생성모델링

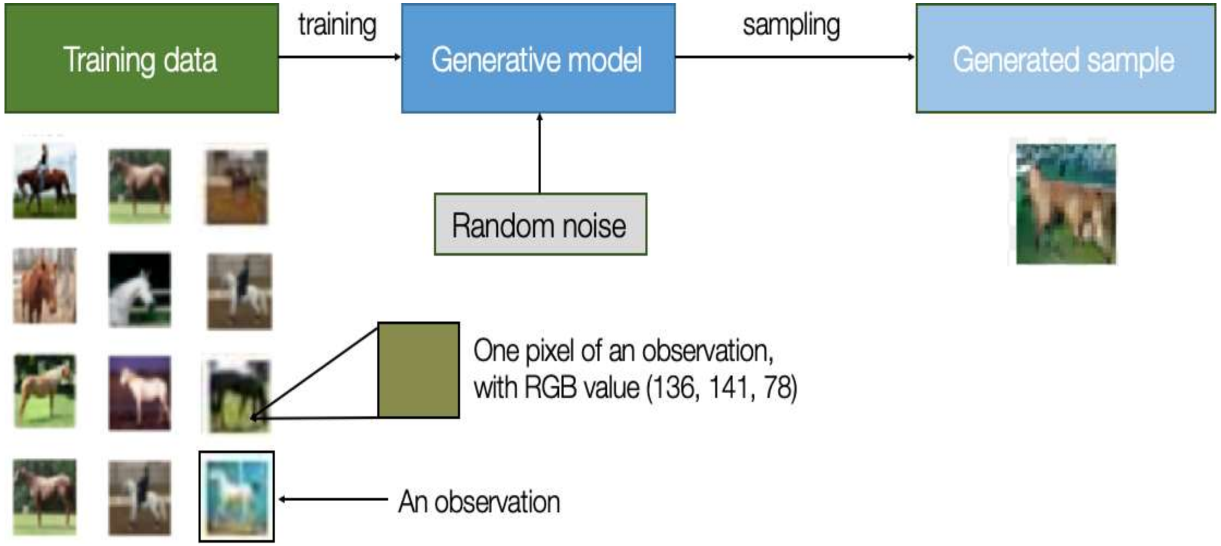
Chapter 1

학습 목표는 아래와 같다.

- **What to learn**
  - 생성 모델링 의미
    - 판별 모델링과의 차이
  - 학습에 필요한 수학 개념 및 프레임 워크
  - 확률 기반의 생성 모델 첫번째 : 나이브 베이즈 예제
  - 나이브 베이즈 예제의 실패 원인 분석

1.1 생성모델링이란?

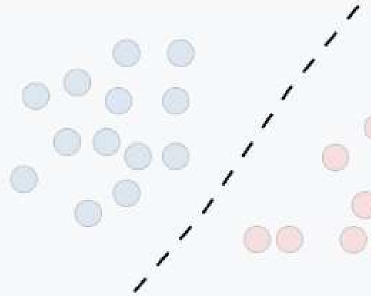
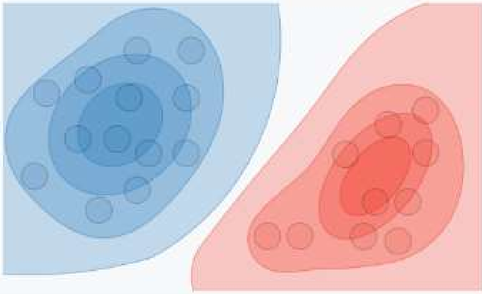
확률 모델의 관점에서, 생성 모델링이란 데이터셋을 생성하는 방법을 기술한 것이다.

