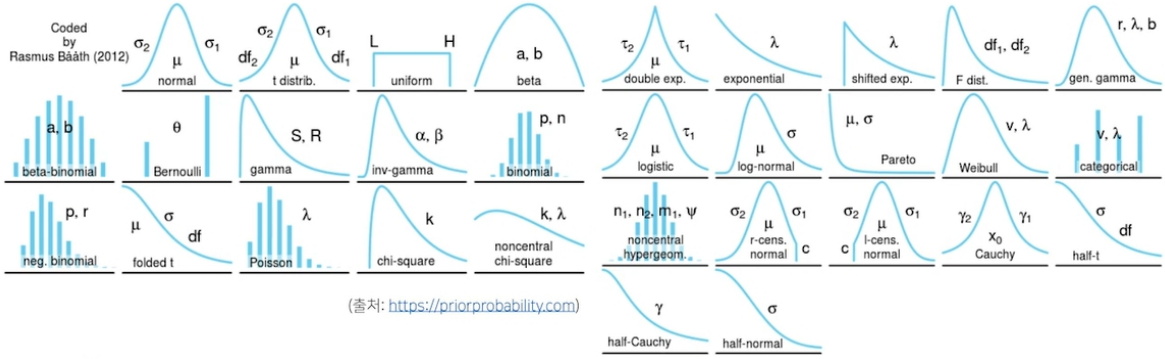
4. 통계학

4.1 통계적 모델링

- 통계적 모델링은 적절한 가정 위에서 확률분포를 추정(inference)하는 것이 목표이다.
- 이는 기계학습과 통계학이 공통적으로 추구하는 목표이다.
- 데이터를 통해서 정답의 분포를 확실하게 알 수는 없다.
- 또한 사용할 수 있는 분포도 매우 다양하다.



- 그러나 유한한 개수의 데이터만 관찰해서 모집단의 분포를 정확하게 알아낸다는 것은 불가능하므로, **근사적으로 확률분포를 추정**할 수 밖에 없다.
- 예측모형의 목적은 분포를 정확하게 맞추는 것 보다는 데이터와 추정 방법의 불확실성을 고려해서 위험을 최소화하는 것

4.1.1 모수적(parametric) 방법론

- 데이터가 특정 확률분포를 따른다고 선험적으로(a priori) **가정한 후** 그 분포를 결정하는 모수(parameter)를 추정하는 방법론
- ex) 정규분포를 가지고 확률분포를 모델링할 경우, 정규분포의 두 가지 모수인 평균과 분산을 추정하는 방법을 통해 데이터를 학습한다.

4.1.2 비모수(noparametric) 방법론

- 특정 확률분포를 **가정하지 않고** 데이터에 따라 모델의 구조 및 모수의 개수가 유연하게 바뀌는 방법론
- 기계학습의 많은 방법론은 비모수 방법론에 속한다.
- (주의)
    - 비모수 방법론이라고 해서 모수가 없는 것은 아니다.
    - 모수가 무한히 많거나 모수가 데이터에 따라서 바뀌는 경우가 비모수 방법론이다.

4.2 확률분포 가정하기 (예제)

4.2.1 확률분포를 가정하는 방법 : 히스토그램 모양 관찰

- 데이터가 2개의 값(0 또는 1)만 가지는 경우 $\rightarrow$ 베르누이분포
- 데이터가 n개의 이산적인 값을 가지는 경우 $\rightarrow$ 카테고리분포, 다항분포
- 데이터가 $[0, 1]$ 사이의 실수값을 가지는 경우 $\rightarrow$ 베타분포
- 데이터가 0 이상의 값을 가지는 경우 $\rightarrow$ 감마분포, 로그정규분포 등
- 데이터가 $\mathbb{R}$ 전체에서 값을 가지는 경우 $\rightarrow$ 정규분포, 라플라스분포 등

- 단, 기계적으로 확률분포를 가정해서는 안되며, **데이터를 생성하는 원리를 먼저 고려하는 것이 원칙**이다.
- 각 분포마다 검정하는 방법들이 있으므로 모수를 추정한 후에는 반드시 검정을 해야 한다.

4.3 데이터로 모수를 추정

- 데이터의 확률분포를 가정했다면 모수를 추정해볼 수 있다.

