2020.03.11 발제

모두의연구소 풀잎스쿨 낯선수학 10-1주차 발제

- 발제자: 박진형

오늘의 목표

- 확률분포의 추정과 모수추정에 대한 기본 개념을 습득한다.
- MLE - 최대가능도 추정법에 대한 기본 개념을 습득한다.
- 베이즈 추정에 대한 기본 개념을 습득한다.

9-1. 확률분포의 추정

- 분석할 데이터는 어떤 확률변수로부터 실현된 표본이다.
- 우리의 관심은 표본 뒤의 데이터를 만들어내는 확률 변수의 분포이다.
- 확률분포를 알아내는 일은 다음 두 작업으로 나뉜다.
  - 확률변수가 우리가 배운 베르누이분포, 이항분포, 카테고리분포, 정규분포 중 어떤 확률분포를 따르는가?
  - 데이터로부터 해당 확률분포의 모수의 값을 구한다.
- 모수의 값으로 가장 가능성이 높은 하나의 숫자를 찾아내는 작업을 모수추정이라고 한다. 다음의 방법이 있다.
  - 모멘트 방법
  - 최대가능도 추정법
  - 베이즈 추정법

모멘트 방법

- 표본자료에 대한 표본모멘트가 확률분포의 이론적 모멘트와 같다고 가정하여 모수를 구한다.

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

- N은 데이터의 개수, xi는 표본 데이터다.

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

- 분산이다. 2차 모멘트라고도 부른다.
- 베르누이분포의 모수 추정
  - 1의 개수에서 전체 데이터 개수를 나눈다.

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

- 정규분포의 모수 추정

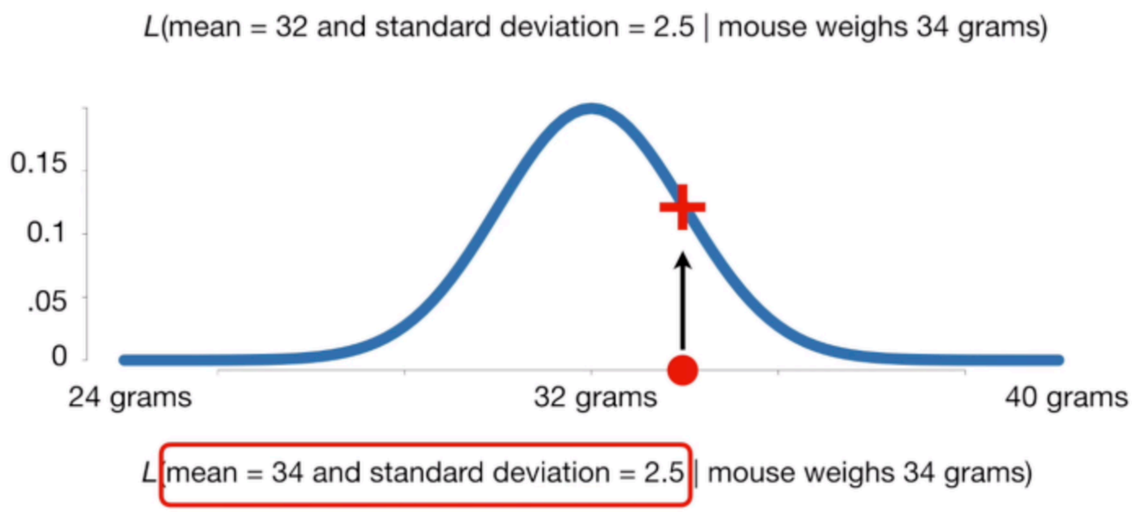
$$E[X] = \mu \triangleq \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

