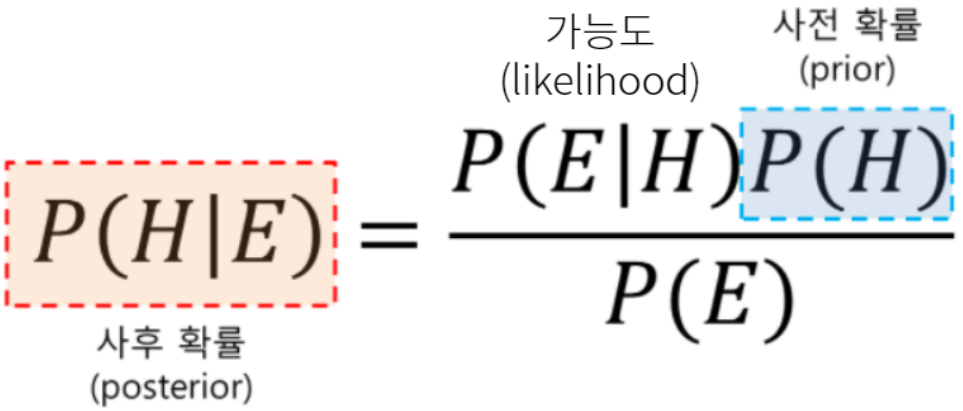
[ISL]4.Classification(2)

태그

선형판별분석(LDA)과 이차판별분석(QDA)은 통계학과 머신러닝에서 널리 사용되는 분류 기법이다. 두 방법 모두 베이즈 정리를 기반으로 하며, 각 클래스의 관측치가 가우시안 분포를 따른다고 가정한다. 그러나 LDA는 공분산 구조의 차이가 심하게 난다면 실행할 수 없다.

▼ 💡 베이즈 정리

베이즈 정리는 새로운 정보를 토대로 어떤 사건이 발생했다는 주장에 대한 신뢰도를 갱신해 나가는 방법, 사전 확률과 조건부 확률을 토대로 사후 확률을 추론하는 과정이다.



H: Hypothesis 어떤 사건이 발생했다는 주장
E: Evidence 새로운 정보

확률	의미
사전확률(Prior)	어떤 사건이 발생한 확률
가능도(Likelihood)	사건이 발생했다는 가정 하에 새로운 정보가 관측될 확률
사후확률(Posterior)	새로운 정보에 의해 갱신된 사건이 발생할 확률

4.4.3 QDA

독립변수 x가 실수이고, 확률분포가 다변량 정규분포라고 가정, => 클래스별로 다른 공분산 구조를 갖는다.

각 클래스의 관측치는 다변량 가우시안 분포를 따름

각 클래스마다 고유한 공분산 행렬(Σk) 사용

각 클래스별로 고유한 평균 벡터(μk) 존재

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k \\ &= -\frac{1}{2}x^T\mathbf{\Sigma}_k^{-1}x + x^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log\pi_k\end{aligned}$$

QDA Classifier 역시 추정값들을 위 식에 대입해 최댓값을 갖는 관측치 X=x를 할당한다. 위의 식을 보면 Σk가 클래스별로 다르고 x에 대한 최고차항이 2차항($x^T\Sigma k^{-1}x$)이므로 따라서 이차 결정경계를 만듦

"Quadratic"이란 이름은 QDA의 판별함수가 이차함수 형태이기 때문에 붙은 것

그렇다면 왜 두가지 방법을 제시할까? **편향 분산 트레이드오프때문**

편향 (Bias): 모델이 얼마나 단순화되어 있는지

- 높은 편향 = 너무 단순한 모델 = 중요한 패턴을 놓침
- 낮은 편향 = 복잡한 모델 = 더 많은 패턴을 포착 가능

분산 (Variance): 모델 예측이 얼마나 불안정한지

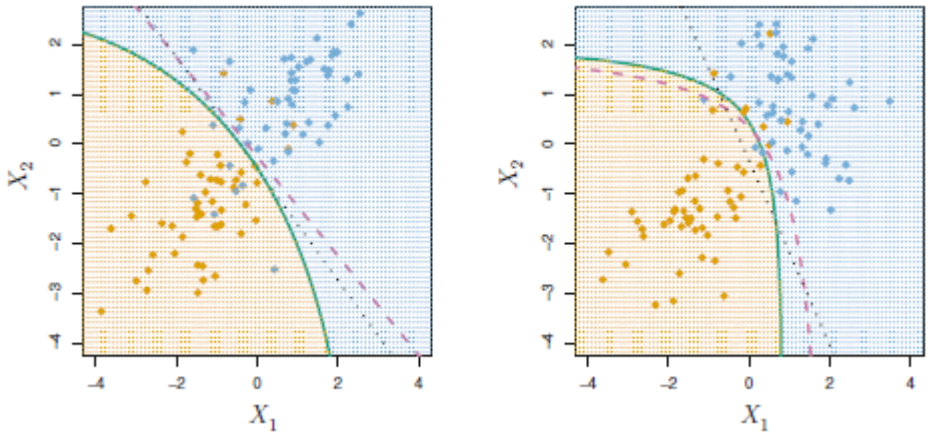
- 높은 분산 = 데이터 변화에 민감 = 예측이 불안정
- 낮은 분산 = 데이터 변화에 안정적 = 예측이 일관적

