Naive Bayesian

나이브 베이즈는 **스팸 메일 필터, 텍스트 분류,감정 분석,추천 시스템** 등에 광범위하게 활용되는 분류 기법이다.

머신러닝을 통해 어떤 동물의 사진이 있을 때 그 동물이 호랑이인지 고양이인지 얼룩말인지 등을 구분할 수 있다. 사전에 수많은 호랑이, 고양이, 얼룩말 사진에 대해 학습을 시킨다. 다양한 자세, 표정, 생김새, 털의 색깔 등을 가진 호랑이, 고양이, 얼룩말에 대해 '이 사진은 호랑이고, 이 사진은 고양이야'라고 학습시키는 것이다. 학습된 머신러닝 모델은 이제 호랑이, 고양이, 얼룩말을 정확히 분류할 수 있다. 이제는 학습시 사용되었던 사진이 아닌 새로운 사진을 갖다 줬을 때도 정확히 분류할 수 있다. 이렇게 사전 **데이터를 기반으로 충분히 학습시키는 방법을 지도학습(Supervised learning)이라고 한다.**

지도학습을 하기 위한 첫 스텝은 Feature와 Label을 파악하는 것이다. Label은 우리가 원하는 분류 결과다. 위 예시에서는 호랑이, 고양이, 얼룩말이 Label이다. 이 Label 결과에 영향을 주는 요소를 Feature라고 한다. 동물의 자세, 표정, 생김새, 털의 색깔 등이 바로 Feature다. 즉 수많은 동물의 자세, 표정, 생김새, 털의 색깔(Feature)을 기반으로 그 동물이 호랑이인지 고양이인지 얼룩말인지(Label) 분류를 하는 것이다.

나이브 베이즈 분류는 지도학습의 일종이다. 따라서 Feature와 Label이 필요하다. Feature에 따라 Label을 분류하는데 베이즈 정리를 사용하는 것이 특징이다. 또한 모든 Feature가 서로 독립(independent)이어야 한다는 가정이 필요하다.

Tennis playing으로 예시를 들어보자!

	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Today is Sunny,Mild,High and Strong

- Will he play tennis today?

How can you predict his tennis playing?

- Let's use Probability

  P(yes|sunny,mild,high,strong)

  P(no|sunny,mild,high,strong)

Naive bayes는 주로 nominal value에 사용된다.(discrete하고 대소관계X)

Naive Bayesian Classifier

- Probability with the **strongest assumption** on independence

If Output is given,
Inputs are independent from each other.

$$P(yes \mid sunny, mild, high, strong) = \frac{P(sunny, mild, high, strong \mid yes)P(yes)}{P(sunny, mild, high, strong)}$$

$$P(no \mid sunny, mild, high, strong) = \frac{P(sunny, mild, high, strong \mid no)P(no)}{P(sunny, mild, high, strong)}$$

$P(yes), P(no)$

$P(sunny, mild, high, strong \mid yes)$
$= P(sunny \mid yes)P(mild \mid yes)P(high \mid yes)P(strong \mid yes)$

$P(sunny, mild, high, strong \mid no)$
$= P(sunny \mid no)P(mild \mid no)P(high \mid no)P(strong \mid no)$

By assumption of indepenence

$P(sunny, mild, high, strong) = ??$

Do we need to evaluate this?
Can we evaluate this?

- Let's ESTIMATE the followings, but How?>>아래식들은 이 tabel로부터 estimate를 해야한다.

$P(yes)$,
$P(no)$,
$P(sunny \mid yes)$,
$P(mild \mid yes)$,
$P(high \mid yes)$,
$P(strong \mid yes)$,
$P(sunny \mid no)$,
$P(mild \mid no)$,
$P(high \mid no)$,
$P(strong \mid no)$

	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

$$P(sunny, mild, high, strong) = \alpha$$

$$P(yes \mid sunny, mild, high, strong)$$
$$= \frac{P(yes)P(sunny \mid yes)P(mild \mid yes)P(high \mid yes)P(strong \mid yes)}{P(sunny, mild, high, strong)}$$
$$= \frac{1}{\alpha}\left(\frac{9}{14} \times \frac{2}{9} \times \frac{4}{9} \times \frac{3}{9} \times \frac{3}{9}\right) = \frac{0.007055}{\alpha}$$

$$P(no \mid sunny, mild, high, strong)$$
$$= \frac{P(no)P(sunny \mid no)P(mild \mid no)P(high \mid no)P(strong \mid no)}{P(sunny, mild, high, strong)}$$
$$= \frac{1}{\alpha}\left(\frac{5}{14} \times \frac{3}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5}\right) = \frac{0.04114}{\alpha}$$

- Can we evaluate $P(sunny, mild, high, strong)$

$$P(yes \mid sunny, mild, high, strong) + P(no \mid sunny, mild, high, strong) = 1$$

$$= \frac{0.007055}{\alpha} + \frac{0.04114}{\alpha} = 1$$

$$\alpha = P(sunny, mild, high, strong) = 0.048195$$

그런데 굳이 알파값을 구해야하나? 왜냐면 우리는 strong assumption을 주어서 한거기 때문에 저 알파값이 무조건 옳은것이 아니다. 또한 위 예시에서 tennis를 칠지 안칠지를 구하는 것에 알파값은 필요가 없다.

