What is the basic assumption of the K Nearest Neighbors algorithm?

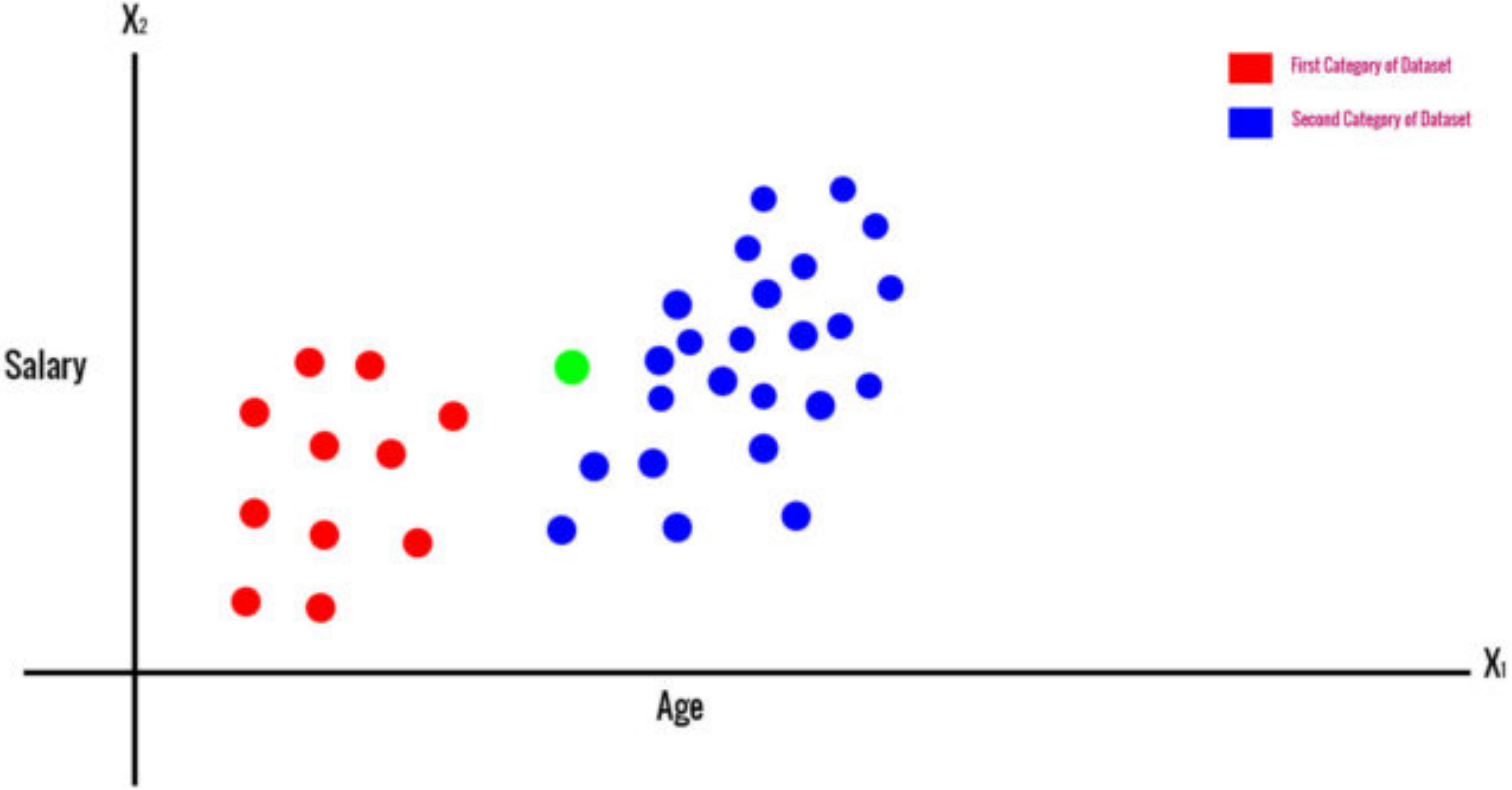
The majority of the labels of the observations ‘closest’ to our unlabeled instance are the likeliest label for our unlabeled data

How do we determine ‘closeness’?