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

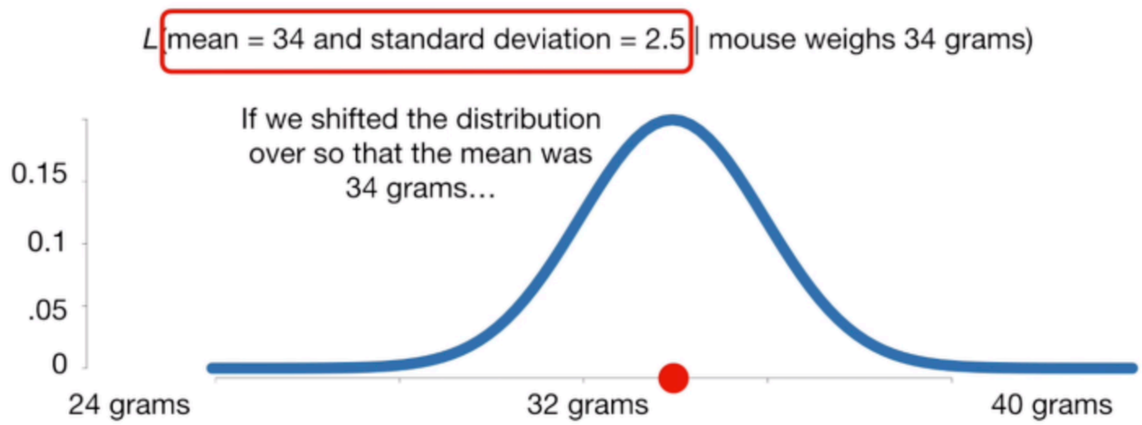
가능도함수

> 아래의 자료는 다음의 유튜브 링크 에서 참조하였다.

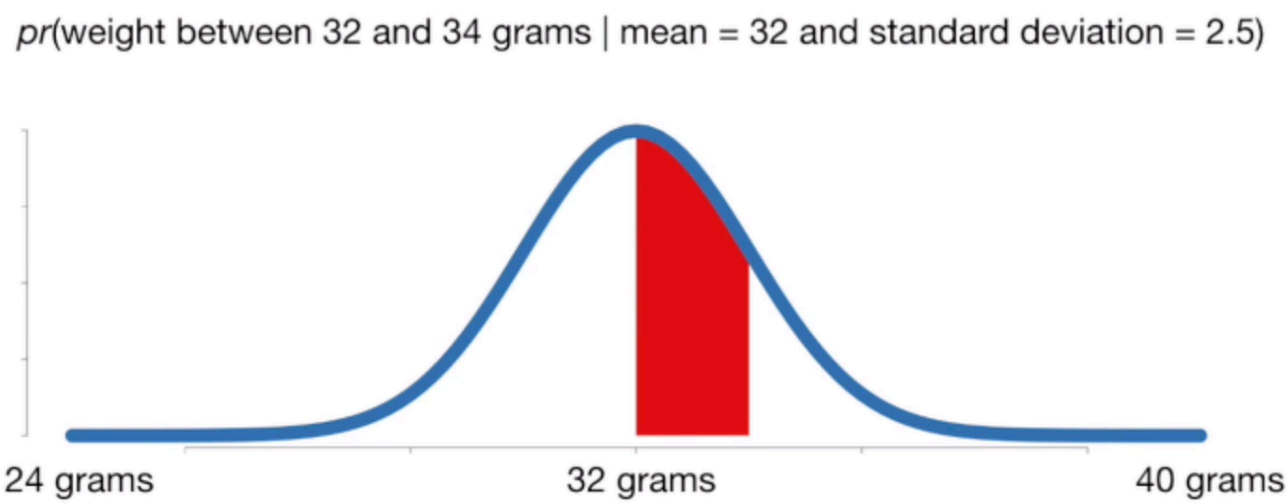
- 가능도는 주어진 표본에서 가장 가능한(likely) 모수를 추정하는 척도이다.
  - 어떤 값이 관측되었을 때, 이것이 어떤 확률 분포에서 왔을 지에 대한 확률이다. (확률의 확률)
  - 관측값이 고정되고 그것이 주어졌을 때 해당 확률분포에서 나왔을 확률을 구한다.
  - 내가 쥐를 하나 골라 달았는데 34g의 무게가 나왔다. 이 관측 결과가 평균이 32이고 분산이 2.5인 확률에서 나왔을 확률은 0.12이다.



  - 평균을 34로 조절하니 가능도가 높아졌다.



- 확률은 모수(parameter)가 특정값(fixed value)으로 정의가 되어 있을 때 모수를 찾는다.
    - 어떤 고정된 분포에서 이것이 관측될 확률이다.



