

Ⅲ. 지도 학습 (Supervised Learning)

- Classification

④ Naïve Bayes

Supervised Learning

Regression

Linear Regression

Ordinary Least Squares
Regression

LOESS (Local Regression)

Neural Networks

Classification

Decision Trees

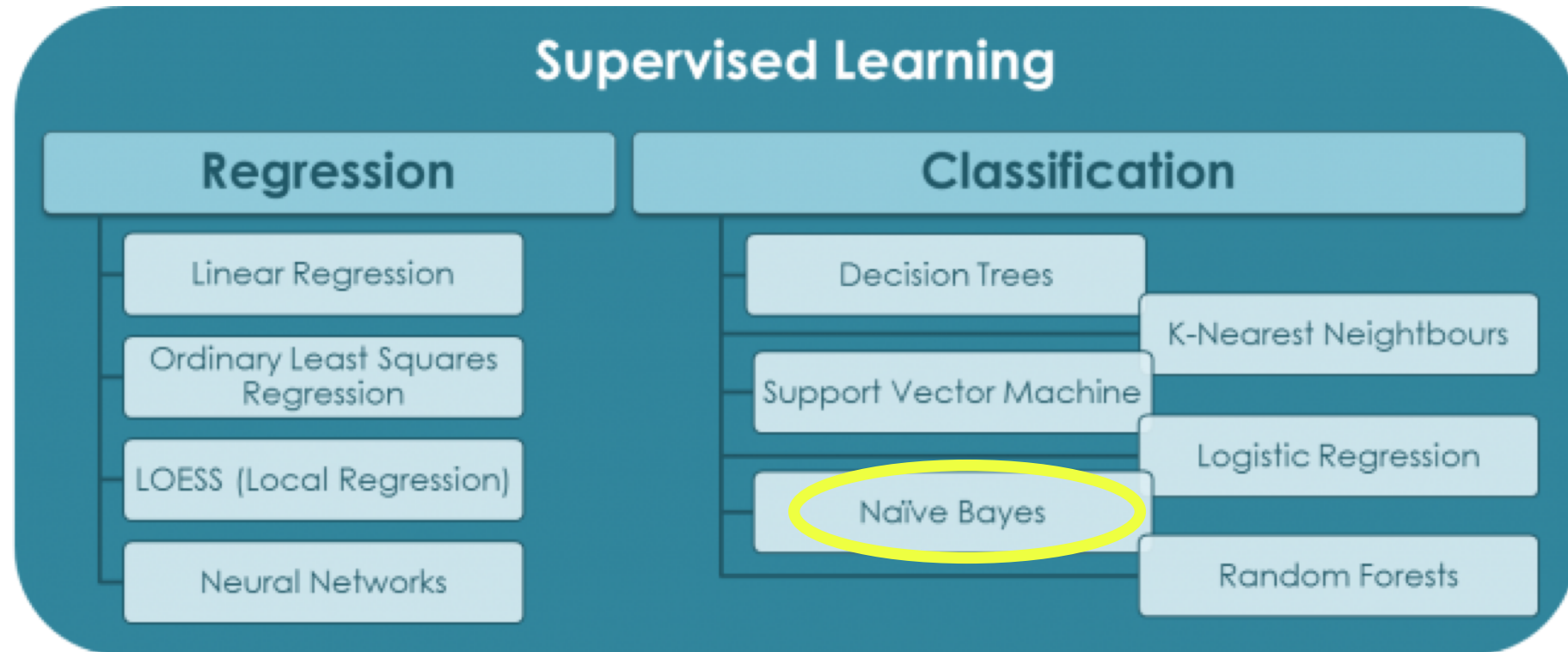
Support Vector Machine

Naïve Bayes

K-Nearest Neighbours

Logistic Regression


Random Forests

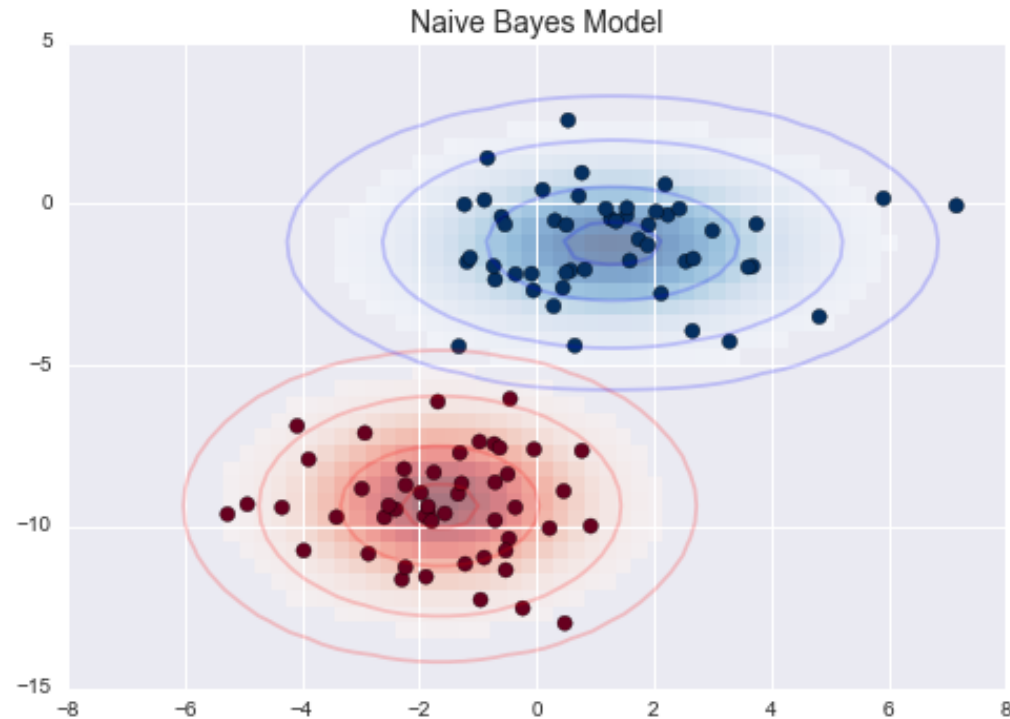


Naïve Bayes란?

- Bayesian rule에 근거한 classifier
- Naïve Bayes는 일종의 확률 모델로, 약간의 가정을 통해 문제를 간단하게 푸는 방법을 제안



$$P(A | B) = \frac{P(B | A)P(A)}{P(B)}$$
Two small, identical images of a cat's face are placed at the bottom left and bottom right of the equation.



Naïve Bayes란?

- 만약 데이터의 feature가 3개 있고, 각각이 binary라고 해보자.
- 예를 들어 남자인지 여자인지, 성인인지 아닌지, 키가 큰지 작인지 등의 feature를 사용해 사람을 구분해야 한다면, 이 경우 적어도 8개의 데이터는 있어야 모든 경우의 수를 설명할 수 있게 된다.