- Distance - euclidean, manhattan, etc.
- Similarity

Naïve Bayes Classifier

Bayes' Theorem:



Bayes' Theorem:

Likelihood

Class Prior Probability

$$P(c \mid x) = \frac{P(x \mid c)\,P(c)}{P(x)}$$

Posterior Probability

Predictor Prior Probability

$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Bayes' Theorem: example

- Machine1: 30 laptops/hr → P(Machine1) = 30/50 = 0.6
- Machine2: 20 laptops/hr → P(Machine2) = 20/50 = 0.4

- **Out of all produced parts:** 1% are defective → P(Defect) = 1%
- **Out of all defective parts:**

  50% came from machine1 → P(Machine1|Defect) = 50%

  50% came from machine2 → P(Machine2|Defect) = 50%

- What is the profitability that a laptop produced by machine2 is defective? → **P(Defect|Machine2) = ?**

Bayes' Theorem: example

- Machine1: 30 laptops/hr ------------------→ P(Machine1) = 30/50 = 0.6
- Machine2: 20 laptops/hr ------------------→ P(Machine2) = 20/50 = 0.4

- **Out of all produced parts:** 1% are defective ----→ P(Defect) = 1%
- **Out of all defective parts:**

  50% came from machine1 ------------------→ P(Machine1|Defect) = 50%

  50% came from machine2 ------------------→ P(Machine2|Defect) = 50%

- What is the profitability that a laptop produced by machine2 is defective? --→ **P(Defect|Machine2) = ?**

Bayes' Theorem: example

- Machine1: 30 laptops/hr ------------------→ P(Machine1) = 30/50 = 0.6
- Machine2: 20 laptops/hr ------------------→ P(Machine2) = 20/50 = 0.4

- **Out of all produced parts:** 1% are defective ----→ P(Defect) = 1%
- **Out of all defective parts:**
  50% came from machine1 ------------------→ P(Machine1|Defect) = 50%
  50% came from machine2 ------------------→ P(Machine2|Defect) = 50%

- What is the profitability that a laptop produced by machine2 is defective? --→ **P(Defect|Machine2) = ?**

Answer:

$$P(Defect\,|\,Machine2) = \frac{P(Machine2\,|\,Defect) * P(Defect)}{P(Machine2)} = \frac{0.5 * .01}{.4} = 0.0125 = 1.25\,\%$$

Bayes’ Theorem: example

$$\textbf{P(Defect|Machine2)} = \frac{P(Machine2\,|\,Defect) * P(Defect)}{P(Machine2} = \textbf{1.25\%}$$

To make sense of the Bayes’ theorem, let’ take the following example:

- 1000 laptops
- 400 came from machine2
- 1% have a defect = 10
- 50% of the 10 came from machine2 = 5
- the question: % defective parts from machine2 = 5/400 = 1.25%

Quick Exercise:

- Using Bayes Theorem, P(Defect|Machine1) = ?

- Machine1: 600 laptops — P(Machine1) = 600/1000 = 0.6
- Machine2: 400 laptops — P(Machine2) = 400/1000 = 0.4

- **Out of all produced parts:** 10 are defective — P(Defect) = 10/1000 = 1%
- **Out of all defective parts:**

  5 came from machine1 — P(Machine1|Defect) = 5/10 = 50%

  5 came from machine2 — P(Machine2|Defect) = 5/10 = 50%

$$P(Defect\,|\,Machine1) = \frac{P(Machine1\,|\,Defect) * P(Defect)}{P(Machine1)} = \frac{0.5 * .01}{.6} = 0.0083 = .83\%$$

