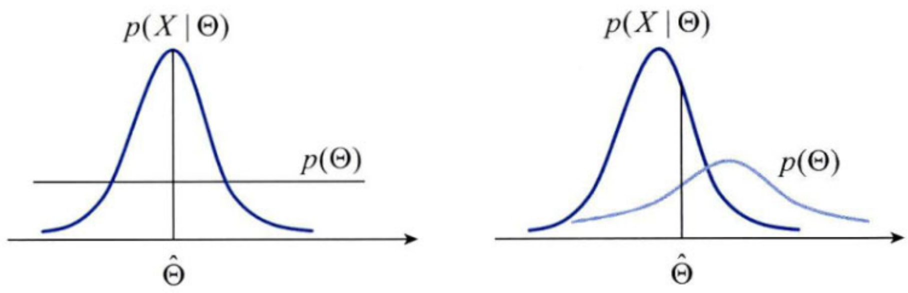
Likelyhood(우도;가능도)
가능도 : - 어떤 결과(증거, 데이터, 사건) X가 발생했을 때 이 X값이 나오게 하는 θ의 가능성 - 주어진 데이터 x에 대해, 모수(parameter) θ를 변수로 둔 함수 - 데이터가 주어져있는 상황에서, θ를 변형시킴에 따라 값이 바뀌는 함수 - 모수 θ를 따르는 분포가 데이터 x를 관찰할 가능성 **가능도함수** 확률변수 $X_1, X_2, \cdots, X_n$의 결합 확률밀도함수가 $f(x_1, x_2, \cdots, x_n; \theta)$라고 하자. 결합 확률밀도함수 $f(x_1, x_2, \cdots, x_n; \theta)$는 고정된 모수 θ에 대하여 $(x_1, x_2, \cdots, x_n)$의 함수로 사용된다. 그러나 반대로 $f(x_1, x_2, \cdots, x_n; \theta)$를 관측치 $X_1 = x_1, \cdots, X_n = x_n$이 주어졌을 때 모수 θ의 함수로 생각해 볼 수 있다. 즉, $X_1, X_2, \cdots, X_n$의 가능도함수 $L(\theta)$를 다음과 같이 정의한다. $L(\theta) = L(\theta; x_1, x_2, \cdots, x_n) = f(x_1, x_2, \cdots, x_n; \theta)$ Discrete probability distribution : $L(\theta \vert x) = p_\theta(x) = P_\theta(X = x)$ Continuous probability distribution : $L(\theta \vert x) = f_\theta(x)$
MLE(최대우도추정)
최대우도방법(Maximum Likelihood Estimator) - 확률 P(X;θ) - Likelihood P(θ;X)
최대가능도 추정량 랜덤표본의 가능도 함수 $L(\theta; x_1, x_2, \cdots, x_n)$을 최대화하는 θ의 값을 $\hat{\theta} = \hat{\theta}(X_1, X_2, \cdots, X_n)$라고 할 때, $\hat{\theta}$을 모수 θ의 최대가능도 추정량이라 한다. - 최대가능도 추정량을 계산하는 데 있어서 로그함수를 이용 - 로그함수는 단조 증가함수로 L(θ)를 최대로 만드는 θ는 Log(L(θ)) 또한 최대로 만들게 됨 - 이를 로그가능도함수라고 합니다.
MLE 3가지 가정: 아무런 가정도 없었던 최소제곱법(LSE)와는 다르게 MLE는 아래와 같은 가정이 필요 1. 오차의 평균 0 2. 오차의 분산 σ^2 3. 오차는 정규분포를 따른다. MLE는 3가지 특징 1. 일치성(not unbaised) 2. 점근 효율성 3. 점근 정규성

베이지안과 우도(likelihood)와의 관계

$$\underbrace{P(\theta|X)}_{\text{A posterior probability}} = \frac{\overbrace{P(X|\theta)}^{\text{Likelihood}}\overbrace{p(\theta)}^{\text{A prior probability}}}{\underbrace{P(X)}_{\text{Marginal probability}}}$$

Bayes Rule을 이용한 방식의 가장 큰 단점들 중 하나는 Likelihood의 Probability Distribution을 알아야 함.
하지만 대부분의 경우는 모름.. 그래서 우도를 이용하여 사전확률을 추정하는 방식이 필요
MLE는 MAP의 특수한 경우로써 a prior probability가 uniform distribution이라고 가정한 것



(a) ML 방법　　(b) MAP 방법

- 위 그림은 `ML`과 `MAP`를 비교합니다.
- **ML에서는 사전확률이 균일**하다고 가정합니다. 따라서 **likelihood가 최고인 점**을 찾으면 그것이 바로 최적해 θ가 됩니다.
- 하지만 MAP 에서는 **사전 확률이 균일하지 않습니다.** 따라서 **사전확률이 최적해에 영향**을 미치게 됩니다. 이러한 차이점이 있습니다.

