9.1 확률분포의 추정

["분석하고자 하는 데이터는 어떤 확률변수로부터 실현된 표본이다"]

= 데이터 분석의 첫번째 가정

"우리가 정말 관심있는 것" = 표본 데이터 이면에서 데이터를 만들어내는 확률변수의 분포

* 확률론적 관점에서, 데이터는 확률변수의 분포를 알아내기 위한 실현의 참고 자료

* 데이터 표본으로부터 확률변수의 분포를 알아내고자 하는 것

= 추정 (확률분포의 추정)

확률분포의 결정

1. 확률변수가 우리가 배운 베르누이분포, 이항분포, 정규분포 등의 기본 분포 중 어떤 확률분포를 따르는지 알아낸다.
2. 데이터로부터 해당 확률분포의 모수의 값을 구한다.

1번 —
- 데이터가 생성되는 원리 파악
- 데이터의 특성 분석
- 히스토그램 시각화를 통한 확률분포의 모양을 살펴보고 힌트 얻기

- 데이터가 0 또는 1 뿐이다 — 베르누이 분포
- 데이터가 카테고리 값이어야 한다 — 카테고리 분포
- 데이터가 0과 1 사이 실수값이어야 한다 — 베타분포
- 데이터는 항상 0 또는 양수이어야 한다 — 로그정규분포, 감마분포, F, chi2, 지수, 하프 코시 분포 등
- 데이터가 크기 제한이 없는 실수다 — 정규분포, 스튜던트 t 분포, 코시분포, 라플라스 분포 등

(연습문제 9.1.1)
사이킷런 보스턴 집값 데이터에서 각각의 feature가 어떤 확률분포와 적합한지 설명하라.

모수추정 방법론

두번째 작업, 즉 모수의 값으로 가장 가능성이 높은 하나의 숫자를 찾아내는 작업
= 모수추정 (parameter estimation)

- 모멘트 방법 (이번절에서 학습)
- 최대가능도 추정법
- 베이즈 추정법

모멘트 방법 method of moment

모멘트 방법은 표본자료에 대한 "표본모멘트가 확률분포의 이론적 모멘트와 같다고 가정"하여 모수를 구한다.

$$\mu = E[X] \triangleq \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

위 식에서 N은 데이터의 갯수, x_i는 표본 데이터다.

2차 모멘트 (분산)의 경우에는 다음과 같다.

$$\sigma^2 = E[(X-\mu)^2] \triangleq s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2$$

예제 : 베르누이분포의 모수추정

※ N_1 = 1이 나온 횟수

$$E[X] = \mu \triangleq \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{N_1}{N}$$

예제: 정규분포의 모수추정

모멘트 방법으로 정규분포의 모수 μ, σ^2를 구하면 다음과 같다.

$$E[X] = \mu \triangleq \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$E[(X-\mu)^2] = \sigma^2 \triangleq s^2 = \frac{1}{N-1}\sum_{i=1}^{N} (x_i - \bar{x})^2$$

예제: 베타분포의 모수추정

모멘트방법으로 베타 분포의 모수 a, b를 구하면 다음과 같다.
이 경우에는 모수와 모멘트 간의 관계를 이용하여 비선형 연립 방정식을 풀어야 한다.

$$E[X] = \frac{a}{a+b} \triangleq \bar{x} \quad \cdots\cdots ①$$

$$E[(X-\mu)^2] = \frac{ab}{(a+b)^2(a+b+1)} \triangleq s^2 \quad \cdots ②$$

이 비선형 연립방정식을 풀어 해를 구하면 다음과 같다.

$$a = \bar{x}\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$$

$$b = (1-\bar{x})\left(\frac{\bar{x}(1-\bar{x})}{s^2} - 1\right)$$

* seaborn.distplot()을 이용한 간단한 모수추정방법

```
sns.distplot(x, kde=False, norm_hist=True,
             fit = sp.stats.beta )
```

↳ 이 확률변수로 모수를 추정하고,
해당 pdf 그래프를 추정 그려준다.