- 그런데 보통 데이터를 설명하는 feature의 개수는 이보다 훨씬 많은 경우가 많다.
- 예를 들어 feature가 10개 정도 있고 각각이 binary라면, 제대로 모든 데이터를 설명하기 위해서는 2^{10} , 약 1000개 이상의 데이터가 필요하다.
- 즉, 필요한 데이터의 개수가 feature 혹은 데이터의 dimension에 exponential하다.
- 이런 경우 그냥 Bayes rule을 사용해 분류를 하게 되면 overfitting이 되거나 데이터 자체가 부족해 제대로 된 classification을 하기 어려울 수 있다.

Naïve Bayes란?

- Naïve bayes는 이런 문제를 해결하기 위해 **새로운 가정을 하나** 하게 된다.
- 바로 모든 feature들이 i.i.d.하다는 것이다. (i.i.d는 independent and identically distributed의 준말로, 모든 feature들이 서로 independent하며, 같은 분포를 가진다는 의미이다.)
- 당연히 실제로는 feature들이 서로 긴밀하게 관련되어 있고 다른 분포를 가질 것이므로 이 가정은 틀린 가정이 될 수 있다.
- 그러나 만약 모든 feature가 i.i.d.하다고 가정하게 된다면 **우리가 필요한 최소한의 데이터 개수는 feature의 개수에 exponential하게 필요한게 아니라 linear하게 필요하게 된다.**
- **간단한 가정으로 모델의 complexity를 크게 줄일 수 있는 것이다.**
- 때문에 Naïve Bayes 뿐 아니라 많은 모델에서 실제 데이터가 그런 분포를 보이지 않더라도 그 **데이터의 분포를 특정한 형태로 가정하여 문제를 간단하게 만드는 기술을 사용**한다.

Naïve Bayes란?