최대 가능도 추정법(MLE)

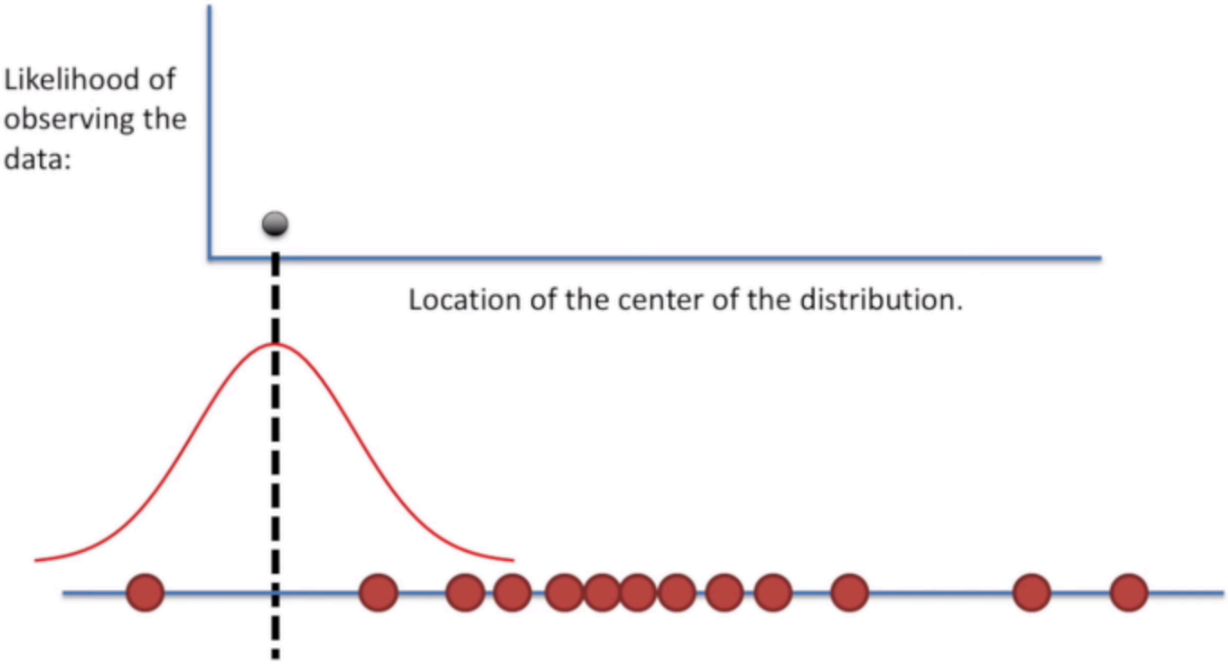
- 최대가능도 추정법(MLE)은 주어진 표본에 대해 가능도를 가장 크게 하는 모수 θ 를 찾는 방법이다.

이해하기

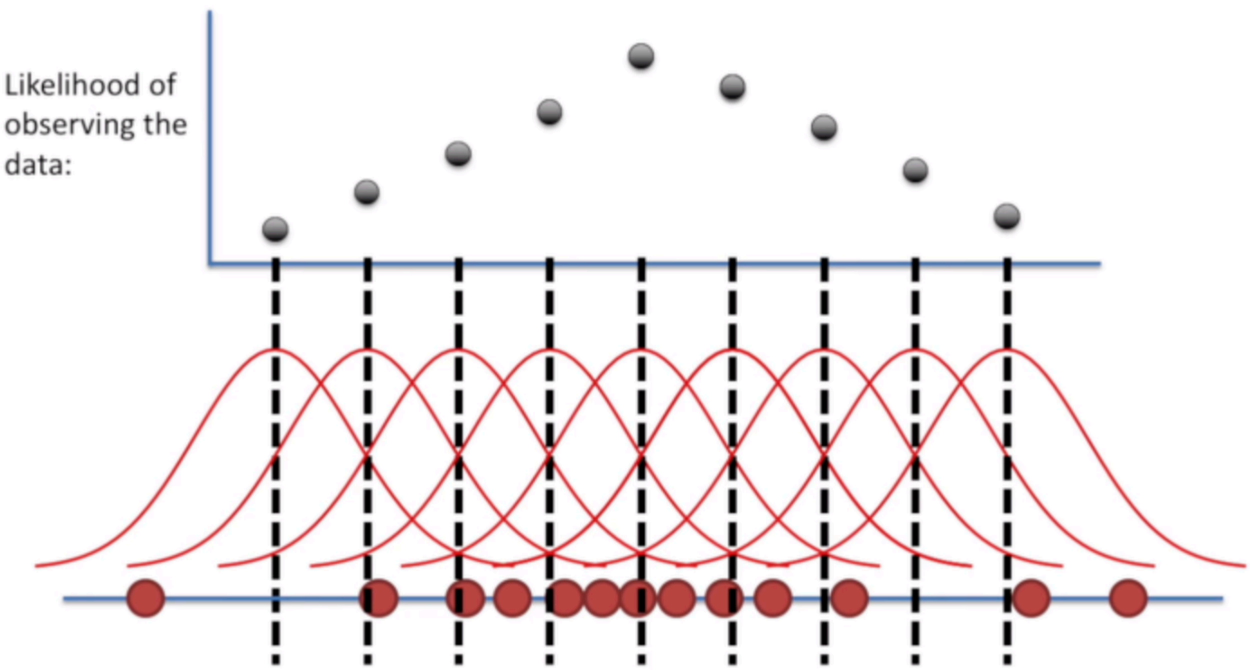
> 아래 내용은 다음의 유튜브 링크를 참조하였다.