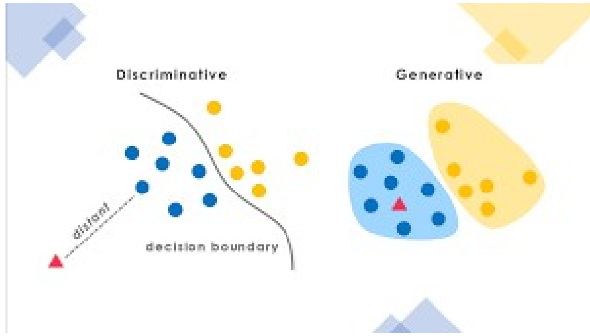


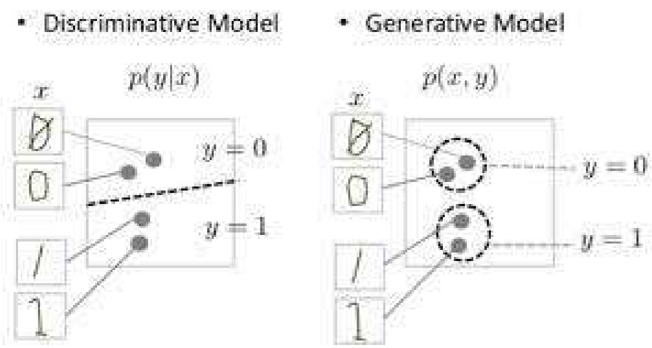
- 각 샘플은 많은 특성으로 구성
  - 이미지 생성 문제일 경우 특성은 개별 픽셀 의미
- 생성 모델은 결정적이 아닌 확률적
  - 훈련 데이터셋에 있을 것 같은 새롭고 완전히 다른 샘플을 생성하는 것이 목표

1.1 생성모델링이란?

생성 모델링 VS 판별 모델링

	Discriminative model	Generative model
Goal	Directly estimate $P(y\|x)$	Estimate $P(x\|y)$ to then deduce $P(y\|x)$
What's learned	Decision boundary	Probability distributions of the data
Illustration		
Examples	Regressions, SVMs	GDA, Naive Bayes

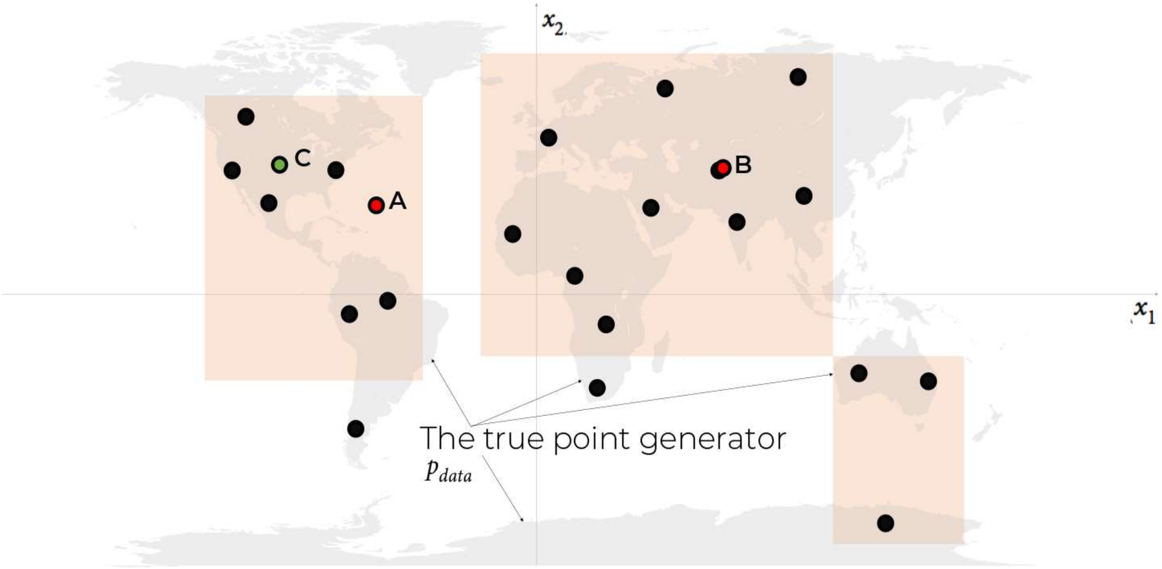




1.1 생성모델링이란?