Formal Description of NBC

- Assume target function f : X→V
- What is the most probable value of f(x) →where x is (a1,a2,...an)
- Let's estimate as follows:

$$f(x) = \arg\max_{v \in V} P(v \mid a_1, a_2, \cdots, a_n)$$

  - What does it mean?

v는 target value이다. tennis를 예시로하면 저값이 0.5이상이면 치고 미만이면 안친다 이런느낌

  - It is equivalent

$$f(x) = \arg\max_{v \in V} \frac{P(a_1, a_2, \cdots, a_n \mid v)P(v)}{P(a_1, a_2, \cdots, a_n)}$$
$$= \arg\max_{v \in V} P(a_1, a_2, \cdots, a_n \mid v)P(v)$$

- However, We don't know the values of

$$P(a_1, a_2, \cdots, a_n \mid v), P(v)$$

- Let's estimate those with GIVEN DATA
  - Still, we may have enough data to estimate

$$P(a_1, a_2, \cdots, a_n \mid v)$$

  - How many data do we need?
- **Let's add an assumption(a strong assumption)**

$$P(a_1, a_2, \cdots, a_n \mid v) = \prod_i P(a_i \mid v)$$

If Output is given,
Inputs are independent from each other.

What's next?

- Estimate those based on given example

$$P(v) \text{ and } P(a_i) \text{ for } i = 1, \cdots, n$$

- Putting all the probabilities and choose v

$$f(x) = \arg\max_{v \in V} P(v) \prod_i P(a_i \mid v)$$

Performance

- Conditional indepengdence assumption is often violated but it is **easy to implement and works surprisingly well anyway**

Discussion

- Let's do it once more Hwith the following data

	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Rain	Mild	High	Weak	Yes
4	Rain	Cool	Normal	Strong	No
5	Sunny	Mild	High	Weak	No
6	Rain	Mild	Normal	Weak	Yes
7	Overcast	Mild	High	Strong	Yes
8	Rain	Mild	High	Strong	No

- Today is Sunny, Mild, High and Strong
  - Will he play tennis today?

$P(yes) = 3/8$

$P(no) = 5/8$

$P(sunny \mid yes) = 0/3$

$P(mild \mid yes) = 3/3$

$P(high \mid yes) = 2/3$

$P(strong \mid yes) = 1/3$

$P(sunny \mid no) = 3/5$

$P(mild \mid no) = 2/5$

$P(high \mid no) = 4/5$

$P(strong \mid no) = 3/5$

	Outlook	Temp	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Rain	Mild	High	Weak	Yes
4	Rain	Cool	Normal	Strong	No
5	Sunny	Mild	High	Weak	No
6	Rain	Mild	Normal	Weak	Yes
7	Overcast	Mild	High	Strong	Yes
8	Rain	Mild	High	Strong	No

→ ??

$$P(yes \mid sunny, mild, high, strong)$$
$$= \frac{P(yes)P(sunny \mid yes)P(mild \mid yes)P(high \mid yes)P(strong \mid yes)}{P(sunny, mild, high, strong)}$$
$$= \frac{1}{\alpha}\left(\frac{3}{8} \times \frac{0}{3} \times \frac{3}{3} \times \frac{2}{3} \times \frac{1}{3}\right) = 0$$

$$P(no \mid sunny, mild, high, strong)$$
$$= \frac{P(no)P(sunny \mid no)P(mild \mid no)P(high \mid no)P(strong \mid no)}{P(sunny, mild, high, strong)}$$
$$= \frac{1}{\alpha}\left(\frac{5}{8} \times \frac{3}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5}\right) = \frac{0.072}{\alpha}$$

어? 이러면 무조건 안치네?

If the size of given data or the number of instances for a specific value is small, NBC gives a poor performance

When n_ai,v is small:

$$P(a_i \mid v_j) = \frac{n_{a_i,v_j} + m \cdot p}{n_{v_j} + m}$$

n_v : number of training examples for which v=v_j

$n_{a_i,v}$: number of examples for which v=v_j and a=a_i

p is a prior estimate

m is weight given to prior

m은 적당한수>>p에대한 가중치

P는 p(ai | v)가 일어날확률을 사람이 예측한값: 0<p<1