Naïve Bayes Classifier

Using Naive Bayes Classifier

Advantages	Disadvantages
Works well for many features	Needs large amount of training data
Fast to calculate	Continuous data must be preprocessed to be used
Handles rare events, categorical, and missing data well	Assumes features are independent and equally important

Naïve Bayes example1

New instance: (Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong)

Day	Outlook	Temp	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

- $p(yes) = \frac{9}{14} \quad and\ p(no) = \frac{5}{14}$

Outlook	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	3/5

Temp	Yes	No
Hot	2/9	3/5
Mild	4/9	2/5
Cool	3/9	2/5

Humidity	Yes	No
High	3/9	4/5
Normal	6/9	1/5

Wind	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

PlayTennis = No

$$p(yes|new\ Instance) = p(yes) * p(sunny|yes) * p(cool|yes) * p(high|yes) * p(strong|yes) = 0.0053$$

$$p(no|new\ Instance) = p(no) * p(sunny|no) * p(cool|no) * p(high|no) * p(strong|no) = 0.0206$$

$$v_{NB}(yes) = \frac{v_{NB}(yes)}{v_{NB}(yes)+v_{NB}(no)} = 0.205 \qquad v_{NB}(no) = \frac{v_{NB}(no)}{v_{NB}(yes)+v_{NB}(no)} = 0.795$$

New Instance: (Outlook = sunny, Temperature = cool, Humidity = High, Wind = strong)

PlayTennis = No

Day	Outlook	Temp	Humidity	Wind	PlayT
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$p(yes) = \frac{9}{14} \quad and \quad p(no) = \frac{5}{14}$$

Outlook	Yes	No
Sunny	2/9	3/5
Overcast	4/9	0/5
Rainy	3/9	2/5

Temp	Yes	No
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Yes	No
High	3/9	4/5
Normal	6/9	1/5

Wind	Yes	No
Strong	3/9	3/5
Weak	6/9	2/5

p(yes|new Instance) = p(yes) * p(sunny|yes) * p(cool|yes) * p(high|yes) * p(strong|yes) = $\frac{9}{14} * \frac{2}{9} * \frac{3}{9} * \frac{3}{9} * \frac{3}{9} = 0.0053$

p(no|new Instance) = p(no) * p(sunny|no) * p(cool|no) * p(high|no) * p(strong|no) = $\frac{5}{14} * \frac{3}{5} * \frac{1}{5} * \frac{4}{5} * \frac{3}{5} = 0.0206$

Normalized: $v_{NB}(yes) = \frac{v_{NB}(yes)}{v_{nb}(yes) + v_{nb}(no)} = \frac{0.0053}{0.0053 + 0.0206} = 0.205$ $\quad v_{NB}(no) = \frac{v_{NB}(no)}{v_{nb}(yes) + v_{nb}(no)} = \frac{0.0206}{0.0053 + 0.0206} = 0.795$

Gaussian Naïve Bayes: example2

New instance: <Height(ft) = 6, Weight(lbs) = 130), Foot size(inch) = 8> **Sex = ?**

Person	Height (ft)	Weight (lbs)	Foot size (inches)
Male	6.00	180	12
Male	5.92	190	11
Male	5.58	170	12
Male	5.92	165	10
Female	5.00	100	6
Female	5.50	150	8
Female	5.42	130	7
Female	5.75	150	9

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

Male:

$$Mean\ (Height) = \frac{(6+5.92+5.58+5.92)}{4} = 5.855$$

$$Variance\ (Height) = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$= \frac{(6-5.855)^2+(5.92-5.855)^2+(5.58-5.855)^2+(5.92-5.855)^2}{4-1}$$

