* 베이즈 정리   Bayes formula

- $P(y, x) = P(x \mid y)P(y)$ // 일반적으로 x와 y가 같이 일어난 결합확률과

  y와 x가 같이 일어날 결합확률은 같음.

  $= P(x, y)$

  $= P(y \mid x)P(x)$

- $\boxed{P(y \mid x)} = \dfrac{P(x \mid y)P(y)}{P(x)}$

likelihood

prior probability

posterior probability

x : 이미 알고있는 사건

y : 추정해야할 사건

조건부 확률   p(추정해야할 사건 | 이미 알고 있는 사건) = $\boxed{P(y \mid x)}$

우도   p(이미 알고있는 사건 | 추정해야할 사건) = $P(x \mid y)$ = 역확률문제.

$= \mathcal{L}(y \mid x)$ // 때로 우도를 $\mathcal{L}$기호이용해

x와 y 위치 바꿔 표현하기도 함

* ML에 적용 $X \in R^4$ $y \in \{class1, class2, class3\}$

- 특징 벡터 X. 부류 y.

$$X = (x_1, x_2, x_3, x_4)^T \xrightarrow{\text{posterior 추정}}$$

7.0 3.2 4.7 1.4

$p(y = class1 \mid x) = 0.18$

$p(y = class2 \mid x) = 0.72$

$p(y = class3 \mid x) = 0.10$

$\xrightarrow{\text{argmax}}$ $y = class2$.

- 패턴인식

$$\hat{y} = \underset{y}{\text{argmax}}\, \boxed{P(y \mid x)}$$

posterior

- 실제로 $\boxed{P(y \mid x)}$를 직접 추정하는 것은 아주 단순한 상황을 빼고는 불가능.

  X가 만드는 공간은 무수히 많은 점을 가지므로, 이 모든 점에 대해 일일이 확률을 표현할 수 없음

⇨ $P(y)$와 $P(x \mid y)$를 구할 수 있다면, bayes formula 이용해 $\boxed{P(y \mid x)}$를 간접적으로 계산가능

prior   likelihood.

① 사전확률 prior $p(y)$

- 무작위로 n개 sample을 수집하여, 각 클래스별로 n_1, n_2, n_3개 라면

$$p(y=c_i) = \frac{n_i}{n}$$

c_i = i번째 class

② 우도 likelihood. $p(x|y)$

- $p(x|y)$는 $\boxed{p(y|x)}$보다 훨씬 추정하기 쉬움.

y의 class가 고정된 셈이므로 다른 class의 sample을 모두 배제한 채
y에 속한 sample만 가지고 확률 분포를 추정하면 되기 때문.

- 우도 추정 방법 likelihood estimation
  density estimation.
  - 여러가지 확률 밀도 추정 방법 존재. sample이 밀집된 곳이 확률은 높고. 듬성한 곳 확률은 낮게 추정.
  - 파젠창 방법 (어떤 점에 일정한 크기 마스크를 씌우고, 점 속으로 들어오는 sample 개수 세어 확률 추정)
    - 가우시안 같은 매개 변수로된 확률분포함수 가정하고 매개변수 추정.
    - 가우시안을 여러개 혼합해 정확도 높이는 방법

* 최대 우도 추정 Maximum likelihood estimation MLE.

- 우도를 이용한 목적함수는 딥러닝에서 널리 사용
- 추정해야 할 매개변수 집합 θ.

데이터 X가 주어졌을 때, X를 발생시켰을 가능성을 최대로 하는 매개변수 θ의 값을 찾아라.
우도 최대나 하는 해 구해라

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\, P(X \mid \theta) \quad \text{likelihood}$$

이미 알고 있는 사건 추정해야 할 사건

- 데이터 $X = \{x_1, x_2, \dots x_n\}$ 는 독립 동일분포 (independent and identically distributed) 이므로, sample을 독립적으로 다룰 수 있음.

$$P(X \mid \theta) = p(x_1, x_2, \dots x_n \mid \theta) = \prod_{i=1}^{n} P(x_i \mid \theta)$$

- sample 개수인 n이 보통 수천을 넘으므로, 확률을 n번 곱하면 너무 작은 값이 되어 계산 시, 버려질수있음
- 최대 로그 우도 추정 maximum log likelihood estimation.

$$\hat{\theta} = \underset{\theta}{\arg\max}\, \log P(X \mid \theta) = \underset{\theta}{\arg\max} \sum_{i=1}^{n} \log P(x_i \mid \theta)$$

(log는 단조증가 함수 이므로, 기존 식의 최적해 $\hat{\theta}$와 로그 적용식 최적해 $\hat{\theta}$는 동일)

* ML에 적용 (딥러닝에 적용)

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(X \mid W)$$

추정해야 할 매개변수.
신경망 가중치 집합.

ex. 신경망 784-100-100-100-10 구조, 매개변수 개수 99.400개

목적함수와 최적화 알고리즘을 이용해 매개변수의 최적의 조합을 찾아냄.