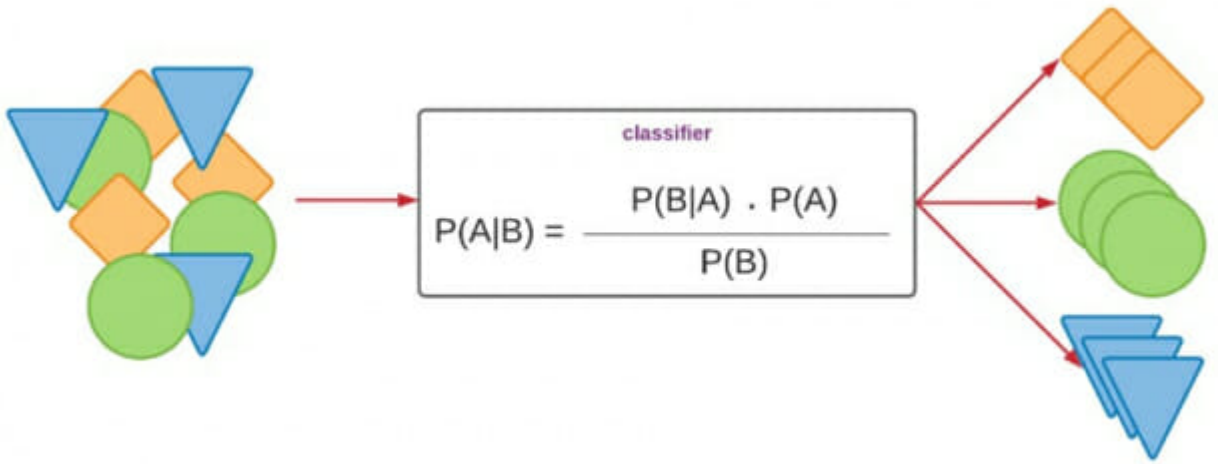
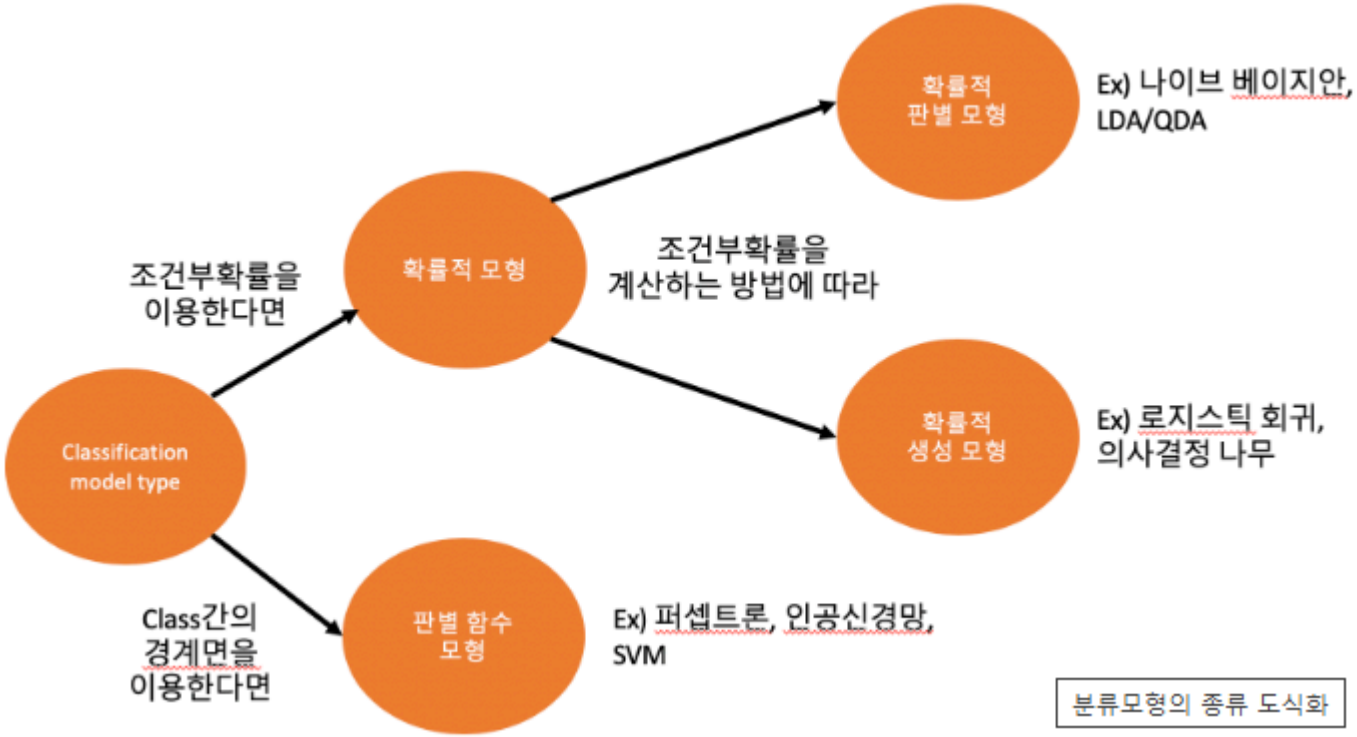
나이브 베이즈 정리(Naive Bayes Theorem)-추론

