Sampling

Sampling (샘플링)

: 임의의 확률 분포 $p(x)$로부터 표본을 추출하는 작업

- Ex) 주사위를 여러번 던질 때 도출된 앞면의 숫자 = 표본
- 높은 차원의 최적화 문제를 풀 때 샘플링 필요

Candidate Sampling

: 자연어처리 딥러닝 모델 생성 시 소프트맥스 확률값을 구할 때 계산량을 줄이도록 하는 방식

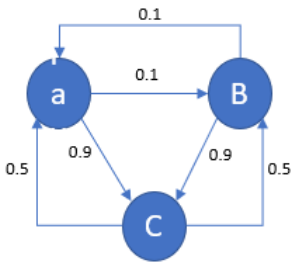
- 말뭉치의 모든 단어를 쓰지 않고 몇 개만 뽑아서 소프트맥스 확률값 계산 후, 해당 단어들과 관계된 파라미터에 대해서만 업데이트

Negative Sampling

: candidate sampling + 노이즈 ("negative") 샘플인지의 여부를 가림

- 노이즈 샘플 여부는 이진분류문제로 접근 (0이면 노이즈로)
- "negative sample" = 사용자가 정한 윈도우 내에 등장하지 않는 단어들
- word2vec에 쓰임
  - 중심 단어와 positive 샘플의 단어 벡터끼리는 유사하게 업데이트
  - 중심 단어와 negative샘플의 단어 벡터끼리는 멀게 업데이트

Markov Chain (마르코프 연쇄, Markov Process)



$$P = \begin{bmatrix} 0 & 0.1 & 0.9 \\ 0.1 & 0.0 & 0.9 \\ 0.5 & 0.5 & 0.0 \end{bmatrix}$$

: 마르코프 성질을 가진 이산 확률 과정. 즉,

$$P(X_{n+1} = X | X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = P(X_{n+1} = x | X_n = x_n)$$

을 만족하는 확률변수들의 수열

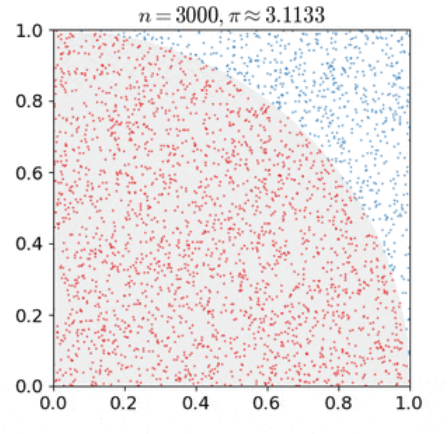
- **Markov Property (마르코프 성질)** = 과거와 현재 상태가 주어졌을 때의 미래 상태의 조건부 확률 분포가 현재 상태에 의해서만 결정됨 (즉 각 상태는 바로 이전의 상태에만 영향을 받음)
- 연속적인 현상을 단순히 표현할 때 마르코프 체인을 가정하여 사용 (기회비용을 절감하면서 실제 예측에 근접한 효과를 낼 수 있음)

Monte Carlo Method (몬테카를로 방법)

: 반복적으로 무작위 샘플링된 난수를 이용하여 함수의 값을 확률적으로 근사하는 알고리즘

- **모든 샘플이 독립이고, 생성될 확률도 랜덤**

- Ex) 원주율 구하기: 수많은 난수 페어 (x, y)를 발생시켜 원 안에 위치한 쌍들의 비율을 계산해 면적 구하기



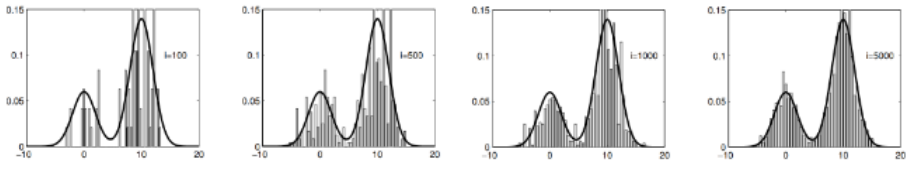

```
import random

n=1000
count=0
for i in range(n):
  x=random.random()
  y=random.random()
  if(x*x+y*y<1):
    count=count+1
a=4*count/n
print(a)
```