4.3.1 정규분포의 모수

- 정규분포의 모수는 평균 μ 과 분산 σ^2 이다.
- 이 두 가지 모수를 추정하는 통계량(statistic)은 다음과 같다.

표본평균 ($\bar{X}$)

- 모집단의 평균을 추정하는 통계량
- $\bar{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$
  - 주어진 데이터의 산술평균
- $\mathbb{E}[\hat{X}] = \mu$
  - 표본평균의 기대값은 원래 데이터에서 관찰되는 모집단의 평균과 일치한다.

표본분산 (S^2)

- 모집단의 분산을 추정하는 통계량
- $S^2 = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_i)^2$
  - 주어진 데이터에서 표본평균을 뺀 값의 제곱의 산술평균
  - 표본분산을 구할 때 N이 아니라 $N-1$ 로 나누는 이유는 불편(unbiased) 추정량을 구하기 위해서이다.
    - 표본분산의 기대값을 취했을 때 모집단의 분산과 일치하기 위해서 사용
- $\mathbb{E}[S^2] = \sigma^2$
  - 표본분산의 기대값은 원래 데이터에서 관찰되는 모집단의 분산과 일치한다.

4.3.2 통계량(statistic)의 활용

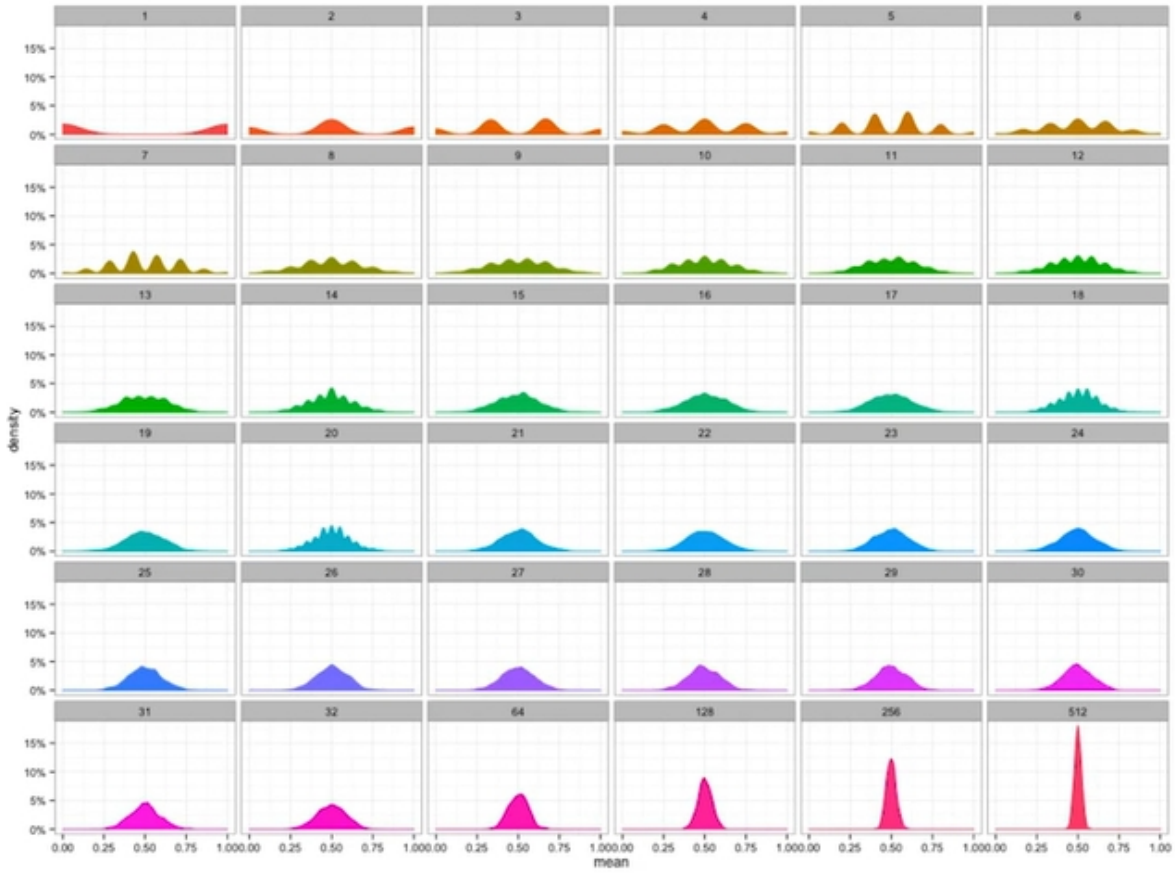
- 이렇게 구한 표본평균과 표본분산을 가지고 주어진 데이터의 확률분포의 모수를 추정해볼 수 있다.
- 추정된 모수를 가지고 원래 데이터의 성질 및 정보를 취합할 수 있다.
- 이를 통해 예측을 하거나 의사결정을 내릴 때 이와 같은 통계량을 사용해볼 수 있다.