- 쥐의 무게를 여러 번 관측했다고 가정하자. 아래와 같은 붉은 점에 바로 관측된 결과값이다.
- 이 때 이렇게 관측될 가능성이 가장 큰 확률분포는 무엇일까? 를 풀어내는 것이 Maximum Likelihood이다.

- 쥐의 관측 결과가 대략적으로 정규분포일 것을 가정하자. 만약 해당 왼쪽에 평균이 치우친 정규분포일 때 총 가능도가 검은 점과 같다고 하자.



- 정규분포의 평균을 조금씩 키울 때마다 가능도가 어떻게 변화하는지 확인할 수 있다. 우리가 수집한 관측값들이 나올 수 있는 가장 가능한 확률분포는 가능도가 가장 큰, 검은 점이 가장 높게 솟아 있는 정규분포에서 왔다고 추정할 수 있다.



이야기로 풀어가는 최대가능도 추정법

> 아래의 내용은 다음 자료 를 참고했습니다.

- 우리는 일반적으로 동전 하나를 던졌을 때 앞 또는 뒤가 나올 확률은 같다고 가정하여 0.5라고 생각한다.
- 하지만 이 0.5라는 확률 값은 정확한 값이 아니라 우리가 가정한 값이다. 때문에 몇 번 의 수행 결과로 동전의 앞면이 나올 확률 P(H)를 정하고자 한다.

- 만약 100번의 동전던지기를 수행했을 때, 앞면이 56번 나왔다면 '동전을 던졌을 때 앞면이 나올 확률'은 얼마라고 얘기 할 수 있을까? 이 문제에 대한 해답을 구하는 것이 MLE이다.
- 좀 더 자세하게 설명하면 P(E|T)와 비례하는 L(T|E)에서, 표본 E는 앞면이 56 번 나왔다는 사실이고, 이론 T를 변화시키면서 어느 이론이 가장 그 확률이 높은지 찾는 과정이다.

$$L(P(H) = 0.5|E) = 100!/56!44! * 0.5^{56} * 0.5^{44} \approx 0.0389$$

  - 가정을 바꾸어가면서 계산을 반복해보면, 다음의 결과를 얻을 수 있다.

T	Likelihood
P(H) = 0.48	0.0222
P(H) = 0.50	0.0389
P(H) = 0.52	0.0587
P(H) = 0.54	0.0739
P(H) = 0.56	0.0801
P(H) = 0.58	0.0738
P(H) = 0.60	0.0576
P(H) = 0.62	0.0378

- 위 표에서 가장 높은 Likelihood를 가지는 이론 T는 P=0.56 일 때이다. 때문에 우리는 동전의 앞 면이 나올 확률이 0.5 라는 것을 전혀 모르는 상황에서 위와 같은 증거가 있을 때에는, "동전을 던져서 앞면이 나올 확률은 0.56 이다" 라고 말할 수 있다.
- 이 이야기의 교훈은?
  - 최대 우도 추정법은 다음과 같이 구한다.
    - 모델을 설정한다.
    - 그 모델에서 본인이 목격한 사건들의 발생 확률 식을 설정한다.
    - 그 확률을 최대로 높이는 모델 변수를 구한다.
  - 언제 사용하는가?
    - P(H)에 대한 확률을 모르는데, 어떠한 데이터가 주어진 경우, 이 데이터를 통해서 확률 P(H)를 추정할 때 사용한다.

일반화

- 확률질량함수 $f0$ 가 있다고 가정해보자. $X = (x1, x2, x3, ..., xn)$ 는 해당 확률로 측정되는 데이터이다.
- 만약 observation x가 주어진다면, θ의 값만 알 수 있다면 바로 $f(x|\theta)$의 값을 계산할 수 있다.
- Likelihood는 다음과 같이 표현할 수 있다.
  - $L(\theta; x_1, x_2, ..., x_n) = L(\theta; X) = f(X|\theta) = f(x_1, x_2, ..., x_n|\theta)$
- Maximum Likelihood Estimation (MLE)는 θ를 추정하는데, Likelihood를 최대로 만드는 값으로 선택한다.

