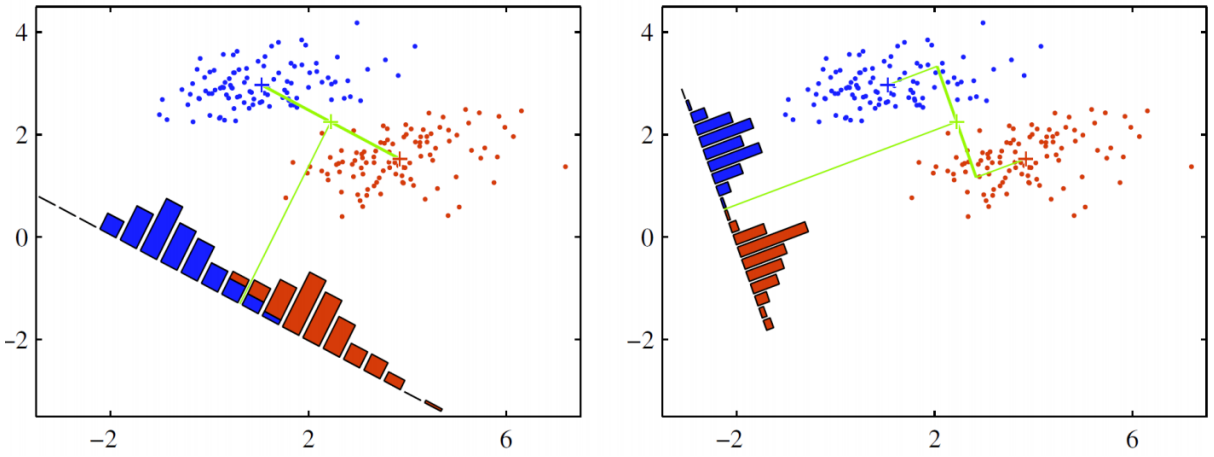
LDA-Bayesian Perspective



LDA는 위 그림처럼 두 데이터를 잘 구분할 수 있는 **직선**을 찾는 걸 목표로 한다.

그러기 위해선 두 클래스의 평균이 서로 멀도록 해야되며, 클래스 간 분산은 크도록, 클래스 내 분산은 작도록 하는 방향으로 직선을 찾아야 한다.

이 직선을 선형 결정 경계라 칭한다.

PRML 4.2 에 따르면, 확률적 관점으로 전환하여 데이터의 분포에 대한 간단한 가정에서 선형 결정 경계가 어떻게 생기는지 보여준다.

이를 위해서 k개의 클래스를 가진 C_k 가 있다고 가정한다. x는 새로운 데이터라 가정한다.

클래스 조건부 밀도 $P(x|C_k)$와 클래스 사전 확률 $P(C_k)$을 모델링하고, 베이즈 정리를 통해 사후 확률 $P(C_k|x)$을 계산한다.

여기서 사후확률 $P(C_k|x)$는 '새로운 데이터 x가 주어졌을 때, 이 데이터가 C_k클래스에 속할 확률이다.'

클래스가 2개 즉, k = 2일때를 예시로 들어보자, 클래스가 2개이기 때문에 $p(C_1|x)$ 또는 $p(C_2|x)$ 하나 구하면 된다.

그렇다면 $P(C_1|x)$를 구해보자

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x)}$$
$$= \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

여기서 로짓변환을 사용한다.

$p(C_1|x)$의 분자와 분모에 로그를 취하고 이를 비율로 나타내면, 로그 오즈(log odds) 또는 로짓(logit) 변환이라고 하는 것을 얻을 수 있다.

$$a = \ln\left(\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}\right)$$

여기서 a가 0보다 크다면 $P(C_1|x)$가 $P(C_2|x)$ 보다 크다. 따라서 새로운 데이터 x는 C_1 클래스에 속하게 된다. 그 증명과정을 아래에 서술하겠다.

즉, a는 베이지안 관점에서 결정경계가 된다.

$$a > 0$$
$$\ln\left(\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}\right) > 0$$
로그 함수의 성질에 의해
$$\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} > 1$$
$$P(x|C_1)P(C_1) > P(x|C_2)P(C_2)$$
여기서 양변에 $P(x)$를 나누어 준다면
$$P(C_1|x) > P(C_2|x)$$
가 된다.

a가 0이라면 $P(C_1|x) = P(C_2|x)$가 되므로 임의로 설정을 해줘야 한다.