LDA	p(p+1)/2	분산 낮음, 편향 높음
QDA	kp(p+1)/2	분산 높음, 편향 낮음

QDA: 클래스별로 다른 공분산 행렬을 추정/ 더 많은 파라미터를 추정해야 함⇒ 파라미터 수 ↑

언제 어떤 방법을 사용해야 할까?

- 훈련 데이터가 충분히 많을 때 → QDA 선호
  - 높은 분산을 감당할 수 있음
  - 낮은 편향의 이점을 활용 가능
- 훈련 데이터가 적을 때 → LDA 선호
  - 낮은 분산이 더 중요
  - 높은 편향을 감수하더라도 안정적인 예측 선호



보라색 : 베이즈 경계

검은색 : LDA

초록색 : QDA

왼쪽의 그림은 정규분포를 따르는 두 개의 클래스인데, 주황색과 파란색 클래스 모두 X1과 X2 사이의 상관계수가 0.7로 공통, 베이즈 결정경계가 선형, 이 경우는 공통 분산을 가정하는 LDA를 쓰는 것이 더 좋다.

실제로 LDA는 베이즈 경계와 상당히 비슷하다는 것을 알 수 있다.

오른쪽의 그림은 마찬가지로 정규분포를 따르는 두 개의 클래스인데, X1과 X2 사이의 상관계수가 다르다.

주황색 클래스의 상관계수는 0.7, 파란색 클래스의 상관계수는 -0.7이다.

베이즈 결정경계가 2차형인걸 볼 수 있다.

이 경우는 각 클래스 별로 다른 분산을 가지고 있으므로 QDA가 더 좋다.

4.4.4 Naive Bayes

나이브 베이즈 분류는 베이즈 정리에 기반한 통계적 분류 기법

지금까지 사후 확률 Pk(x)를 π1,...,πK와 f1(x),...,fK(x)로 표현했는데, 실제로 이 공식을 사용하려면 π1,...,πK와 f1(x),...,fK(x)의 추정값이 필요하다.

사전 확률 π1,...,πK 를 추정하는 것은 K번째 클래스에 속하는 훈련 샘플의 비율로 간단하게 추정할 수 있지만, f1(x),...,fK(x)를 추정하는 것은 매우 복잡한 일이다.

- 사전확률: "보기 전에 추측한 확률"
- 사후확률: "본 후에 다시 계산한 확률"

사후확률을 위해서는 사전확률과 밀도함수가 필요하다.

1) 사전확률 추정

전체 이메일 1000개 중:

- 스팸 메일 300개
- 정상 메일 700개

따라서:

- 스팸 사전확률(π1) = 300/1000 = 0.3
- 정상 사전확률(π2) = 700/1000 = 0.7

=> 비교적 쉬움

2) 밀도함수 추정

스팸 메일의 특성 분포:

- 텍스트 길이는 어떤 분포를 따르는지
- 특정 단어들의 등장 확률 분포
- 이미지 수의 분포 - 링크 수의 분포
- 이러한 모든 특성들 간의 관계

=> 많은 변수들의 복잡한 관계를 고려해야하기 때문에 어렵다.

그래서 LDA와 QDA에서는 정규분포라는 가정을 세웠지만 나이브 베이즈는 좀더 강력하게 "각 클래스 내에서 예측변수들이 서로 독립적이다"라는 가정을 새운다.

fk(x) = fk1(x1) × fk2(x2) × ... × fkp(xp) (여기서 fkj는 k번째 클래스에서 j번째 예측변수의 밀도함수이다)

기본적으로 p-차원 밀도 함수를 추정하는 것이 어려운 이유는 각 예측 변수의 **주변 분포**뿐만 아니라 예측 변수 간의 **결합 분포**까지 고려해야 하기 때문이다. 하지만 우리는 예측 변수들이 서로 독립적이라고 가정했다.

