나이브베이즈&군집분석

나이브 베이즈에 필요한 배경지식

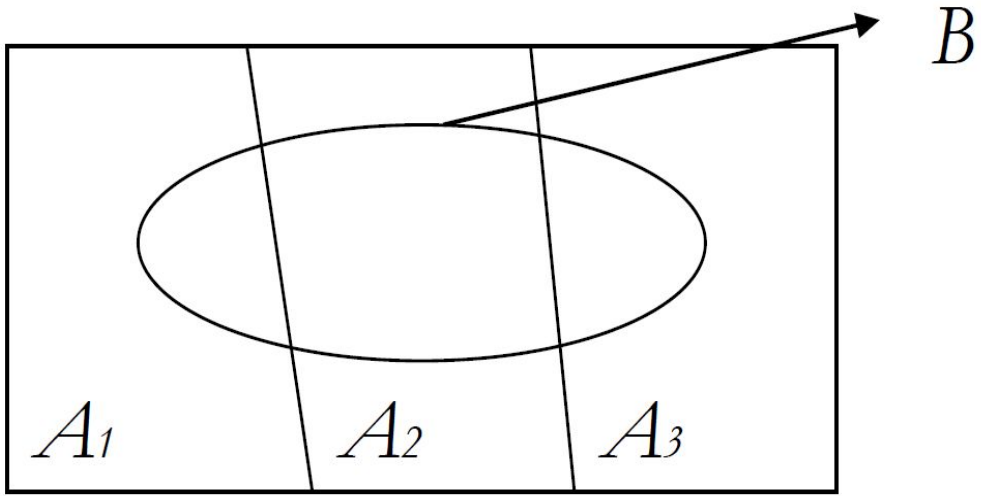
베이즈 정리**(Naive Bayes)**

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

조건부 확률 **P(A|B)**는 **B**가 일어났을때 **A**가 일어날 확률을 말합니다.
베이즈 정리는 **P(A|B)**의 추정이 **P(A&B)**와 **P(B)**의 확률에 의해 정해야 한다는 정리입니다.

베이즈 정리(Naive Bayes)

일반적으로 사건 **A1, A2, A3**가 서로 배반**(mutually exclusive)**이고 **A1, A2, A3**의 합집합이 표본공간**(sample space)**과 같으면 사건 **A1, A2, A3**는 표본공간 **S**의 **분할(disjoint)**이라고 정의합니다. 우리가 관심있는 사건 **B**가 나타날 확률을 그림과 식으로 나타내면 다음과 같습니다.



전확률 공식

전확률 공식은 베이즈 공식이라고도 불리며 다음과 같은 식을 가지고 있습니다.

$$P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3) = \sum_{i=1}^{3} P(A_i)P(B|A_i)$$

여기서 P(An)은 사전확률이라고 물리며 , P(B|An)은 우도(likelihood)라 불립니다

만약 P(A1|B)를 구하고 싶으면

다음과 같은 식으로 구합니다.

P(B)가 일어나고 진행되기 때문에

P(A1|B)는 사후확률이라고 합니다

$$P(A_1|B) = \frac{P(A_1)P(B|A_1)}{P(B)}$$

$$= \frac{P(A_1)P(B|A_1)}{P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + P(A_3)P(B|A_3)}$$

사후확률 구하기 예시

만약 오늘 날씨가 좋고, 바람이 많이 불지 않고,

기압은 높은데,

온도가 낮다면

오늘은 비가 올것인가? 안올것인가?

이 질문을 식으로 바꿔보면.

비 와 변수들간 관계표									
	날씨가 좋은가?		바람이 많이 부는가?		기압이 높은가?		온도가 높은가?		
	Yes	No	Yes	No	Yes	No	Yes	No	계
비 온날	2	6	6	2	8	0	5	3	8
안온날	8	4	2	10	2	10	6	6	12
계	10	10	8	12	10	10	11	9	20

사후확률 구하기 예시

사전 조건에 맞는 경우는 비가오고

안오고에 따라 두가지로 나눌수있다.