이제 조금 더 일반화 하기 위해 $a = \ln\left(\frac{p(x|C_1)p(C_1)}{p(x|C_2)p(C_2)}\right)$식을 변형해 보겠다.

$$a = \ln\left(\frac{p(x|C_1)\,p(C_1)}{p(x|C_2)\,p(C_2)}\right)$$

$$e^{a} = \frac{p(x|C_1)\,p(C_1)}{p(x|C_2)\,p(C_2)}$$

$$e^{-a} = \frac{p(x|C_2)\,p(C_2)}{p(x|C_1)\,p(C_1)}$$

$$1+e^{-a} = \frac{p(x|C_1)\,p(C_1) + p(x|C_2)\,p(C_2)}{p(x|C_1)\,p(C_1)}$$

$$\frac{1}{1+e^{-a}} = \frac{p(x|C_1)\,p(C_1)}{p(x|C_1)\,p(C_1) + p(x|C_2)\,p(C_2)}$$

$$= \frac{p(x|C_1)\,p(C_1)}{p(x)} = p(C_1|x)$$

$$a = \ln\left(\frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}\right)$$

$$e^{a} = \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

여기서 분자와 분모를 뒤집어준다면

$$e^{-a} = \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}$$

$$1 + e^{-a} = \frac{P(x|C_2)P(C_2) + P(x|C_1)P(C_1)}{P(x|C_1)P(C_1)}$$

$$\frac{1}{1 + e^{-a}} = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)} = P(C_1|x)$$

$\frac{1}{1+e^{-a}}$는 시그모이드 함수 $\sigma(a)$가 된다. 시그모이드 함수는 입력 a가 0보다 크면 함수값이 0.5보다 크고 0보다 작으면 0.5보다 작아진다. 그리고 함수값의 범위는 0부터 1까지 이다.

즉, 새로운 데이터 x가 클래스 C_1에 속할 사확률인 $P(C_1|x)$가 시그모이드 함수의 형태로 바뀌지며, 새로운 데이터 x를 넣었을때 그 함수값이 0.5보다 크다면 그 새로운 데이터 x는 클래스 C_1에 속하게 된다. 즉, 확률의 형태로 바뀐다.

LDA with bayesian perspective

$$P(C=k \mid X=x) = \frac{P(X=x \mid C=k)\,P(C=k)}{P(X=x)} = \frac{f_k(x)\,\pi_k}{\sum_{i=1}^{K} f_i(x)\,\pi_i}$$

$f_k(x)$는 특정 class k에 속할 X의 확률을 알려주는 함수

π_k는 사전확률

$f_k(x)$ 다변량 정규 분포

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}} |\Sigma_k|^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1} (x-\mu_k)\right]$$

$$a = \ln \frac{P(C=k \mid X=x)}{P(C=l \mid X=x)} = \ln \frac{P(X=x \mid C=k)\,P(C=k)}{P(X=x \mid C=l)\,P(C=l)}$$

$$= \ln \frac{f_k(x)}{f_l(x)} + \ln \frac{\pi_k}{\pi_l}$$

Σ 공분산 행렬 모두 동일

$$\Sigma_k = \Sigma_l = \Sigma$$

$$\ln \frac{f_k(x)}{f_l(x)} = \ln \frac{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)\right]}{(2\pi)^{\frac{p}{2}} |\Sigma|^{\frac{1}{2}} \exp\left[-\frac{1}{2}(x-\mu_l)^T \Sigma^{-1} (x-\mu_l)\right]}$$

$$= \Sigma^{-1}\left[-\frac{1}{2}(x-\mu_k)^T(x-\mu_k) + \frac{1}{2}(x-\mu_l)^T(x-\mu_l)\right]$$

$$= \Sigma^{-1}\left[-\frac{1}{2}(x^Tx - 2x^T\mu_k + \mu_k^T\mu_k) + \frac{1}{2}(x^Tx - 2x^T\mu_l + \mu_l^T\mu_l)\right]$$

$$= \Sigma^{-1}\left[x^T\mu_k - x^T\mu_l - \frac{1}{2}(\mu_k^T\mu_k - \mu_l^T\mu_l)\right]$$

$$= x^T\Sigma^{-1}(\mu_k - \mu_l) - \frac{1}{2}(\mu_k + \mu_l)^T\Sigma^{-1}(\mu_k - \mu_l)$$

$$\frac{P(C=k \mid X=x)}{P(C=l \mid X=x)} = \ln \frac{\pi_k}{\pi_l} + x^T\Sigma^{-1}(\mu_k - \mu_l) - \frac{1}{2}(\mu_k + \mu_l)^T\Sigma^{-1}(\mu_k - \mu_l)$$

새로운 data X에 대한 선형판별식

$P(C=k \mid X=x) > P(C=l \mid X=x)$ 라면 k범주 할당

특정 class k에 대한 선형판별식 $f_k(x)$

$$f_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln \pi_k$$

$$C(x) = \arg\max_k f_k(x)$$

새로운 data x가 어떤 class에 속하게 될지 가장 확률이 높은지 알 수 있음