각 클래스 내에서 p개의 변수들이 독립적이라고 가정함으로 변수들 간의 관계를 고려할 필요가 없어진다.

why=? "관계가 없다"고 가정했기 때문

그러나 현실적으로 정말 변수들이 독립적일 수 는 없다. => 그렇다면 어떤 경우에 좋은 성능을 보일까

1. 데이터가 적을 때 (데이터 수(n)가 변수 수(p)에 비해 충분히 크지 않을 때)

2. 각 클래스 내 예측변수들의 결합분포를 효과적으로 추정하기 어려울 때

Naive Bayes의 실제 구현 방법

밀도함수 추정방법

A. 양적 변수의 경우

1. 정규분포 가정방법

- $X_j|Y = k \sim N(\mu_{jk}, \sigma^2_{jk})$ - 각 클래스 내에서 변수가 정규분포를 따른다고 가정

- QDA와 비슷해 보이지만, 예측변수들이 독립적이라고 가정

- 공분산 행렬이 대각행렬이라는 점이 다름

2. 비모수 가정방법

- 히스토그램 사용: 각 클래스 내에서 변수의 히스토그램 생성 ,같은 구간에 속하는 관측치의 비율로 확률 추정

- 커널 밀도 추정: 히스토그램을 부드럽게 만든 버전

B. 질적(Qualitative) 변수의 경우

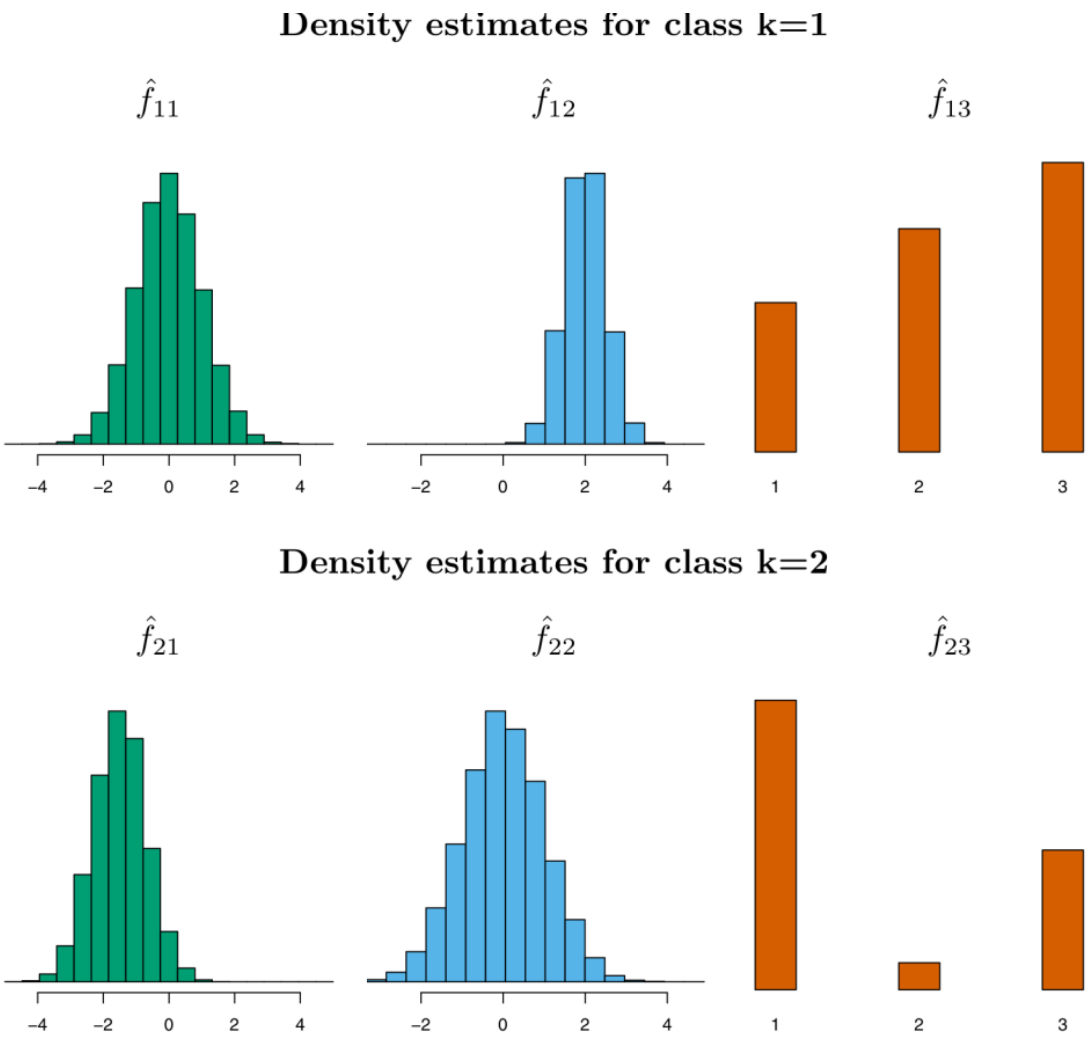
단순히 각 범주의 비율을 계산

- 100개 관측치 중
- 값이 1인 경우: 32개 → 0.32
- 값이 2인 경우: 55개 → 0.55
- 값이 3인 경우: 13개 → 0.13