(~부정의 의미)

비가 올 확률

$$= \frac{P(비 \mid 날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도)}{P(비 \mid 날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도) + P(\sim 비 \mid 날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도)}$$

구해야 할 것

- $P(비 \mid 날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도)$
- $P(\sim 비 \mid 날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도)$

사후확률 구하기 예시

$$P(\text{비} \mid \text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})$$

$$= \frac{P(\text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도} \mid \text{비})P(\text{비})}{P(\text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})}$$

$$= \frac{P(\text{날씨}|\text{비})\ P(\sim\text{바람}|\text{비})\ P(\text{기압}|\text{비})\ P(\sim\text{온도}|\text{비})P(\text{비})}{P(\text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})}$$

$$\fallingdotseq P(\text{날씨}|\text{비})\ P(\sim\text{바람}|\text{비})\ P(\text{기압}|\text{비})\ P(\sim\text{온도}|\text{비})P(\text{비})$$

비 와 변수들간 관계표

	날씨가 좋은가?		바람이 많이 부는가?		기압이 높은가?		온도가 높은가?		
	Yes	No	Yes	No	Yes	No	Yes	No	계
비 온날	2/8	6/8	6/8	2/8	8/8	0/8	5/8	3/8	8
안온날	8/12	4/12	2/12	10/12	2/12	10/12	6/12	6/12	12
계	10	10	8	12	10	10	11	9	20

사후확률 구하기 예시

$$P(비 \mid 날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도)$$

$$= \frac{P(날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도 \mid 비)P(비)}{P(날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도)}$$

$$= \frac{P(날씨|비)\ P(\sim 바람|비)\ P(기압|비)\ P(\sim 온도|비)P(비)}{P(날씨 \cap \sim 바람 \cap 기압 \cap \sim 온도)}$$

$$\fallingdotseq P(날씨|비)\ P(\sim 바람|비)\ P(기압|비)\ P(\sim 온도|비)P(비)$$

1. $P(날씨|비)\ P(\sim 바람|비)\ P(기압|비)\ P(\sim 온도|비)P(비)$

$= (2/8) * (2/8) * (8/8) * (3/8) * (8/20)$

$= 0.009375$

2. $P(날씨|\sim 비)\ P(\sim 바람|\sim 비)\ P(기압|\sim 비)\ P(\sim 온도|\sim 비)P(\sim 비)$

$= (8/12) * (10/12) * (2/12) * (6/12) * (12/20)$

$= 0.33333$

사후확률 구하기 예시

비가 올 확률

$$= \frac{P(\text{비} \mid \text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})}{P(\text{비} \mid \text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도}) + P(\sim\text{비} \mid \text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})}$$

$$= \frac{0.009375}{0.009375+0.333333} = 0.027 \rightarrow 2.7\%$$

비가 안 올 확률

$$= \frac{P(\sim\text{비} \mid \text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})}{P(\text{비} \mid \text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도}) + P(\sim\text{비} \mid \text{날씨} \cap \sim\text{바람} \cap \text{기압} \cap \sim\text{온도})}$$

$$= \frac{0.333333}{0.009375+0.333333} = 0.973 \rightarrow 97.3\%$$

사후확률 구하기

이렇게 복잡한 경우를 거쳐 사후 확률을 구하는 이유

보통의 문제에서는 사후확률을 아는 경우는 거의 없고 실험을 통한 사전 확률만을 알기 때문에 이러한 방식으로 사후확률은 구한다.

나이브 베이즈 모델 예시

문서 이진분류 문제를 예로 들어보겠습니다.

우리가 풀려는 문제는 문서 d가 주어졌을 때 범주 c_1 혹은 c_2로 분류하는 것입니다. 지금까지 설명한 베이즈 법칙을 다시 쓰면 아래와 같습니다.

$$P(c_1|d) = \frac{P(c_1,d)}{P(d)} = \frac{\frac{P(c_1,d)}{P(c_1)} \cdot P(c_1)}{P(d)} = \frac{P(d|c_1)P(c_1)}{P(d)}$$

$$P(c_2|d) = \frac{P(d|c_2)P(c_2)}{P(d)}$$

나이브 베이즈 모델 예시

P(Ci):사전확률 **->** 문서전체에서 주제 **ci**가 나올 확률

p(d|Ci):우도**(likelihood),**사전확률**(ci**가 주제일때 문서가 **d**일 경우**)**

p(Ci|d):사후확률**(**문서 **d**가 주어졌을때 **ci**에 속하는 값**)**

베이즈 모델의 기준 **P(c1|d)/P(d)**와 **P(c2|d)/P(d)**를 기준으로 큰쪽으로 **c1** 일지 **c2**일지 구합니다**.**

결론적으로 사후확률 **p(Ci|d)**는 **P(ci|d)/P(d)**에 달려있음을 알 수 있습니다**.**

나이브 베이즈 모델 예시

이번에는 문서 **d**가 단어 **w1,w2**로 구성되어있다고 가정하겠습니다.