- 만약 우리가 선택하는 값을 $\theta\hat{}$ 라고 적는다면, MLE는 다음과 같은 방식으로 값을 찾는다.
  - $\theta\hat{} = argmax_\theta L(\theta; X) = argmax_\theta f(X|\theta)$
- 우리의 관측값이 독립적이라면, $f(X|\theta) = \prod_i f(x_i|\theta)$ 이 된다. 독립사건이므로 곱해나간다.
  - 로그 가능도 함수를 사용하면, 즉 로그를 취하면 곱셈을 덧셈으로 바꿀 수 있어 편리하다.

한계

- MLE는 관찰 값에 따라 너무 민감하게 변한다는 단점을 가지고 있다.
- 동전 던지기는 확률 과정이기 때문에 극단적인 경우로 n번을 던져서 앞면이 n번이 나올 수가 있다.
- 이 경우 MLE는 이 동전은 앞면만 나오는 동전이라고 판단해버린다.
- 스팸필터를 만드는데 연속으로 스팸이 아닌 메일이 n개가 들어왔다고 해서 모든 메일이 스팸이 아니라고 할 수는 없다.

기타

- 가능도함수와 확률밀도함수의 차이를 요약하면 다음과 같다.
  - 확률밀도함수는 모수의 값을 이미 알고 있는 경우, 변수 x 의 상대적 확률(모수)를 구하는 것이다. 적분하면 1이 나온다.
  - 가능도함수는 x가 이미 발생했고 값을 이미 알고 있다. 이 때 모수 값을 구하는 것으로 적분하면 전체 면적이 1이 아닐 수 있다.

베이즈 추정 Baysian Estimation

Bayes' Rule

Bayes' Rule

$$p(\theta|x) = \frac{p(\theta)f(x|\theta)}{p(x)}$$

- 베이즈 확률에는 크게 두 가지의 요소가 존재한다.
  - $P(A)$: A의 사전확률 (a prioi). 어떠한 사건에 대한 정보가 없을 때의 확률.
  - $P(A|B)$: B에 대한 A의 사후확률 (posteriori). B라는 정보가 주어졌을 때의 확률.
- 좌변에 있는 수식은 Posterior이다. 어떤 관측치 X 가 주어졌을 때 모수 θ 를 가지는 확률모형이다.
- Prior는 확률모형으로, 알려지지 않았으므로 해당 모수의 분포를 파악해둔다. (Prior Distribution)
  - 따라서 작위성이 들어감은 당연하다.
- Likelihood는 가능도이다. 주어지는 parameter에 따라 관측치를 얻을 수 있는 확률을 의미한다.
- Posterior 분포는 Prior와 Likelihood의 곱에 비례한다고 얘기할 수 있다.
  ($Posterior \propto Prior \ * Likelihood$)

- 예시
  - 내 친구 A가, 자신이 카페에 갔었는데 모르는 사람이 자기 번호를 따갔다고 자랑을 하였다.
  - 이 때, 그 번호를 알아 간 사람이 여자일 확률 $P(W)$는 0.5(남자일 확률 $P(M)$ = 0.5) 이다.
  - 그런데 친구가 그 번호를 알아간 사람의 머리카락 길이가 어깨 아래까지 길었다고 주장했다.
    - 머리가 어깨 아래까지 길었다는 사건: L
    - 아무 조건이 없이 누군가 번호를 알아갔다고 할 때의 $P(W)$는 0.5 이지만, 여기에 한가지 정보가 추가 되었을 때의 확률 $P(W)$는 당연히 변화한다.
  - 그렇다면, 조건부 확률 계산 방식에 의해 다음과 같이 계산이 가능하다.

$$P(W|L) = P(W \cap L)/P(L)$$

  - 그런데 실제 세계에서는 번호를 알아간 사람이 여자일 확률과 머리가 어깨 아래까지 길었다는 사건의 교집합을 찾기 어려울 것이다. 다시 말해서, $P(W \cap L)$ 의 사건은 변수가 많아지면 계산하기가 복잡해질 것이다.
  - 따라서 다음과 같은 식으로 변환하여 계산할 수 있다.

$$P(W|L) = P(L|W)P(W)/P(L)$$

  - 위 식의 장점은 무엇인가? $P(L|W)$ 만 알게 되면 $P(W|L)$ 를 계산할 수 있기 때문이다.