4.3.2 표집분포 (sampling distribution)

- **통계량(표본평균, 표본분산)**의 확률분포를 **표집분포**라고 부른다. (표본들의 분포 x)
- (주의)
    - 표본분포(sample distribution)와 표집분포(sampling distribution)은 다른 것이다.

4.3.3 중심극한정리 (Central Limit Theorem)

- **표본평균의 표집분포**는 N이 커질수록 정규분포 $\mathcal{N}(\mu, \sigma^2/N)$ 를 따른다.
- 이를 중심극한정리 라고 부른다.
- 모집단의 분포가 정규분포를 따르지 않아도 성립한다.
    - 모집단의 분포가 정규분포를 따르지 않으면 표본분포(sample distribution)은 N이 커져도 정규분포를 따르지 않는다.
    - 모집단의 분포가 정규분포를 따르지 않아도 표본평균의 표집분포(sampling distribution)은 N이 커지면 정규분포를 따른다.
- ex) 베르누이 확률분포(이항분포)를 따르는 확률변수의 분포
    - 처음 데이터를 모았을 때 표본평균의 분포를 확인해보면 처음에는 양 극단으로 나뉜 데이터가 관찰된다.
    - 데이터를 모으면 모을수록 베르누이분포들의 표본평균의 표집분포는 정규분포를 따른다.
    - 평균값은 하나의 값에 몰려 있다.
    - 반면 분산은 점점 짧아진다.
        - 데이터의 개수 N이 증가하면 표본분산의 σ^2/N 이 0으로 간다.



4.4 최대가능도 추정법 (MLE)

- 표본평균이나 표본분산은 정규분포 뿐만이 아니라 다른 분포에서도 계산할 수 있는 중요한 통계량 중 하나이다.
- 하지만 확률분포마다 사용하는 모수가 다르므로 확률분포의 성질을 추정하는 모수를 결정하는 적절한 통계량이 달라지게 된다.
- 이론적으로 가장 가능성이 높은 모수를 추정하는 방법 중 하나는 **최대가능도 추정법(Maximum Likelihood Estimation, MLE)**이다.
- 최대가능도 추정법을 이용해서 주어진 확률분포를 어떤 식으로 가정하느냐에 상관없이 이론적으로 가능성이 가장 높은 모수를 추정한다.

$$\hat{\theta}_{\text{MLE}} = \underset{\theta}{\operatorname{argmax}}\, L(\theta\,;\,\mathrm{x}) = \underset{\theta}{\operatorname{argmax}}\, P(\mathrm{x}\,|\,\theta)$$

4.4.1 가능도(likelihood) 함수 : $L(\theta\,;\,\mathrm{x})$

- 모수 θ 를 따르는 분포가 x 를 관찰할 **가능성**을 뜻한다. (모수 θ 에 대한 **확률로 해석하면 안된다.**)
- 가능도 함수는 확률질량(밀도)함수와 같은 것이지만 **관점의 차이**가 있다.
  - 확률질량(밀도)함수
    - 모수 θ 가 주어져 있을 때 데이터 x 에 대한 함수
  - 가능도 함수
    - 주어진 데이터 x 에 대해서 모수 θ 를 변수로 둔 함수
    - 데이터가 주어진 상황에서 모수 θ를 변형시킴에 따라 값이 바뀌는 함수