나이브 베이지 분류(Naive Bayesian Classification)란?

- 데이터의 특징을 가지고 각 클래스(레이블)에 속할 확률을 계산하는 조건부 확률 기반의 분류 방법
- 데이터의 특징이 모두 상호 독립적이라는 가정하에 확률 계산을 단순화 ⇒ 나이브(naïve)하다
- Bayes Theorem에 의해 데이터의 특징을 통해 클래스 전체의 확률 분포 대비 특정 클래스에 속할 확률을 구하는 것
- 나이브 베이지 분류를 통해 데이터 특징이 하나 이상일 때 나이브 베이지 공식으로 해당 데이터가 어떤 레이블에 속할 확률이 가장 높은지를 알 수 있음




확률적
판별 모형
Ex) 나이브 베이지안,
LDA/QDA
조건부확률을
이용한다면
확률적 모형
조건부확률을
계산하는 방법에 따라
Classification
model type
확률적
생성 모형
Ex) 로지스틱 회귀,
의사결정 나무
Class간의
경계면을
이용한다면
판별 함수
모형
Ex) 퍼셉트론, 인공신경망,
SVM
분류모형의 종류 도식화

---

1) 결합 확률

$$P(X,Y) = P(X \mid Y) \cdot P(Y)$$

두 가지 이상의 사건이 동시에 발생하는 확률

$$P(X,Y) = P(X \mid Y) \cdot P(Y) = \frac{P(X \cap Y)}{P(Y)} \cdot P(Y) = P(X \cap Y) = P(X) \cdot P(Y)$$

나이브 베이즈 알고리즘에서는 X, Y 사건이 독립이라고 가정한다.

---

2) 결합확률과 베이지 정리

$$P(A \mid X,Y) = \frac{P(X,Y \mid A) \cdot P(A)}{P(X,Y)} \; (by\ Bayes'\ Theorem)$$

P(A)	데이터가 레이블 A에 속할 확률
P(X, Y | A)	레이블이 A인 데이터가 X, Y 특징을 모두 가질 확률
P(X, Y)	데이터가 X, Y 특징을 모두 가질 확률
P(A | X, Y)	X, Y 특징을 모두 갖는 데이터가 레이블 A에 속할 확률

$$P(X,Y) = P(X) \cdot P(Y) \qquad P(X,Y \mid A) = P(X \mid A) \cdot P(Y \mid A)$$

어떤 한 데이터가 각 특징을 갖는 사건끼리는 서로 독립이라고 베이즈 정리에서 전제한다.

레이블이 A인 데이터가 특징 X를 가질 사건과 특징 Y를 가질 사건은 서로 독립이다.

$$P(A \mid X,Y) = \frac{P(X \mid A) \cdot P(Y \mid A) \cdot P(A)}{P(X) \cdot P(Y)}$$

★ 데이터가 각 특징을 갖는 사건끼리 독립인 것이지, 특징과 레이블이 서로 독립이라는 의미가 아니다!