Maximum A Posterior(MAP)

- 베이즈 추정법은 모숫값이 가질 수 있는 모든 가능성의 분포를 계산하는 작업이다.
- 예를 하나 들어보자. 일반적으로 동전을 던졌을 때 질량의 분포가 균등하게 있다고 가정하자.
  - 동전을 100번 던졌다고 가정하자. 그 결과로 70번의 앞면이 나왔다.
  - 이 경우, MLE에서는 앞면이 나올 확률을 0.7로 가정하여 그 과정에 대한 확률을 구했다.
  - 어, 그런데 질량의 분포가 균등하게 있으면 앞면이 나올 확률은 0.5임은 자명하다. 이 사전확률은 무시된 것이다.
- MAP에서는 Posterori를 활용한다. 다음 식에 사전에 우리가 알고 있는 정보를 대입한다.

$$P(T|E) = P(E|T)P(T)/P(E)$$

- 알고자 하는 확률 '100번 동전을 던진 시행에서 70번의 앞면이 나왔을 때, 동전의 앞면이 나올 확률' 공식에, 이론 T에 특정 가정을 대입하고, 그에 대한 확률을 계산한다. 만약 T = 0.5 라는 가정을 기준으로 계산하면,

$$P(T = 0.5|E = 0.7) = P(E = 0.7|T = 0.5)P(T = 0.5)/P(E = 0.7)$$

- 여기서 $P(E = 0.7|T = 0.5)$는 Likelihood Function에서 구할 수 있고, $P(E = 0.7)$는 변수에 특정 값을 대입했을 때 나온 값이 아닌 상수이다.
- 즉, 우리가 이전에 알고 있던 '동전을 던졌을 때 앞면이 나올 확률은 0.5 이다' 라는 명제에 대한 사전 확률인 $P(T = 0.5)$만 주어진다면 Posteriori(사후확률)를 계산할 수 있다.

언제 사용하나

- MAP 방법은 θ가 주어지고, 그 θ에 대한 데이터들의 확률을 최대화하는 것이 아니라, 주어진 데이터에 대해 최대 확률을 가지는 θ를 찾는다. 수식으로 표현하면 다음과 같다.

$$\hat{\theta} = argmax_\theta L(x;\theta)$$

- 우리가 원하는 것은 general한 설명이다. 여러 모수들 중에서 데이터가 주어졌을 때 가장 확률이 높은 θ를 고를 수 있다면 가장 좋은 결과를 얻을 수 있을 것이다.

MLE vs MAP

> 다음의 예시는 해당 블로그 에서 참조했습니다.

> "한 아저씨는 한강공원을 산책하고 있습니다. 산책 도중 버려진 아이폰을 하나 발견했습니다. 자세히 보니 아이폰은 작년 말에 나온 11 Pro였습니다. 생각해보니, 애플은 뭇 젊은 친구들에게 인기가 많은 스마트폰입니다. (20대 구매율 90%) 아저씨는 경찰서에 이를 돌려주고 연락처를 남겼습니다. 아이폰의 주인이 20대일 확률은 얼마일까요?"

- MLE와 MAP는 서로 다른 주장을 펼칩니다.
  - MLE: 가장 큰 Likelihood를 비교하자! $P(자전거브랜드|남자), P(자전거브랜드|여자)$ 의 확률을 비교하자!
  - MAP: 가장 큰 Posterior를 비교하자! $P(남자|자전거브랜드), P(여자|자전거브랜드)$ 의 확률을 비교하자!
  - MLE는 순수히 남, 여 중에서 해당 자전거 브랜드를 갖고 있을 확률을 구한다.
  - 반면에 MAP는 자전거 구매자의 성비까지 함께 고려한 확률을 구한다.
  - MAP로 구하기

  $P(남자|자전거브랜드) = p(남,자전거브랜드)/p(자전거브랜드) = p(남,자전거브랜드)p(남,자전거브랜드) + p(여,자전거브랜드)$

베이지언 확률

> 아래의 예시는 다음의 블로그 링크 를 참조했습니다.

