생성적 분류와 식별적 분류

- Machine Learning에는 다양한 계파가 존재
  - 생성적 분류
  - 식별적 분류
  - 빈도주의 및 베이지안 접근법
- 예시: 패턴 x의 유형 y를 예측하는 분류
  - 패턴 x : sample x
  - 유형 y: label y
- 용어 정리
  - 클래스 사후 확률: 사후확률로 보면 됨
  - 데이터 생성 확률: 결합확률(=우도x사전 확률)임
  - [참고] Bayes Rule

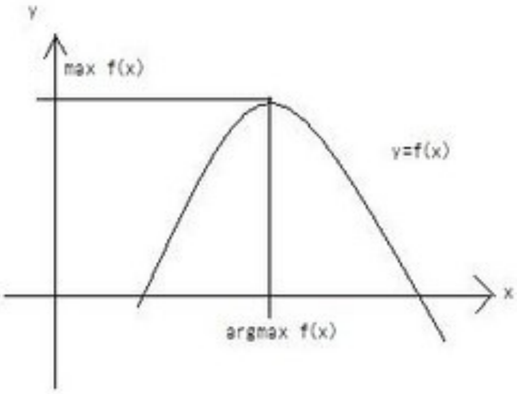
$$p(y|x) = \frac{p(x|y)p(y)}{p(x)}$$

식별적 분류 (discriminative learning algorithms)

- 패턴 x가 주어졌을 때, 유형 y의 조건부 확률 $p(y|x)$가 최대가 되도록 하는 유형 $\hat{y}$을 구할 수 있으면, 패턴 인식을 수행할 수 있음

$$\hat{y} = arg \max_y p(y|x)$$

- $arg \max_y p(y|x)$: $p(y|x)$를 최대로 하는 y의 값
  - $arg \max$: 최대값을 얻게 하는 인수
  - ^: 추정으로 얻은 값
  - [참고] $max_y p(y|x)$:$p(y|x)$의 y에 대한 최대값



- $p(y|x)$: 클래스 사후 확률
- **식별적 분류**
  - **클래스 사후 확률** $p(y|x)$을 **직접 훈련 데이터로부터 학습**하는 접근법

클래스 사후 확률과 결합 확률의 관계

- 클래스 사후 확률은 y의 함수로 다음과 같이 나타낼 수 있음
  - 패턴 x와 유형 y의 결합 확률 $p(x, y)$에 비례함

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \propto p(x|y)p(y) = \mathbf{p(x, y)}$$

  - [참고] 곱 규칙(product rule): $p(y, x) = p(x|y)p(y) = p(x, y)$
  - 분모 $p(x)$는 y가 2 클래스일 때, $p(x) = p(x|y=1)p(y=1) + p(x|y=0)p(y=0)$로 분자인 $p(x|y), p(y)$로 표현할 수 있음. 클래스 사후 확률을 계산하는 데 실제로 필요하지 않음.

생성적 분류 (generative learning algorithms)

- 클래스 사후 확률 $p(y|x)$가 최대가 되도록 하는 유형 $\hat{y}$은 결합확률 $p(x, y)$를 최대로 하는 y로 대신 구할 수 있음

$$\hat{y} = arg\max_y p(x|y)p(y) = arg\max_y p(x, y)$$

- $p(x, y)$: 데이터 생성 확률
- **생성적 분류**
  - **데이터 생성 확률 $p(x, y)$을 추정하는 방법**으로 패턴 인식을 수행하는 접근법

생성적 분류 vs 식별적 분류?