P(ci|d)=P(ci|w1,w2)로 식을 바꿀 수 있겠습니다. 이러한 상황에서 아까 봤던 식을 사용하면 다음과 같은 식을 얻습니다.

$$P(ci|d)= P(ci|w1,w2) \propto P(w1,w2|ci)P(ci) \propto P(w1,w2|ci)$$

나이브 베이즈 분류기는 각 단어가 독립**(independent)**임을 가정합니다.

모델 이름에 나이브라는 말이 붙은 이유이기도 합니다. 이에 따라 식을 다시 쓸 수 있습니다.

$$P(w1,w2)=P(w1)\cdot P(w2)$$

$$P(w1,w2|ci)=P(w1|ci)\cdot P(w2|ci)$$

나이브 베이즈로 spam메일 분류하기

데이터는 **TYPE** 과 **TEXT** 컬럼으로 구성되어 있으며 총 **5559**개의 **SMS** 를 포함하고 있습니다.

IDX	type	text
1	ham	Hope you are having a good week. Just checking in
2	ham	K..give back my thanks.
3	ham	Am also doing in cbe only. But have to pay.
4	spam	complimentary 4 STAR Ibiza Holiday or 10,000 cash needs your URGENT
5	spam	okmail: Dear Dave this is your final notice to collect your 4* Tenerife Holiday or ·····
~		
5558	spam	SMS. ac JSco: Energy is high, but u may not know where 2channel it. 2day ur
5559	ham	Shall call now dear having food

나이브 베이즈로 spam메일 분류하기

하기위한 핵심 전처리: 숫자 제거, 소문자화, 불용어 제거, 구두점 제거 형태소 분석

idx	check	good	thanks	pay	~
1	1	1	0	0	
2	0	0	1	1	~
3	0	0	0	0	
~					

나이브베이즈 모델 종류

나이브 베이즈 모델의 종류에는 3가지가 있다.

1.가우시안 나이브 베이즈 모델

연속적인 값을 처리할때 전형적으로 각각의 값들이 가우스 분포를 따른다고 가정하고 진행하는 모델.

가우스 분포(정규 분포):평균과 표준편차에 대하여 모양이 결정된다.

$$P(x_d \mid y = k) = \frac{1}{\sqrt{2\pi\sigma_{d,k}^2}} \exp\left(-\frac{(x_d - \mu_{d,k})^2}{2\sigma_{d,k}^2}\right)$$

나이브베이즈 모델 종류

2.베르누이 나이브 베이즈 모델

$$P(x_d \mid y = k) = \mu_{d,k}^{x_d}(1 - \mu_{d,k})^{(1-x_d)}$$

$$P(x_1, \ldots, x_D \mid y = k) = \prod_{d=1}^{D} \mu_{d,k}^{x_d}(1 - \mu_{d,k})^{(1-x_d)}$$

3.다항 분포 가능도 모델

$$P(x_1, \ldots, x_D \mid y = k) \propto \prod_{d=1}^{D} \mu_{d,k}^{x_{d,k}}$$

$$N_k = \sum_{d=1}^{D} x_{d,k}$$

$$\sum_{d=1}^{D} \mu_{d,k} = 1$$

판별 분석

판별 분석은 두개 이상의 모집단에서 추출된 표본들이 지니고 있는 정보들을 이용하여 표본들이 어디에서 추출된건지 기준을 정하는 분석법을 말한다.

예를 들어 은행에서 채무자가 대출금을 갚을것인지 아닌지에 대한 여부를 판단하기 위해

과거에 대출금을 갚지 않은 사람의 정보 유형을 참고하여 담보 신청시 신청자의 정보 유형을 과거의 유형과 비교하여 판단할수 있겠습니다.

판별 변수의 주요 개념

1.판별 변수

판별 변수는 어떤 집단에 속하는지 판단 할때 사용되는 변수로서 독립변수중 판별력이 높은 변수를 뜻합니다.

판별 변수를 선택하는데 중요하게 생각해야 할것은 다른 독립변수들 간의 상관관계입니다.

상관관계가 높은 변수는 전부다 선택하지 않도록하고 상관관계가 적은것 끼리 판단해야 효과적인 판별함수를 만들 수 있습니다.