- 눈 앞에 한 명의 사람이 서 있다. 이 사람은 커튼에 가려져 있다. 우리가 볼 수 있는 것은 형상 뿐이다.
- 이 사람은 철수인가, 영희인가? 확률적 classification 문제가 된다.
- 조건부 확률로 생각해 보면 $P(눈앞의사람|철수)$와 $P(철수|눈앞의사람)$의 2가지 확률을 생각할 수 있다.
- $P(눈앞의사람|철수)$는 철수가 눈 앞의 사람의 형상을 나타낼 확률이다. 이는 likelihood이다.
- $P(철수|눈앞의사람)$은 눈앞의사람의 형상을 나타내는 사람이 철수일 확률이다. 이는 posterior 확률이다.

- MLE는 가능한 사람들에 대해서 $P(\text{눈앞의사람}|\text{철수}), P(\text{눈앞의사람}|\text{영희}), ...$ 등을 각각 계산해서 그중 likelihood가 가장 높은 사람을 선택하는 방식이다.
- MAP는 posterior 확률 $P(\text{철수}|\text{눈앞의사람}), P(\text{영희}|\text{눈앞의사람}), ...$ 을 계산해서 가장 확률이 높은 사람을 선택하는 방식이다.
- $P(\text{철수}|\text{눈앞의사람})$은 다음과 같이 계산한다.

  $$P(\text{철수}|\text{눈앞의사람}) = \frac{P(\text{눈앞의사람}|\text{철수})P(\text{철수})}{P(\text{눈앞의사람})} \quad \text{--- (1)}$$

  - 식 (1)에서 P(눈앞의사람)은 고정된 상수값이다.
  - MAP에서는 결국 $P(\text{눈앞의사람}|\text{철수})P(\text{철수}), P(\text{눈앞의사람}|\text{영희})P(\text{영희}), ...$ 들 중에서 가장 높은 값을 나타내는 사람을 선택하는 것이 된다.
  - 결과적으로 ML와 MAP의 차이는 뒤에 $P(\text{철수}), P(\text{영희}), ...$ 과 같은 확률을 곱하느냐 곱하지 않느냐의 차이가 된다.
  - 앞서 살펴보았듯이 $P(\text{철수}), P(\text{영희}), ...$ 와 같이 클래스 고유의 확률값을 prior 확률이라고 부릅니다.
    - 만일 눈 앞의 사람형상이 남자인지 여자인지를 구분하는 문제라고 생각하면 prior 확률이 이해가 갑니다.
    - $P(\text{눈앞의사람})$은 가능한 사람이 철수, 영희, 까치라고 할 때 다음과 같이 계산된다.

      $$P(\text{눈앞의사람}) = P(\text{눈앞의사람}|\text{철수})P(\text{철수}) + P(\text{눈앞의사람}|\text{영희})P(\text{영희}) + P(\text{눈앞의사람}|\text{까치})P(\text{까치})$$
- ML와 MAP 중 어느것이 보다 정확한 방법인가?
  - MAP이 보다 정확한 방법이다.
  - 커튼에 비친 사람의 형상을 보고 그 사람이 누구인지 맞춘다고 했을 때, $P(\text{철수}|\text{눈앞의사람}), P(\text{영희}|\text{눈앞의사람}), ...$ 가 가장 높은 사람을 선택하는 것이 원래의 문제를 있는 그대로 푸는 것이다.
  - ML로 문제를 푸는 것은 모든 prior 확률이 같다는 가정($P(\text{철수}) = P(\text{영희}) = ...$)하에 일종의 근사적인 해를 구하는 것이다.
  - 하지만 MAP로 문제를 풀더라도 식(1)에 의해서 likelihood 식으로 변환하여 계산하기 때문에 어느 경우나 결국은 likelihood의 계산이 필요하게 됩니다.
- $likelihood\ P(\text{눈앞의사람}|\text{철수})$ 를 직접 계산하기 위해서는 철수에 대한 사진 수만장을 찍고 여기서 나온 형상들의 확률 분포를 구해야 한다.
  - 그러나 이러한 접근은 상당히 비효율적이며 형상의 같고 다름에 대한 기준도 모호하다.
- 따라서, 대부분의 실제 문제에 있어서는 사람의 형상에 대한 주요 특징(키, 머리크기, 허리둘레 등)을 선별하여 이 특징들에 대한 확률분포를 이용한다.
- 즉, $P(\text{철수}|\text{눈앞의사람}), P(\text{눈앞의사람}|\text{철수})$ 대신에 $P(\text{철수}|\text{키}, \text{머리크기}, \text{허리둘레}, ...), P(\text{키}, \text{머리크기}, \text{허리둘레}, ...|\text{철수})$를 이용하여 classification 문제를 푼다.