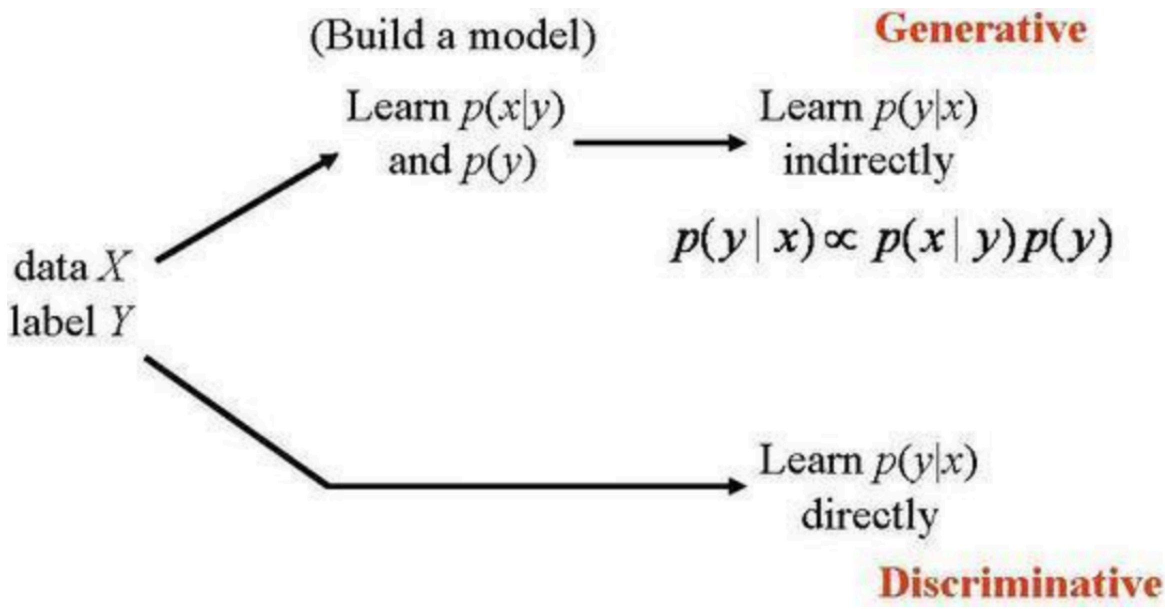
> "제한된 정보만으로 어떤 문제를 풀 때, 그 과정에서 원래의 문제보다 일반적인 문제를 풀지 말고, 가능한 원래 문제를 직접 풀어야 한다." - Vladimir N. Vapnik(SVM 발명자)
>
> 왜냐하면 현재 가진 정보가 일반적인 문제를 풀기에는 불충분하더라도 목표가 되는 문제에 대한 해를 직접 구하는 데 충분할 수 있기 때문이다.

- 데이터 생성 확률 $p(x, y)$을 알고 있다면 아래 식에 따라 사후 확률 $p(y|x)$을 구할수 있음

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

- 그러나, 사후 확률 $p(y|x)$을 알아도, 데이터 생성 확률 $p(x, y)$을 알아내는 것은 불가능 함

- 데이터 생성 확률 $p(x, y)$에 대한 추정은 사후 확률에 대한 추정보다 일반적인 문제(어려운 문제)에 해당함
- **패턴 인식에서는 클래스 사후 확률 $p(y|x)$ 만 알아도 가능하므로, 식별적 분류가 더 바람직한 접근법**
- 다양한 실제 문제에서는 데이터 생성 확률 $p(x, y)$에 대한 선험적 지식을 얻을 수 있는 경우, 생성적 분류가 더 바람직한 접근법
  - 예: 음성 인식 경우, 발성기관 구조나 발음의 원리를 알아보는 것으로 데이터 생성 확률에 대한 지식을 얻을 수 있음



빈도주의 및 베이지언 접근법

- 파라미터 θ를 갖는 모델 $p(x, y; \theta)$를 이용하여, 데이터 생성 확률 $p(x, y)$를 추정하는 문제를 생각

빈도 주의(Frequentism)

- 파라미터 θ를 **결정론적인 변수**로 보고, 주어진 훈련 표본 데이터 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$을 사용하여, 파라미터 θ를 학습함
- 예: 최대 우도 추정(Maximum Likelihood Estimation, MLE) 학습 방법
    - 훈련 데이터 $\mathcal{D}$가 생성될 확률이 가장 높은 파라미터 θ를 학습함

$$\max_{\theta} \prod_{i=1}^{N} q(x_i, y_i; \theta)$$

- 빈도주의에서는 훈련 데이터 $\mathcal{D}$로 부터 정확도 높은 θ를 어떻게 학습할 것인지가 연구 주제임

베이지안 접근법(Bayesian Approach)

- 파라미터 θ를 **확률 변수**로 보고, 그 사전 확률 $p(\theta)$를 상정하여 훈련 데이터 $\mathcal{D}$에 대한 **사후 확률** $p(\theta|\mathcal{D})$를 구함
- 베이즈 정리를 이용해 사후 확률 $p(\theta|\mathcal{D})$은 사전 확률 $p(\theta)$을 써서 다음 식으로 표현
- 주어진 훈련 표본 데이터 $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$을 사용하여, 파라미터 θ를 학습함

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{\prod_{i=1}^n q(x_i, y_i|\theta)p(\theta)}{\int \prod_{i=1}^n q(x_i, y_i|\theta)p(\theta)d(\theta)}$$

- 원리적으로 사전 확률 $p(\theta)$가 주어진다면, 위의 식으로 사후 확률 $p(\theta|\mathcal{D})$ 계산 가능
- 베이지안 접근법에서는 사후 확률을 어떻게 효율적으로 계산할 것인가가 연구 주제임

참고

식별적 접근법, 빈도주의적 학습 방법 위주

[출처]

max, argmax 그림

generative, discriminative 그림

그림과 수식으로 배우는 통통 머신러닝

cs229n Lecture Notes2 Part4. Generative Learning algorithms