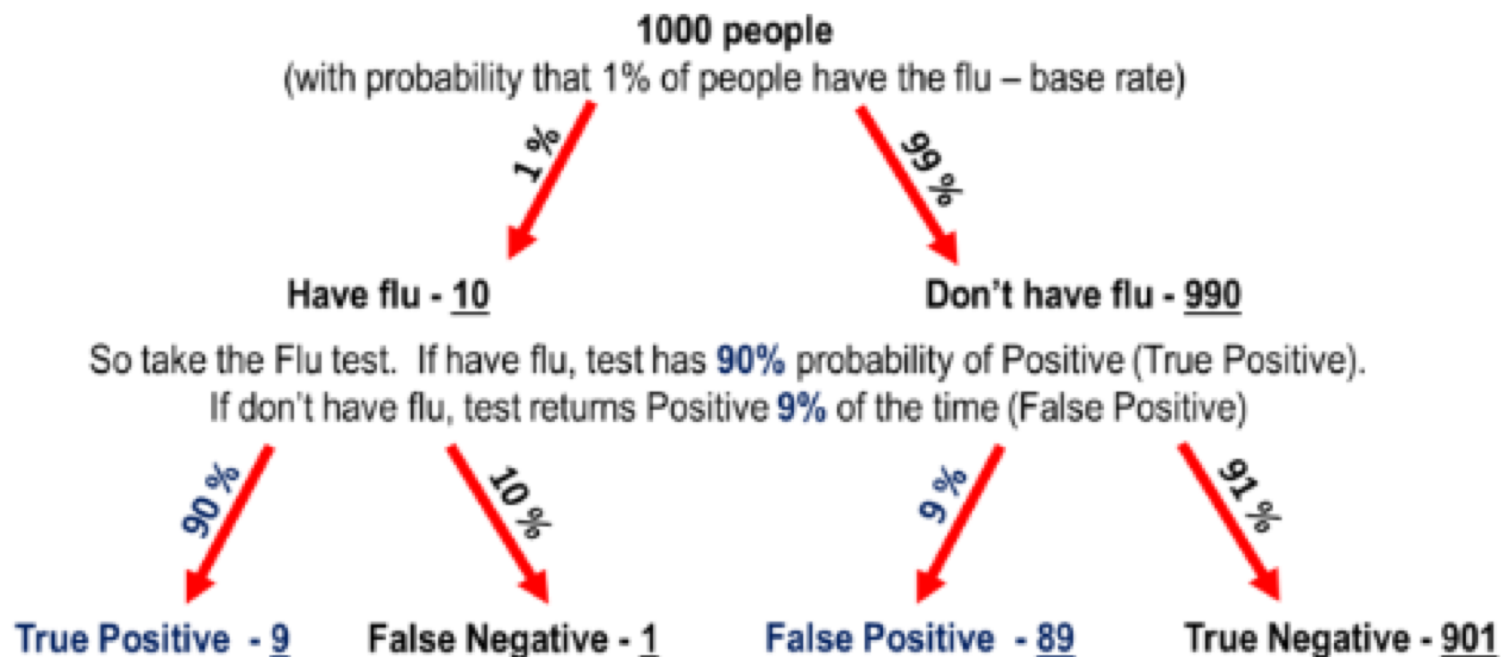
- 조금 더 엄밀하게 수식을 사용해 설명을 해보자.
- 우리가 가지고 있는 input data 를 $x=(x_1,...,x_n)$ 이라고 가정해보자.
- 즉 우리는 총 n개의 feature를 가지고 있다고 가정해보자 (보통 n은 데이터의 개수를 의미하지만, wikipedia의 notation을 따라가기 위하여 이 글에서도 dimension을 나타내기 위해 n을 사용하였다).
- 그리고 Class의 개수는 k라고 해보자.
- **우리의 목표는 $p(C|x_1,...,x_n)=p(C|x)$ 를 구하는 것이다.**
- 즉, 1부터 k까지의 class 중에서 가장 확률이 높은 class를 찾아내어 이를 사용해 classification을 하겠다는 것이다.
- Bayes rule을 알고 있으므로 이 식을 bayes rule을 사용해 전개하는 것은 간단하다.

- **$p(C|x)=p(C)p(x|C)p(x)$**

Naïve Bayes란?

- 이 때 분모에 있는 데이터의 확률은 normalize term이기 때문에 모든 값을 계산하고 나서 한 번에 계산하면 되므로 우리는 $p(x,C)=p(C)p(x|C)$, 다시 말해 **prior와 likelihood를 계산해야만 한다.**
- 그러나 이 값은 joint probability이므로 데이터에서부터 이 값을 알아내기 위해서는 '엄청나게 많은' 데이터가 필요하다. 구체적으로는 앞서 말한 것 처럼 dimension에 exponential하게 많은 데이터 개수를 필요로 한다.
- 그러나 만약 우리가 x 가 모두 indepent하다고 가정한다면 간단하게 다음과 같은 식으로 나타낼 수 있다.
- **$p(C)p(x_1,...,x_n|C)=p(C)p(x_1|C)p(x_2|C)...\dots=p(C)\prod p(x_i|C)$**
- 따라서 normalize term을 z 로 표현한다면, 우리가 구하고자 하는 최종 posterior는 **$p(C|x)=\frac{1}{z}p(C)\prod p(x_i|C)$** 로 나타낼 수 있게 된다.

Naïve Bayes 예시



The probability that you have the flu given you tested **Positive** =

$$\frac{\text{\# of people who have flu and tested True Positive (9)}}{\text{\# of people who have flu with True Positive (9) + \# of people who don't have flu with False Positive (89)}}$$

9 % probability you have flu