4. 분류

분류기 종류

- 로지스틱 회귀(logistic regression)
- 선형판별분석(linear discriminant analysis)
- K-최근접이웃(knn)

#More Computing#

- 일반화가법모델(generalized additive model)
- 트리(tree)
- 랜덤포리스트(random forest)
- 부스팅(boosting)
- 서포트 벡터 머신(support vector machine)

MLE

- **Maximum Likelihood Estimate**

- With MLE, we seek a point value for θ which maximizes the likelihood, p(D|θ), shown in the equation(s) above. We can denote this value as θ^. In MLE, θ^ is a point estimate, not a random variable.

- In other words, in the equation above, MLE treats the term p(θ)p(D) as a constant and does NOT allow us to inject our prior beliefs, p(θ), about the likely values for θ in the estimation calculations.

Bayesian Estimate

- Bayesian estimation, by contrast, fully calculates (or at times approximates) the posterior distribution $p(\theta|D)$. Bayesian inference treats θ as a random variable. In Bayesian estimation, we put in probability density functions and get out probability density functions, rather than a single point as in MLE.

MLE와 Bayesian의 관계

- 베이즈 법칙(Bayesian Law) 또는 베이즈 이론(Bayesian Theory) 를 간단히 말하자면

$$P(A|B) = \frac{P(B|A)}{P(B)} \cdot \mathbf{P(A)}$$

- **사전확률 p(A) 과 우도확률 p(B|A)를 안다면 사후확률 p(A|B)를 알 수있다는 것이다.**
- 여기서 P(A|B)는 사후확률 Posterior 또는 Posterior belief
- P(A)는 Prior, 또는 prior belief
- P(B)는 Evidence
- P(B|A)는 Likelihood라고 한다.

2장 : 베이즈 분류기

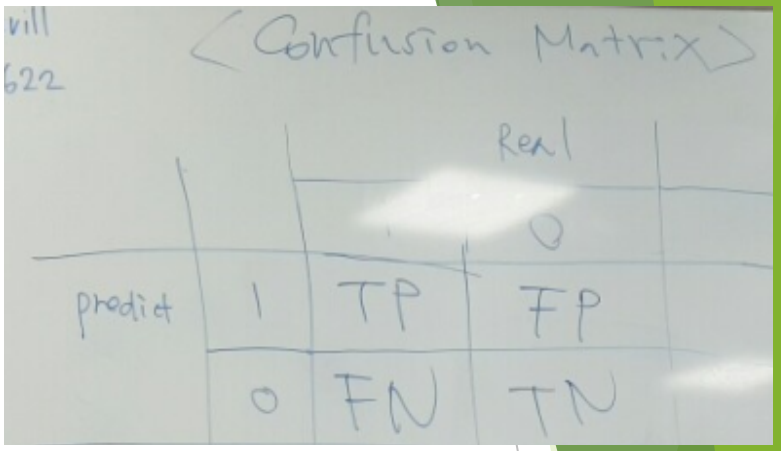
- 오직 두 개의 범주값(1,2)만 가진다고 가정.
- 베이즈 분류기는 $Pr(Y=1|X=x)>0.5$이면 1, 아니면 0이라고 분류
- 베이즈 분류기는 가장 큰 클래스로 관측치를 분류하며 모든 분류기 중에서 오차률이 가장 낮다.

Cf)$Pr(Y=j|X=x)$

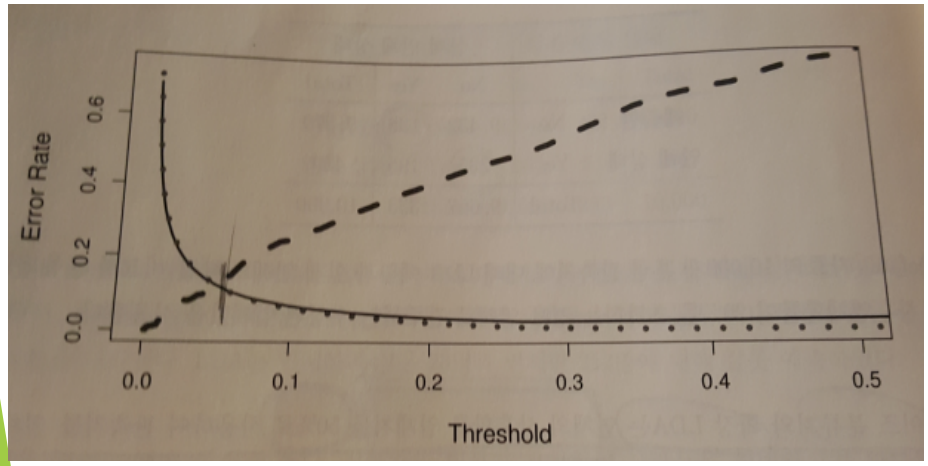
-조건부 확률

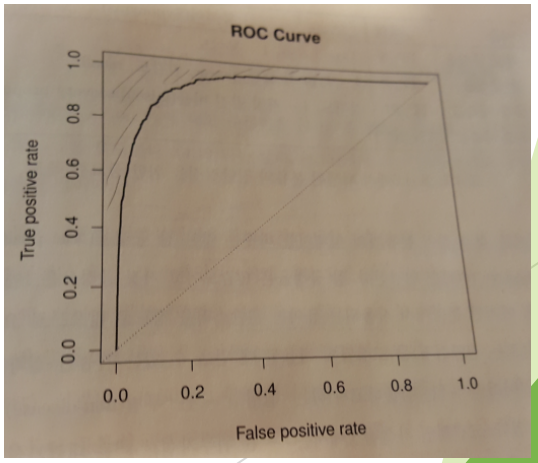
-관측된 설명변수 벡터 x가 주어진 경우에 대해 Y=j일 확률.

ConfusionMatrix (혼돈행렬)



- FP, FN의 두가지 유형의 오류를 범할 수 있다.
- 임계치를 다른 값으로 조정하여 TP FP FN TN의 비율을 조정할 수 있다.
- 임계치에 따른 두 오류의 trade-off를 볼 수 있는 plot





선형판별분석-LDA (linear discriminant analysis)

- 실제환경에서는 베이즈 분류기를 사용할 수 없다. X가 실제로 각 클래스 내의 어떠한 분포를 따른다는 가정이 확실하더라도 파라미터(m(뮤),sigma^2)을 추정해야한다.

- 선형 판별분석 방법은 파이, m(뮤) 그리고 sigma^2에 대한 추정 값을 아래 식에 대입하여 베이즈 분류기를 근사한다.

$$\delta_k(x) = x \cdot \frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

사전분포가 가우스모델(가정)일 경우, LDA는 모든 분류기 중에서 총오류률이 가장 낮은 베이즈 분류기에 근접하고자 한다.

이차판별분석-QDA (Quadratic discriminant analysis)

- 훈련셋이 아주 커 분류기의 분산이 주요 우려사항이 아니거나 K개의 클래스들이 공통의 공분산행렬을 갖는다는 가정이 명백히 맞지 않을 경우에 QDA를 사용한다.

- QDA는 각 클래스의 관측치들이 가우스분포를 따른다고 가정하고, 파라미터들에 대한 추정치를 베이즈 정리에 대입하여 예측한다. 단, 각 클래스가 자체 공분산 행렬을 가진다고 가정한다.