판별 변수의 주요 개념

2.판별 함수

판별함수는 판별변수들을 선형관계로 모은 것으로 집단의 수와 독립변수의 수중 작은 값만큼 도출할 수 있습니다. 판별 함수의 목적은 종속변수의 집단에 대한 예측력을 높이는데 있다. 판별분석이 이용되기 위해서는 각 개체가 어느 집단에 속해 있는지 알려져 있어야 하고 이에 대한 판별식을 만들어 판단하는 과정을 포함하게 된다.

3.판별점수

판별점수는 어떤 대상이 어떤 집단에 속하는지 판별하기 위하여 그 대상의 판별변수들의 값을 판별함수에 대입하여 구한 값을 뜻한다.

판별 변수의 주요 개념

4,표본의 크기

판별 분석시 전체 표본의 크기는 독립변수의 개수보다 3배이상 되어야 합니다. 종속변수의 집단 각각의 표본의 크기 중 최소 크기가 독립변수의 개수보다 커야 합니다.

판별 분석의 단계

1.케이스가 속한 집단을 구분하는데 기여할수 있는 독립 변수를 찾습니다.

2.집단을 구분하는 기준이 되는 독립 변수들의 선형 결합 즉 판별 함수를 도출합니다

3.도출된 판별 함수에 의한 분류의 정확도를 파악합니다.

4.판별함수를 이용하여 새로운 케이스가 속하는 집단을 예측합니다.

판별 분석의 계산 논리

먼저 판별 변수는 주어진 독립 변수의 특성을 바탕으로 종속 변수의 변화와 판단의 방향을 예측하는 것이기 때문에 가장 중요한것은 독립변수의 선별이 가장 중요합니다.

판별 함수에 대한 공식은 다음과 같습니다.

$$Z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

x1,x2,x3…..: 판별 변수

B1,B2,.....:판별계수

Z:판별 점수

판별 분석의 주요 개념

평균값:판별 분석에서는 일차적으로 종속 변수의 값이 정해져 있는 사례수에 대한 평균값을 계산한다.

윌크스의 람다:종속변수의 변수값을 기준으로 분류된 각 독립변수의 평균값이 어느정도 차이가 나는지 분석하는 통계값

집단내 제곱합 / 전체의 제곱합

독립 변수에 대한 람다 값이 1이면 종속 변수의 평균값이 동일하다는 의미이고 람다값이 작으면 종속 변수의 평균값 차이가 크다고 해석한다.

결과적으로 람다값이 작아야 예측력이 높다고 생각한다

판별 분석은 평균값이 차이가 많이나야 독립변수에 대한 예측을 충분히 할수 있기 때문이다

판별 분석의 주요 개념

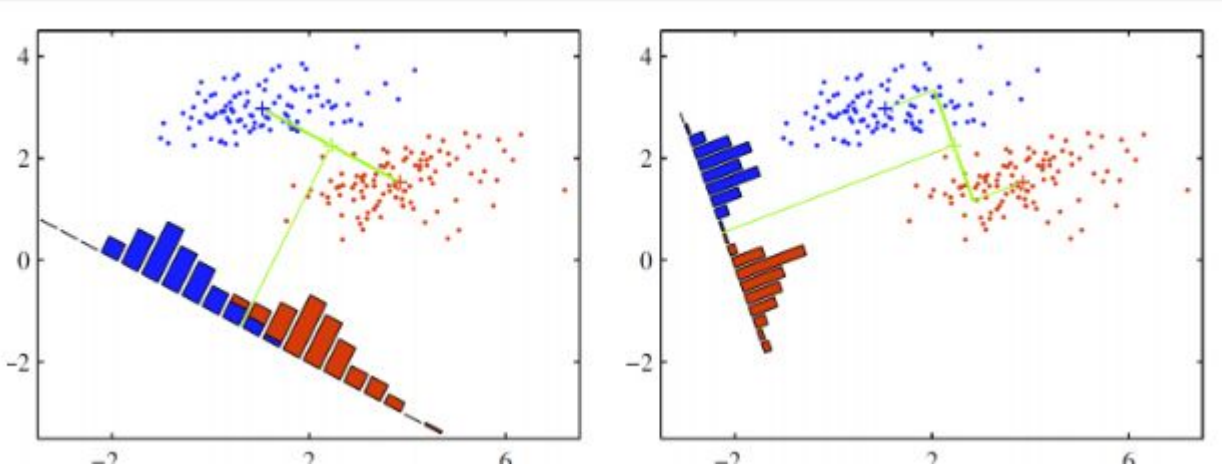
고유값과 에타값:고유값은 판별함수가 어느 정도로 설명력이 있는지에 대한 내용을 분석하는데 사용되는 통계값, 공식은 집단간 제곱합/집단내 제곱합입니다. 고유값이 크면 판별함수의 설명력이 높다고 할수 있겠습니다.