생성 모델링 프레임워크

- Rule 1: It can generate examples that appear to have been drawn from p_{data}.
- Rule 2: It can generate examples that are suitably different from the observations in $\mathbf{X}$. In other words, the model shouldn't simply reproduce things it has already seen.



- A : 규칙 1 위반
- B : 규칙 2 위반
- C : 성공

1.2 확률적 생성 모델

First Probabilistic Generative Model

- Wrodl dataset
  - 5 features, ($accessoriesType$, $clothingColor$, $clothingType$, $hairColor$, $topType$)
  - There are $7 \times 6 \times 3 \times 4 \times 8 = 4{,}032$ different combinations of these features

- The goal of generative modeling
  - to use these observations to build a p_{model} that can accurately mimic the observations produced by p_{data}.

Table 1-1. The first 10 observations in the Wrodler face dataset

face_id	accessoriesType	clothingColor	clothingType	hairColor	topType
0	Round	White	ShirtScoopNeck	Red	ShortHairShortFlat
1	Round	White	Overall	SilverGray	ShortHairFrizzle
2	Sunglasses	White	ShirtScoopNeck	Blonde	ShortHairShortFlat
3	Round	White	ShirtScoopNeck	Red	LongHairStraight
4	Round	White	Overall	SilverGray	NoHair
5	Blank	White	Overall	Black	LongHairStraight
6	Sunglasses	White	Overall	SilverGray	LongHairStraight
7	Round	White	ShirtScoopNeck	SilverGray	ShortHairShortFlat
8	Round	Pink	Hoodie	SilverGray	LongHairStraight
9	Round	PastelOrange	ShirtScoopNeck	Blonde	LongHairStraight

First Probabilistic Generative Model

- Parameter
  - this parametric model would have $d = 4{,}031$ parameters—one for each point in the sample space of possibilities

- maximum likelihood estimate $\hat{\theta}_j = \frac{n_j}{N}$
  - combination 1
  - (*LongHairStraight, Red, Round, ShirtScoopNeck, White*)
  - $\hat{\theta}_1 = \frac{2}{50} = 0.04$

- **Problem**
  - it would assign just as much weight to a random collection of colorful pixels as to a replica of a Picasso painting that differs only very slightly from a genuine painting.

- **To achieve this, we need to choose a different parametric model.**

Naive Bayes

- *Naive Bayes* assumption
  - each feature x_j is *independent* of every other feature x_k.
    - p (x j | x k) = p (x j)

- Parameter
  - This model is defined by only 7 + 6 + 3 + 4 + 8 – 5 = 23 parameters.

Figure 1-8. Ten new Wrodl styles, generated using the Naive Bayes model

1.3. 생성모델의 난관

Why is this?

- 픽셀은 상호 연관성이 크며, 샘플이 나올 수 있는 공간은 매우 거대함.
- 그러므로 픽셀을 독립적으로 샘플링하여 의미 있는 얼굴 이미지를 출력하는 것은 거의 불가능함.
- **위 이유들로 인하여 나이브 베이즈 모델로 실제 이미지를 처리하였을 때 성능이 안좋음.**

1.3. 생성모델의 난관

Representation Learning(표현 학습)

we describe each observation in the training set using **some low-dimensional *latent* space** and then **learn a mapping function**