---

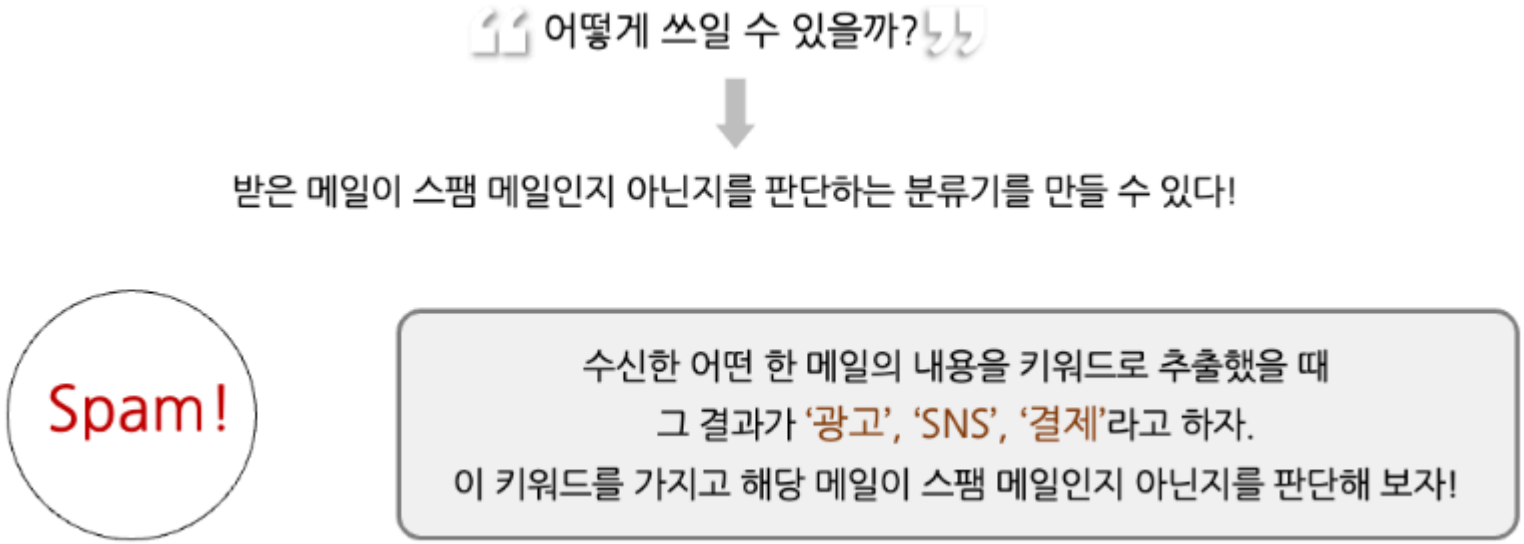
3) 나이브 베이즈 분류기의 원리

- 나이브 베이즈 분류기는 베이즈 정리를 사용하여 주어진 데이터를 특정 클래스에 속할 확률을 계산하고, 가장 높은 확률을 가진 클래스를 선택하는 방식
  - 이때, 모든 특성(feature)이 서로 독립적이라고 가정하며, 이러한 가정이 '나이브'라는 이름에 기인함
- 분류 과정
  - 각 클래스의 사전 확률 P(C)를 계산
  - 각 특성별로 조건부 확률 P(x_i|C)를 계산

- 테스트 데이터의 모든 특성에 대해 조건부 확률을 곱하여 P(x|C)를 구함
- P(x|C)와 P(C)를 곱하여 사후 확률 P(C|x)를 구함
- 가장 높은 확률을 가진 클래스를 선택

---

4) 나이브 베이지 분류 적용 예시



광고(AD), SNS(SNS), 결제(PAY) 키워드를 갖는 메일이 스팸(SPAM)일 확률을 구한다.