다음은 p=3, k=2인 naive Bayes classifier의 예시로

이때 마지막 predictor만 qualitative이고 π1, π2의 estimator는 0.5로 같다고 가정한다

fkj를 추정한 결과가 다음과 같다고 하자



quantitative 변수의 분포인 첫번째, 두번째 그림은 histogram이고

qualitative 변수의 분포는 각 클래스별로 predictor의 비율을 나타낸 것임을 확인할 수 있다

이때 새로운 x* = (0.4, 1.5, 1)^T 이 들어오면 다음과 같은 과정을 거쳐 분류

Let, $\hat{f}_{11}(0.4) = 0.368$, $\hat{f}_{12}(1.5) = 0.484$, $\hat{f}_{13}(1) = 0.226$

$\hat{f}_{21}(0.4) = 0.030$, $\hat{f}_{22}(1.5) = 0.130$, $\hat{f}_{23}(1) = 0.616$

$$P(Y=1 \mid X=x^*) = \frac{0.5 \times 0.368 \times 0.484 \times 0.226}{0.5 \times 0.368 \times 0.484 \times 0.226 + 0.5 \times 0.030 \times 0.130 \times 0.616} = 0.944$$

$$P(Y=0 \mid X=x^*) = \frac{0.5 \times 0.030 \times 0.130 \times 0.616}{0.5 \times 0.368 \times 0.484 \times 0.226 + 0.5 \times 0.030 \times 0.130 \times 0.616} = 0.056$$

naive Bayes에 의하면 x*는 Y=1로 분류할 수 있다

		True default status		
		No	Yes	Total
Predicted default status	No	9615	241	9856
	Yes	52	92	144
	Total	9667	333	10000

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

[threshold = 0.5] 좌측 : naive Bayes / 우측 : LDA

		True default status		
		No	Yes	Total
Predicted default status	No	9320	128	9448
	Yes	347	205	552
	Total	9667	333	10000

		True default status		
		No	Yes	Total
Predicted default status	No	9432	138	9570
	Yes	235	195	430
	Total	9667	333	10000

[threshold = 0.2] 좌측 : naive Bayes / 우측 : LDA

첫번째 사진에서 2종 오류(241명)가 1종 오류(52명)

threshold가 낮아지면서 2종오류가 감소하고 1종오류가 증가하는 것은 naive Bayes, LDA 모두 마찬가지이고

naive Bayes가 **LDA보다 더 낮은 2종오류를 보이는 것**을 알 수 있다

- LDA가 전체 오류율은 더 낮음
- 나이브 베이즈는 실제 채무불이행자를 더 잘 예측함

n=10,000, p=2로 데이터가 충분히 커서 LDA와 비교해 큰 이점이 없었음

나이브 베이즈가 더 효과적일 수 있는 상황

- 특성변수의 수(p)가 더 많을 때
- 표본 크기(n)가 더 작을 때

- 이러한 경우에는 분산 감소(variance reduction)가 더 중요한 이점이 됨

▼ 추가 내용

💡가우시안 나이브 베이즈 (GaussianNB)

- **연속적**인 어떤 데이터에도 적용 가능
- 매우 고차원적인 데이터셋에 많이 사용
- 특성치들이 정규분포를 따른다는 가정 하에 조건부확률을 계산

💡 베르누이 나이브 베이즈 (BernoulliNB)

- 특성치의 출현 여부 = **0과 1 이진 데이터**에 적용
- 모델의 복잡도 조절하는 매개변수 alpha 존재alpha 크면 모델이 복잡도 낮아짐 alpha에 따른 알고리즘 성능 변동은 비교적 크지 않아서 성능 향상에 크게 기여한다고는 할 수 없지만, 정확도를 높일 수는 있다.

💡 다항분포 나이브 베이즈 (MultinomialNB)

- **특성치의 개수**를 활용한 분석 (예를 들면, 문장에 나타난 단어의 횟수)
- 0이 아닌 특성이 비교적 많은 데이터셋에서 베르누이보다 성능이 높다.
- 모델의 복잡도 조절하는 매개변수 alpha 존재 (베르누이와 동일)

💡나이브 베이즈의 0 확률 문제 해결 방안 : 라플라스 스무딩

확률을 계산하는 과정에서 발생할 수 있는 0 확률 문제를 해결하기 위한 방법 중 하나로

라플라스 스무딩은 각 클래스에서 각 특징의 발생 확률을 계산할 때, 해당 특징이 한 번도 발생하지 않아 0이 되는 경우를 방지하기 위해 사용된다. 즉, 학습 데이터에 없는 신규 데이터는 조건부 확률이 0이므로 분류하지 못하기 때문에 라플라스 스무딩 기법으로 보정하여 분류하는것

처음 보는 것도 일어날 수 있다"는 가능성을 열어두는 것