4.4.2 로그가능도 최적화

- 데이터 집합 X 가 **독립적으로 추출**되었을 경우 **로그가능도를 최적화**한다.
  - 데이터 집합 X 의 각 행벡터가 **독립적으로 추출**되었을 경우 가능도 함수를 **확률질량(밀도)함수 $P(\mathrm{x}_i \mid \theta)$ 의 곱셈**으로 표현할 수 있다.
    - 곱셈으로 표현되는 것은 데이터 집합의 확률분포가 독립적으로 추출되었을 경우에 가능한 공식이다.
  - 이 경우 로그 함수의 성질을 이용해서 가능도 함수에 로그를 씌워주게 되면 **로그 가능도**가 된다.
  - 이는 **확률질량(밀도)함수 $P(\mathrm{x}_i \mid \theta)$의 덧셈**으로 표현할 수 있다.

$$L(\theta\,;\,\mathrm{x}) = \Pi_{i=1}^{n} P(\mathrm{x}_i \mid \theta) \quad \Rightarrow \quad log\ L(\theta\,;\,\mathrm{x}) = \sum_{i=1}^{n} log\ P(\mathrm{x}_i \mid \theta)$$

4.4.3 로그가능도를 사용하는 이유

- 로그가능도를 최적화하는 모수 θ 는 가능도를 최적화하는 MLE가 된다.
- 데이터의 숫자가 적으면 상관없지만 만일 데이터의 숫자가 수억 단위가 된다면 컴퓨터의 정확도로는 가능도를 계산하는 것은 불가능하다.
  - 0 ~ 1 사이의 확률을 수억번 곱해주는 연산은 컴퓨터는 **연산 오차** 때문에 연산이 불가능하다.
- 데이터가 독립일 경우, 로그를 사용하면 **가능도의 곱셈을 로그가능도의 덧셈으로 바꿀 수 있기 때문**에 컴퓨터로 연산이 가능해진다.
- 경사하강법으로 가능도를 최적화할 때 미분 연산을 사용하게 되는데, 로그 가능도를 사용하면 가능도를 사용했을 때의 연산량 $O(n^2)$ 을 $O(n)$ 으로 줄여준다.
- 대게의 손실함수의 경우 경사하강법을 사용하므로 **음의 로그가능도(negative log-likelihood)**를 최적화하게 된다.

4.4.4 최대가능도 추정법 예제 : 정규분포

> 정규분포를 따르는 확률변수 X 로부터 독립적인 표본 $\{x_1, \ldots, x_n\}$ 을 얻었을 때 최대가능도 추정법을 이용하여 모수를 추정하면?

$$\hat{\theta}_{\mathrm{MLE}} = \underset{\theta}{\mathrm{argmax}}\ L(\theta\,;\,\mathrm{x}) = \underset{\theta}{\mathrm{argmax}}\ P(\mathrm{x} \mid \theta) = \underset{\mu,\,\sigma^2}{\mathrm{argmax}}\ P(X \mid \mu,\,\sigma^2)$$

- 최대가능도 추정법은 주어진 데이터를 가지고 가능도 함수 $L(\theta\,;\,\mathrm{x})$ 를 최적화하는 모수 θ 를 찾는 것이다.

$$\begin{aligned} L(\theta\,;\,\mathrm{x}) &= \sum_{i=1}^{n} log\ P(x_i \mid \theta) \\ &= \sum_{i=1}^{n} log\ \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{|x_i-\mu|^2}{2\sigma^2}} \\ &= -\frac{n}{2} log\ 2\pi\sigma^2 - \sum_{i=1}^{n} \frac{|x_i-\mu|^2}{2\sigma^2} \end{aligned}$$

- $\theta = (\mu,\,\sigma)$ 에 대해 마지막 수식을 미분해서 최적화를 할 수 있다.

내용 작성 x

4.4.5 최대가능도 추정법 예제 : 카테고리 분포

내용 작성 x

4.5 딥러닝에서 최대가능도 추정법

- 최대가능도 추정법을 이용해서 기계학습 모델을 학습할 수 있다.
- 딥러닝 모델의 가중치 $\theta = (W^{(1)}, \dots, W^{(L)})$ 라 표기했을 때 분류 문제에서 소프트맥스 벡터는 카테고리분포의 모수 $(p_1, \dots, p_k)$ 를 모델링한다.
- 원-핫 벡터로 표현한 정답 레이블 $\mathrm{y} = (y_1, \dots, y_K)$ 을 관찰 데이터로 이용해 확률분포인 소프트맥스 벡터의 로그가능도를 최적화할 수 있다.

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} y_{i,k} log(MLP_\theta(\mathrm{x}_i)_k)$$

4.6 확률분포의 거리

내용 작성 x