★ 단, 메일이 광고, SNS, 결제 라는 각 키워드를 가질 사건은 서로 독립이라고 가정한다.

$$P(SPAM \mid AD, SNS, PAY) = \frac{P(AD \mid SPAM) \cdot P(SNS \mid SPAM) \cdot P(PAY \mid SPAM) \cdot P(SPAM)}{P(AD) \cdot P(SNS) \cdot P(PAY)}$$

P(AD | SPAM)	스팸 메일이 광고라는 키워드를 가질 확률	P(AD)	모든 메일 중 광고라는 키워드를 갖는 메일의 비율
P(SNS | SPAM)	스팸 메일이 SNS라는 키워드를 가질 확률	P(SNS)	모든 메일 중 SNS라는 키워드를 갖는 메일의 비율
P(PAY | SPAM)	스팸 메일이 결제라는 키워드를 가질 확률	P(PAY)	모든 메일 중 PAY라는 키워드를 갖는 메일의 비율

P(SPAM | AD, SNS, PAY) → 광고, SNS, 결제 키워드를 갖는 메일이 스팸일 확률

스팸으로 분류될 확률이 임계치 이상이면 해당 메일을 스팸으로 분류한다!

스팸 메일 분류(이진 분류)에 적용

- 어떤 문장이 주어졌을 때 스팸 메일일 확률은 P(스팸 메일|문장)로, 정상 메일일 확률은 P(정상 메일|문장)로 표기할 수 있음
- 메일의 모든 문장이 2개의 단어로 구성되었다고 가정했을 때 나이브 베이지 분류에서
  - 스팸 메일일 확률
    - P(스팸 메일|문장)= P(단어1|스팸 메일)*P(단어2|스팸 메일)*P(스팸 메일)
  - 정상 메일일 확률
    - P(정상 메일|문장)= P(단어1|정상 메일)*P(단어2|정상 메일)*P(정상 메일)

```
#라이브러리 불러오기
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
```

```
#사이킷런에서 문서 데이터 세트를 가져옴
news = datasets.fetch_20newsgroups()
```

```
#입력 데이터와 타깃을 준비
X, y = news.data, news.target

#데이터 세트를 학습 세트와 테스트 세트로 분할
X_train, X_test, y_train, y_test = train_test_split(X, y, random_state=1)

#학습 세트의 첫 번째 입력 데이터를 확인
print(X_train[0])
```

```
From: clay@rsd.dl.nec.com (Clay Finley)
Subject: Re: Carrying Arms
Nntp-Posting-Host: rsd21.rsd.dl.nec.com
Organization: NEC America, Radio Software Dept
Distribution: usa
Lines: 22

|> In article <1993Apr5.220457.6800@spdc.ti.com> dwhite@epcot.spdc.ti.com (Dan White) writes:
|>
|> >However, haven't we already lost our right to bear arms?
|>
|> >    It seems that in most states, like Texas, a citizen may own a
|> >gun and carry while at his home or business.  But a citizen is severely
|> >restricted from bearing outside these areas.  Here in Texas you cannot
|> >carry in your car except when "traveling" which is usually defined as
|> >"traveling across a county line."  How did this come about?  Are there
|> >any court rulings on the legality of restricting the carrying of a
|> >weapon outside the home?
|>
```

```
In Texas, it is legal to carry handguns while "traveling", and also to and from
sporting activities.          ^^^^^^^^

Chapter 46 of the Texas State Penal Code does NOT restrict long guns.
Therefore, it is legal to carry and transport long guns any place in Texas.

Regards,
Clay
```

```
#텍스트 데이터를 TF-IDF 벡터 값으로 변환
vectorizer = TfidfVectorizer()
X_train_vec = vectorizer.fit_transform(X_train)
X_test_vec = vectorizer.transform(X_test)
```

- TF-IDF에서 TF(Term Frequency)는 단어 빈도, 즉 단어의 출현 빈도를 말함
  - 예를 들어 스티브 잡스가 말한 “the journey is the reward“라는 문장의 단어 빈도는 다음 같이 나타낼 수 있음

단어	the	hello	journey	world	is	reward
출현빈도	2	0	1	0	1	1

```
#다중 분류 나이브 베이지 분류 모델 객체를 생성
model = MultinomialNB(alpha=0.01)
```