에타는 판별 점수와 종속변수 사이의 상관관계를 나타내는 통계값입니다.

에타값은**0**에서 **1**사이에 위치하는데 이 값이 높으면 판별함수와 종속변수 사이의 상관관계가 밀접하기 때문에 판별함수의 설명력이 높아집니다.

판별 함수 예시

선형판별분석 (Linear Discriminant Analysis)



선형판별분석(Linear Discriminant Analysis)

그렇다면 두 범주를 잘 구분할 수 있는 직선은 어떤 성질을 지녀야 할까요?

사영 후 두 범주의 중심(평균)이 서로 멀도록, 그 분산이 작도록 해야할 겁니다. 왼쪽 그림을 오른쪽과 비교해서 보면 왼쪽 그림은 사영 후 두 범주 중심이 가깝고,

분산은 커서 데이터가 서로 잘 분류가 안되고 있는 걸 볼 수가 있습니다.

반대로 오른쪽 그림은 사영 후 두 범주 중심이 멀고, 분산은 작아서 분류가 비교적 잘 되고 있죠. **LDA**는 바로 이런 직선을 찾도록 해줍니다.

선형판별분석(Linear Discriminant Analysis)

$$y = \overrightarrow{w}^T \overrightarrow{x}$$

$$m_1 = \frac{1}{N_1} \sum_{n \in C_1} x_n$$

$$m_2 = \frac{1}{N_2} \sum_{n \in C_2} x_n$$

p차원의 입력벡터 **x**(변수 **p**개)를 **w**라는 벡터(축)에 사영시킨 후 생성되는 **1**차원상의 좌표값(스칼라)를 아래와 같이 **y**라고 정의합니다. 각각 **N1**개와 **N2**개의 관측치를 갖는 **C1**과 **C2** 두 범주에 대해 원래 입력공간(**2**차원)에서 각 범주의 중심(평균) 벡터도 아래와 같이 **m1, m2**라고 정의합니다.

선형판별분석(Linear Discriminant Analysis)

첫번째로 사영후 두 범주의 중심이 멀리 떨어지도록 하는 벡터를 찾아야합니다

m2−m1= wT(m2−m1)mk =wTmk

사영후 각 범주안에 속하는 집단안의 분산은 작을수록 좋습니다. 분산을 구하는 식은 다음과 같습니다

$$s_k^2 = \sum_{n \in C_k} (y_n - m_k)^2$$

선형판별분석(Linear Discriminant Analysis)

앞에 두 과정을 동시에 진행 할려면 다음과 같은 과정을 거칩니다.

두 범주 중심을 분자, 두 범주의 분산을 분모에 넣고 이 식을 최대화

$$J(w) = \frac{(m_1 - m_2)^2}{s_1^2 + s_2^2} = \frac{w^T S_B w}{w^T S_W w}$$

$$S_B = (m_1 - m_2)(m_1 - m_2)^T$$

$$S_W = \sum_{n \in C_1} (x_n - m_1)(x_n - m_1)^T + \sum_{n \in C_2} (x_n - m_2)(x_n - m_2)^T$$

선형판별분석(Linear Discriminant Analysis)

목적함수 J(w)은 w에 대해 미분한 값이 0이 되는 지점에서 최대값을 가집니다. 아래 식과 같습니다.

$$(w^T S_B w) S_W w = (w^T S_W w) S_B w$$

약간의 식을 정리하면 다음과 같은 식이 나옵니다

$$S_W w = \lambda S_B w$$
$$S_B{}^{-1} S_W w = \lambda w$$

새로운 축 w는 SB의 역행렬과 SW를 내적한 행렬의 고유벡터라는 이야기 입니다.

선형판별분석(Linear Discriminant Analysis)

새로운 데이터($\mathbf{x'}$)가 주어지면 이를 $\mathbf{w}$와 내적해 각각의 스코어를 낼 수 있습니다. 그 스코어가 일정값보다 크면 $\mathbf{C_1}$범주, 작으면 $\mathbf{C_2}$ 범주로 분류를 하게 됩니다.