Supervised learning Ⅱ

분류문제에 대한 기본적인 아이디어와 기계학습 소개
(로지스틱 회귀, 선형 판별 분석, SVM)
분류문제는 순서가 없는 개별값 (고양이, 개) 등에 대해
분류를 하고자 하는 것임.
출력변수를 인코딩해서 사용하더라도, 출력값을
통해 모델링하기는 어렵기 때문에, 특정 클래스의
확률을 모델링 함.

→ 클래스 변수 Y가 입력변수 X에 주어진 k와
같을 확률을 특정 함수와 함께 갖도록 설계함.

$Pr[Y=k \mid X] \sim f_k(X)$

$\hat{Y} = \operatorname{argmax}_k Pr[Y=k \mid X]$

$\rightarrow MSE = \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$

MSE는 오류율을 의미하니 언제나 이를 기준으로
사용하지는 않음.

Logistic Regression : 로지스틱 회귀

회귀식에서 모델을 써야 하는 이유를 전개.

→ $Y = 0$ or 1

$P(X) \begin{cases} Pr[Y=1 \mid X] \\ Pr[Y=0 \mid X] = 1 - Pr[Y=1 \mid X] \end{cases}$

위 식으로만 P(X)를 모델링하면 음수가 나올 수 있고
이는 확률로 사용이 불가능 함.

$\rightarrow P(X) \approx e^{\beta_0 + \beta_1 X}$

이와같이 회귀식을 지수의 형태로 표현하여
음수의 값을 갖지 않도록 할 수 있음.

but 1을 초과하는 문제점이 있음.

$\rightarrow P(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$

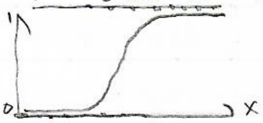
이로써 P(X)는 0~1의 값을 가짐
위 식은 아래와 같이 표현 가능함.

$\log\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X.$

간단한 선형모델로 표현가능한 좌변을
logit of P(X)라고 함.

Sigmoid function

Y가 0 or 1 일때, 로지스틱 회귀선은 아래와 같음.



X가 매우 크거나 작으면 분류하기가 쉬우나
경계에 있는 경우에 분류가 어려워짐.

로지스틱 회귀의 파라미터 찾기

최소제곱법 접근도 가능하나 계산이 어려움

$\rightarrow MSE = \sum_{i=1}^{n} (1 - \hat{Pr}[y_i = k_i \mid X_i])^2$

대신 최대우도 접근법을 사용함.
이는 특정 모델에서 현재 관측된 데이터를 관측할
확률로 정의하는 것임.

$\rightarrow l = \prod_{i: y_i = 0} \frac{1}{1 + e^{\beta_0 + \beta_1 x_i}} \prod_{i: y_i = 1} \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}}$

모집단의 관점에서 가능성을 극대화하는 파라미터를 찾는것

변수가 여러개인 로지스틱 회귀는 아래와 같음.

$\rightarrow \log\left(\frac{P(X)}{1-P(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots \beta_p X_p$

$P(X) = \frac{e^{\beta_0 + \beta_1 X_1 \cdots \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 \cdots \beta_p X_p}}$

LDA (Linear Discriminant Analysis) : 선형 판별 분석

클래스 k의 샘플은 pdf f_k로 부터 생성되었다고 가정함.
다른 클래스의 샘플은 다른 pdf로부터 생성됐다고 가정함.
각 클래스는 관측보다 앞선 사전확률이 있다고 가정하며 π_k로 표현
→ 한 클래스의 샘플을 관측할 확률은 $\pi_k f_k(x)$가 됨.
여러개의 클래스에 대해서 확률을 구해야 하므로
아래와같이 정규화하여 표현 가능함

$\rightarrow P_k(X) = Pr[Y=k \mid X=x] = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$

이는 즉 Y가 k이고 X가 x일때의 확률로서
'posterior probability (사후확률)' 이라고 부른다.