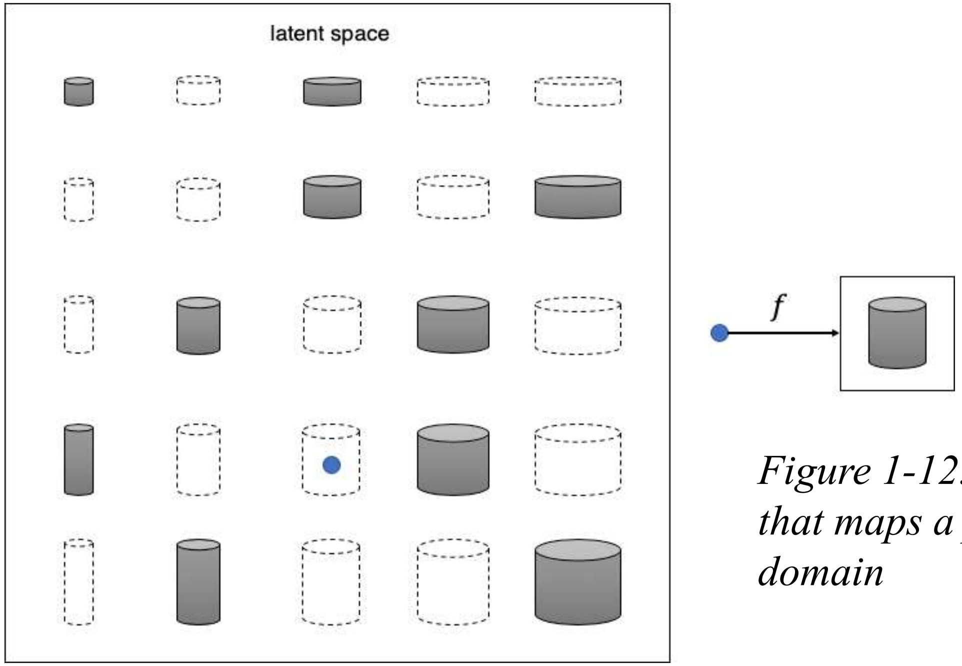


Figure 1-12. The latent space of biscuit tins and the function f that maps a point in the latent space to the original image domain

1.3. 생성모델의 난관

Representation Learning(표현 학습)

representation learning establishes the most relevant **high-level features**

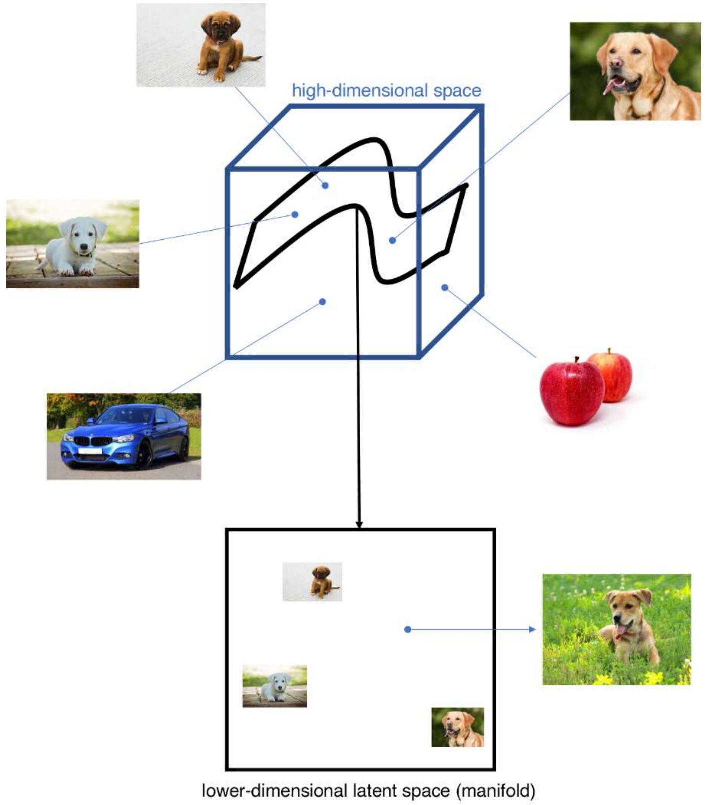


Figure 1-13. The cube represents the extremely high-dimensional space of all images; representation learning tries to find the lower-dimensional latent subspace or manifold on which particular kinds of image lie (for example, the dog manifold)