$$= 0.035055$$

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

Gaussian Naïve Bayes: example2

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$$P(Male) = 4/8 = 0.5$$

$$P(Female) = 4/8 = 0.5$$

$$Posterior\ (Male) = \frac{P(M) * P(H|M) * P(W|M) * P(FS|M)}{Evidence}$$

$$Posterior\ (Female) = \frac{P(F) * P(H|F) * P(W|F) * P(FS|F)}{Evidence}$$

$$\boxed{f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}}$$

$$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2*0.035033}} = 1.5789$$

$$P(W|M) = 5.9881e^{-6}$$

$$P(FS|M) = 1.3112e^{-3}$$

$$P(H|F) = 2.2346e^{-1}$$

$$P(W|F) = 1.6789e^{-2}$$

$$P(FS|F) = 2.8669e^{-1}$$

Gaussian Naïve Bayes: example2

Sex	Mean (height)	Variance (height)	Mean (weight)	Variance (weight)	Mean(foot size)	Variance (foot size)
Male	5.855	0.035033	176.25	122.92	11.25	0.91667
Female	5.4175	0.097225	132.5	0558.33	7.5	1.6667

New Instance to be Classified is:

Sex	Height(ft)	Weight(lbs)	Foot size(inch)
Sample	6	130	8

$$P(Male) = 4/8 = 0.5$$
$$P(Female) = 4/8 = 0.5$$

$$P(H|M) = \frac{1}{\sqrt{2 * 3.142 * 0.035033}} * e^{-\frac{(6-5.855)^2}{2*0.035033}} = 1.5789$$
$$P(W|M) = 5.9881e^{-6}$$
$$P(FS|M) = 1.3112e^{-3}$$

$$P(H|F) = 2.2346e^{-1}$$
$$P(W|F) = 1.6789e^{-2}$$
$$P(FS|F) = 2.8669e^{-1}$$

Female

$$Posterior\ (Male) = \frac{P(M)*P(H|M)*P(W|M)*P(FS|M)}{Evidence} = 0.5 * 1.5789 * 5.9881e^{-6} * 1.3112e^{-3} = 6.1984e^{-9}$$

$$Posterior\ (Female) = \frac{P(F)*P(H|F)*P(W|F)*P(FS|F)}{Evidence} = 0.5 * 2.2346e^{-1} * 1.6789e^{-2} * 2.8669e^{-1} = 5.377e^{-4}$$

Naïve Bayes Classifier

- **Why it's called Naïve?** Because of the assumption that the independent variables are independent from one another since it based on the Bayes theorem. It's called naïve because it's a naïve assumption.
- Independent variables might not be the case. Still, it often gives a good result.
- P(X) - the predictor prior probability, why we can drop it from the Bayes theorem formula? Because it will have the same probability value when applied to either class. For classifier problems, what matters is the comparison between the two numerators.

Naïve Bayes Theorem Classwork

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Based on the dataset above, using the Naïve Bayes theorem classify the new instance: <Red,SUV,Domestic>

Naïve Bayes Theorem Classwork - Solution

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

P(Yes) = 5/10, P(No) = 5/10

Color	Yes	No
Red	3/5	2/5
Yellow	2/5	3/5

Type	Yes	No
Sports	4/5	2/5
SUV	1/5	3/5

Origin	Yes	No
Domestic	2/5	3/5
Imported	3/5	2/5

Based on the dataset above, using the Naïve Bayes theorem classify the new instance: <Red,SUV,Domestic>

$$P(Yes\,|\,Red, SUV, Domestic) = P(YES) * P(Red\,|\,Yes) * P(SUV\,|\,Yes) * P(Domestic\,|\,Yes)$$

$$P(Yes\,|\,Red, SUV, Domestic) = \frac{5}{10} * \frac{3}{5} * \frac{1}{5} * \frac{2}{5} = 0.0240$$

$$\nu_{NB}(Yes) = \frac{.024}{.024 + .072} = 0.25$$

$$P(No\,|\,Red, SUV, Domestic) = \frac{5}{10} * \frac{2}{5} * \frac{3}{5} * \frac{3}{5} = 0.0720$$

$$\nu_{NB}(No) = \frac{.072}{.024 + .072} = 0.75$$