Naive Baysian 방식

- 사람의 키, 머리크기, 허리둘레가 서로 상관관계가 없는 독립변수라고 가정한다.
- 키의 확률분포, 머리크기의 확률분포, 허리둘레의 확률분포를 각각 구한 후 각각의 확률을 서로 곱하여 결합확률을

계산한다.

$$P(키,머리크기,허리둘레|철수)=P(키|철수)P(머리크기|철수)P(허리둘레|철수)$$

Pure Baysian 방식

- 사람이 가질 수 있는 모든 (키,머리크기,허리둘레) 조합에 대하여 확률분포를 계산한 값을 사용한다.
- 키, 머리크기, 허리둘레를 동시에 종속적이라고 가정하고, 3차원의 확률분포에서 계산된 값을 이용한다.

두 방식의 차이를 비교해보자

- 사람의 키는 0 또는 1의 값을 가지고, 머리크기도 0 또는 1의 값만을 갖도록 히스토그램을 구한 후 정규화했다고 가정해보자.
- 철수의 키와 머리크기의 확률분포가 다음과 같다.

`<표1> P(키,머리크기|철수)의 확률분포`

P(키, 머리크기|철수)	키=0	키=1
머리크기=0	0.3	0.1
머리크기=1	0.2	0.4

- 동일한 샘플에 대해 키, 머리크기의 확률분포를 독립적으로 구했다면 그 결과는 다음과 같다.

`<표2> P(키|철수)의 확률분포`

	키=0	키=1
P(키|철수)	0.5	0.5

`<표3> P(머리크기|철수)의 확률분포`

	머리크기=0	머리크기=1
P(머리크기|철수)	0.4	0.6

- 이제 $P(키=0,머리크기=0|철수)$를 구해보자.
  - pure Bayesian으로 구해보면 0.3이다.
  - naive Bayesian으로 구해보면 0.5*0.4 = 0.2이다.
  - 서로 다른 값이 나옴을 확인할 수 있다.
- 이 때, pure Bayesian으로 계산한 0.3이 올바른 확률값이다.
  - naive Bayesian에서 오차가 있는 값이 나오는 이유는 feature간의 상관관계, 즉 키와 머리크기가 가지는 상관관계를 무시하고 확률을 계산했기 때문이다.
  - 만약 상관관계가 없다면 naive Bayesian으로 계산된 확률값도 맞는 값이 된다.
  - 그 점에서 Prior(= P(θ) P(θ))가 Uniform Distribution인 경우에는 모든 Prior값은 $1/n$로 Constant화

되고, MLE는 MAP의 특수한 경우라고 생각해볼 수 있는 것이다.

$$\theta_{MAP} = \arg\max_{\theta} \sum_i \log P(x_i|\theta)P(\theta)$$

$$= \arg\max_{\theta} \sum_i \log P(x_i|\theta)\, const$$

$$= \arg\max_{\theta} \sum_i \log P(x_i|\theta)$$

$$= \theta_{MLE}$$

- 하지만 pure Bayesian 확률분포는 구하기가 어렵기 때문에 오차를 감수하고 간단한 naive Bayesian으로 문제를 단순화시켜서 푸는 경우가 많다.
    - pure Bayesian 방식으로는 3^10 = 59,049 가지 경우에 대한 확률을 계산해야 하지만 naive 방식으로는 3*10 = 30개의 확률값만 계산하면 된다.
- semi-naive Bayesian은 feature들을 먼저 소그룹으로 그룹핑(grouping)을 한 후에 각 그룹 내에서는 feature들 간의 상관관계를 풀(full)로 계산하되, 그룹과 그룹 사이에서는 상관관계가 없는 것으로 확률을 계산하는 방식이다.

Reference

- http://databaser.net/moniwiki/pds/BayesianStatistic/%EB%B2%A0%EC%9D%B4%EC%A6%88_%EC%A0%95%EB%A6%AC%EC%99%80_MLE.pdf
- https://niceguy1575.tistory.com/87
- https://medium.com/@youngji/%EC%B5%9C%EB%8C%80-%EA%B0%80%EB%8A%A5%EB%8F%84-%EB%B0%A9%EB%B2%95-maximum-likelihood-method-a8546e44c1a3
- https://rpubs.com/Statdoc/204928
- https://www.youtube.com/watch?v=sOtkPm_1GYw
- http://sanghyukchun.github.io/58/
- https://jjangjjong.tistory.com/41
- https://forensics.tistory.com/46