|L=1)$
$P(T=1,R=0|L=1)$

Conditionally independent 조건이 있을 때

- $P(T,R|L)=P(T|L)P(R|L)$ 를 계산하기 위한 parameter의 개수 : $(2-1)\times 2 \ + \ (2-1)\times 2 \ = \ 4$

$P(T=0,|L=0)$ $P(R=0,|L=0)$
$P(T=0,|L=1)$ $P(R=0,|L=1)$

Parameters in Naïve Bayes

필요한 parameter수를 계산하자. $\mathbb{x} = (x_1, x_2, \dots, x_d)$라고 하자. 전체 word의 크기를 d, class의 수를 k라고 하자.

$P(\mathbb{x}|y)P(y) = P(x_1, x_2, \dots, x_d|y)P(y)$의 parameters 수 :

$$\left(\mathbf{2}^{\boldsymbol{d}} - \mathbf{1}\right)\boldsymbol{k} + \boldsymbol{k} - \mathbf{1} = \mathbf{2}^{\boldsymbol{d}}\boldsymbol{k} - \mathbf{1}$$

$P(\mathbb{x}|y)P(y) = P(x_1, x_2, \dots, x_d|y)P(y) = P(x_1|y)P(x_2|y) \dots P(x_d|y)P(y)$의 parameters 수 :

$$\boldsymbol{dk} + \boldsymbol{k} - \mathbf{1}$$

Naive Bayes Classifier

Decision rule

- $x_1, x_2, \dots, x_d$: d개의 word를 의미하고, y는 target class를 의미한다.
- 각 클래스에서 d개의 단어가 동시에 발생하는 확률을 구하고, 발생 확률이 가장 높은 클래스를 선택한다.

$$f_{NB}(\mathbb{x}) = \arg\max_y P(x_1, \dots, x_d | y) P(y) = \arg\max_y \prod_{i=1}^{d} P(x_i | y) P(y)$$

Weakness of Naïve Bayes

약점 1 – Naïve Bayes Assumption(가정)

- Naïve Bayes는, data에서 conditional independency 가정이 지켜지지 않더라도, 좋은 성능을 보인다.
- 사실, 특성들(features)은 아래와 같이 conditional independency를 가지고 있지 않음 :

$$P(x_1, \dots, x_d|\mathrm{Y}) \neq \prod_i P(x_i|Y)$$

- 그럼에도 불구하고, Naive Bayes는 가장 많이 사용되어지는 classifier 중 하나임.

Weakness of Naïve Bayes

약점 2 – 불충분한 양의 training data

만약, $y = b$일 때, $x_1 = a$인 training sample를 본 적이 없다고 가정하자.

예를 들어, y= {b=Spam}, x_1 = {a='Earn'} 이라고 하고, Spam 클래스에서는 Earn 이라는 feature가 존재하지 않다고 하자.

$\rightarrow P = (x_1 = a|y = b) = 0$

따라서, 나머지 $x_2, \ldots, x_d$의 값에는 상관없이 항상

$\rightarrow P = (y = b|x_1 = a,\ x_2, \ldots, x_d) = 0$

$\because P(x_1 = a, x_2, \ldots, x_n|y = b) = P(x_1 = a|y = b) \prod_{i=2}^{d} P(x_i|y = b)$

이럴 때는 어떻게 해야 하는가? – **Laplacian Smoothing**

Laplacian Smoothing

Multinomial random variable z라고 하자. z는 1부터 k까지의 값을 가질 수 있다.

m개의 독립인 sample $\{z^{(1)}, \dots, z^{(m)}\}$이 주어졌고, 우리는 이것을 통해서 multinomial distrbution 을 구하고 싶다.

즉, $p(z = 1)$, $p(z = 2)$, … , $p(z = k)$를 구하고 싶다.

추정 값(많은 경우 MLE를 사용한다.)은,

$$p(z = j) = \frac{\sum_{i=1}^{m} I\{z^{(i)} = j\}}{m}$$

이다. 여기서 $I\{.\}$는 지시 함수 이다. 관찰 값 내에서의 빈도수를 사용하여 추정한다.

예를 들어 $\{1, 2, 2, 1, 5, 2, 2, 2, 2, 1\}$ 이면,

$p(z = 1) = 0.3$, $p(z = 2) = 0.6$, $p(z = 5) = 0.1$이다.

Laplacian Smoothing

한 가지 주의 할 것은, 우리가 추정하려는 값은 모집단(population)에서의 모수 $p(z=i)$라는 것이다. 추정하기 위하여 sample을 사용하는 것 뿐이다.

예를 들어, $z^{(i)} \neq 3$ for all $i = 1, \ldots, m$ 이라면, $p(z=3) = 0$ 이 되는 것이다. 10개의 샘플에서 3을 보지 못했다고 해서, $p(z=3) = 0$ 라고 결론 내리는 것이 옳을까?

이것은, 통계적으로 볼 때, 좋지 않은 생각이다. 단지, 표본 집단에서 보이지 않는 다는 이유로 우리가 추정하고자 하는 모집단의 모수 값을 0으로 한다는 것은 통계적으로 좋지 않은 생각이다. (MLE의 약점)

Laplacian Smoothing

이것을 극복하기 위해서는,

① 분자가 0이 되어서는 안 된다.

② 추정 값의 합이 1이 되어야 한다. $\sum_z p(z=j)$=1 (∵ 확률의 합은 1이 되어야 함)

따라서,

$$p(z=j) = \frac{\sum_{i=1}^{m} I\{z^{(i)}=j\} + 1}{m+k}$$

이라고 하자.

Laplacian Smoothing

①의 성립 : sample 내에 j의 값이 없어도, 해당 추정 값은 0이 되지 않는다.

②의 성립 : $z^{(i)} = j$인 data의 수를 n_j라고 하자. $p(z=1) = \frac{n_1+1}{m+k}, \dots, p(z=k) = \frac{n_k+1}{m+k}$ 이다. 각 추정 값을 다 더하게 되면 1이 나온다.

이것이 바로 **Laplacian smoothing**이다.

z가 될 수 있는 값이 1부터 k까지 균등하게 나올 수 있다는 가정이 추가되었다고 직관적으로 알 수 있다.

Laplacian smooting in Naïve Bayes Classifier

Training data : $\{(\mathbb{x}_i, y_i)\}_{i=1}^{n}$, $\mathbb{x}_i = \left(x_1^{(i)}, x_2^{(i)}, \dots, x_d^{(i)}\right)$

$$P(x = a|y = b) = \frac{(the\ number\ of\ j : a \in \mathbb{x}_j\ and\ y_j = b) + 1}{(the\ number\ of\ j : y_j = b) + K}$$

K : x_i의 값이 될 수 있는 words의 수

$$P(y = b) = \frac{(the\ number\ of\ j : y_j = b)\ + 1}{(the\ total\ number\ of\ training\ data) + C}$$

C: Y의 값이 될 수 있는 class의 수