관찰을 통해 Likelihood를 얻었지만, 까다로우며 완벽한 Distribution Function을 얻을 수 있는 것도 아님
그래서 우리는 Data로부터 직접 decision policy를 얻고자 함
정해지지 않은 몇 개의 parameter로 이루어진 함수로 모델링을 한 후에, 이 모델이 주어진 **Data를 가장 잘 설명하도록 parameter들을 구현해내는 방식을 택함**
이러한 방식을 이용하는 대표적인 알고리즘이 바로 **Deep Learning**
Deep Learning의 기본적인 Loss Function들은 대부분 **Maximum Likelihood Estimation(MLE)**과 **Maximum A Posterior(MAP)**를 통해 증명
또한 확률을 기반으로 하기 때문에 이 두 이론을 공부하고 나면 확률과 Deep Learning 사이의 연결고리를 파악하는 데 큰 도움이 될 것
MLE와 MAP에 관해 본격적으로 알아봐야 함!

https://hyeongminlee.github.io/post/bnn001_bayes_rule/ : 놓어 연어

MLE vs. MAP

$$P(\theta|X) = \frac{\overbrace{P(X|\theta)}^{\text{MLE}}\,\overbrace{p(\theta)}^{\text{MAP}}}{P(X)}$$

MLE의 단점을 해결하기 위해 Maximum a Posteriori Estimation(MAP)이라는 방법을 사용하기도 한다. 이 방법은 θ가 주어지고, 그 θ에 대한 데이터들의 확률을 최대화하는 것이 아니라, 주어진 데이터에 대해 최대 확률을 가지는 θ 찾는다. 수식으로 표현하면 다음과 같다.

MLE :

$$\hat{\theta} = \mathrm{argmax}_{\theta} L(\theta; X) = \mathrm{argmax}_{\theta} f(x|\theta)$$

MAP :

$$\hat{\theta} = \arg\max_{\theta} f(\theta|X) = \arg\max_{\theta} \frac{f(X|\theta) f(\theta)}{f(X)} = \arg\max_{\theta} \frac{\mathcal{L}(\theta; X) f(\theta)}{f(X)}$$

MAP : 대표적으로 나이브 베이지안 분류기(y = theta)

$$f^{\mathrm{Naive}}(X) = \underset{j \in \{1,2,...,J\}}{\mathrm{argmax}} P(y=j) \prod_{l=1}^{p} P(X_l | y=j)$$

가정 : conditional independence(독립변수의 조건부 확률에 조건부 독립)

$$P(X|y=j) = P(X_1, X_2, \ldots, X_p | y=j) = \prod_{l=1}^{p} P(X_l | y=j)$$

$$= \prod_{i=1}^{n} P(X_i | y_j) P(y_j)$$

추정방법

나이브 베이즈 분류기를 구하기 위해서는 y의 주변분포 P(y = j)와 특정 범주가 주어졌을 때 독립변수의 조건부 확률 분포 P(X_l|y=j)를 알아야 함. 하지만 대부분의 경우 이를 알 수가 없다. 따라서 데이터를 이용하여 각 확률을 추정해야 한다.

유형

1. 독립변수가 범주형인 경우
2. 독립변수가 연속형인 경우
3. 독립변수가 2개 이상의 특정 카테고리에 대한 빈도수인 경우

$$P(y|x_1, \cdots, x_n) = \frac{P(x_1|y)P(x_2|y) \cdots P(x_n|y)p(y)}{P(x_1)P(x_2) \cdots P(x_n)}$$

여기서 베이즈 법칙에 의하면 X가 주어졌을 때 y의 조건부 확률은 다음과 같이 쓸 수 있다.

$$P(y=j|X) = \frac{P(y=j)P(X|y=j)}{P(X)}$$
$$\propto P(y=j)P(X|y=j)$$

여기서 마지막 비례 표시는 y와 상관없는 부분에 대하여 비례한다는 뜻이다.

나이브 베이즈 예시 : https://m.blog.naver.com/tommybee/222663271120