- 최적화, 확률분포로부터의 추출 등에 활용
- 인공지능 분야에 획기적인 계기 마련
  - 구글 딥마인드 알파고 (표본 추출해서 승률 근사)
- 장점 및 단점
  - 쉽고 빠르게, 그리고 적은 수의 샘플로 근사값 추정 가능
  - but 정확도 떨어짐. 랜덤을 이용한 것이라 샘플을 더 많이 뽑아도 정확도가 많이 올라가지는 않음

MCMC (Markov Chain Monte Carlo)

: 몬테카를로 과정을 마르코프 체인에 적용시킨 것

- i번째 표본을 참고하여 $i+1$번째 표본을 뽑음
  - 모든 샘플이 독립인 MC와는 다르게 이전의 샘플이 **다음 샘플의 추출에 영향을 줌**
- 특정한 확률분포를 샘플링을 통해 모사 가능



- 필요한 만큼 뽑게 되면 확률 분포를 거의 정확하게 모방함
- 확률 기반 시뮬레이션은 대부분 MCMC 활용 (대표적: 베이즈 확률론)

Metropolis-Hastings Algorithm

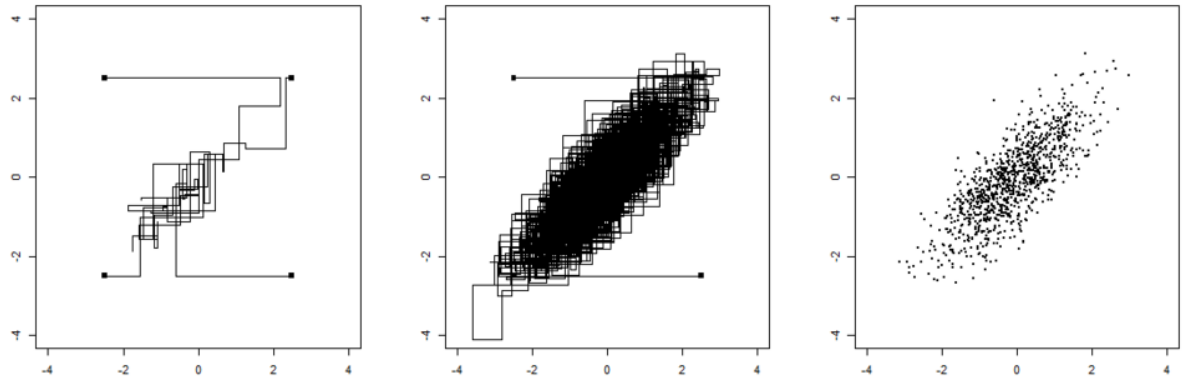
: MCMC의 기본 알고리즘

- 구하고자 하는 분포 $p(x)$와, 그 분포에 비례하는 함수 $f(x)$를 알고 있을 때, 그리고 이 두가지와 도메인을 공유하는 **조건부 확률분포**가 있을 때 $p(x)$에 대한 MCMC 샘플링을 하는 알고리즘

https://www.secmem.org/blog/2019/01/11/mcmc/

Gibbs Sampling

: 복잡한 분포로부터 샘플을 추출할 시 쓰이는 MCMC 알고리



- 다음 샘플은 현재 샘플에 영향을 받지만, 나머지 변수는 그대로 두고 한 변수에만 변화를 줌
- 차원이 많이 큰 경우 사용. 고차원의 문제를 1차원의 문제로 바꾸어 샘플링
- Ex) 확률변수 3개의 결합활률분포 $p(x_1, x_2, x_2)$으로부터 한개의 표본을 얻으려고 할 때
  - 임의 표본 $X^0 = (x_1^0, x_2^0, x_3^0)$ 선택 → 변수 하나만을 변경하여 새로운 표본 X^1 추출

https://m.blog.naver.com/sw4r/221917843395