
بسم الله الرحمن الرحیم

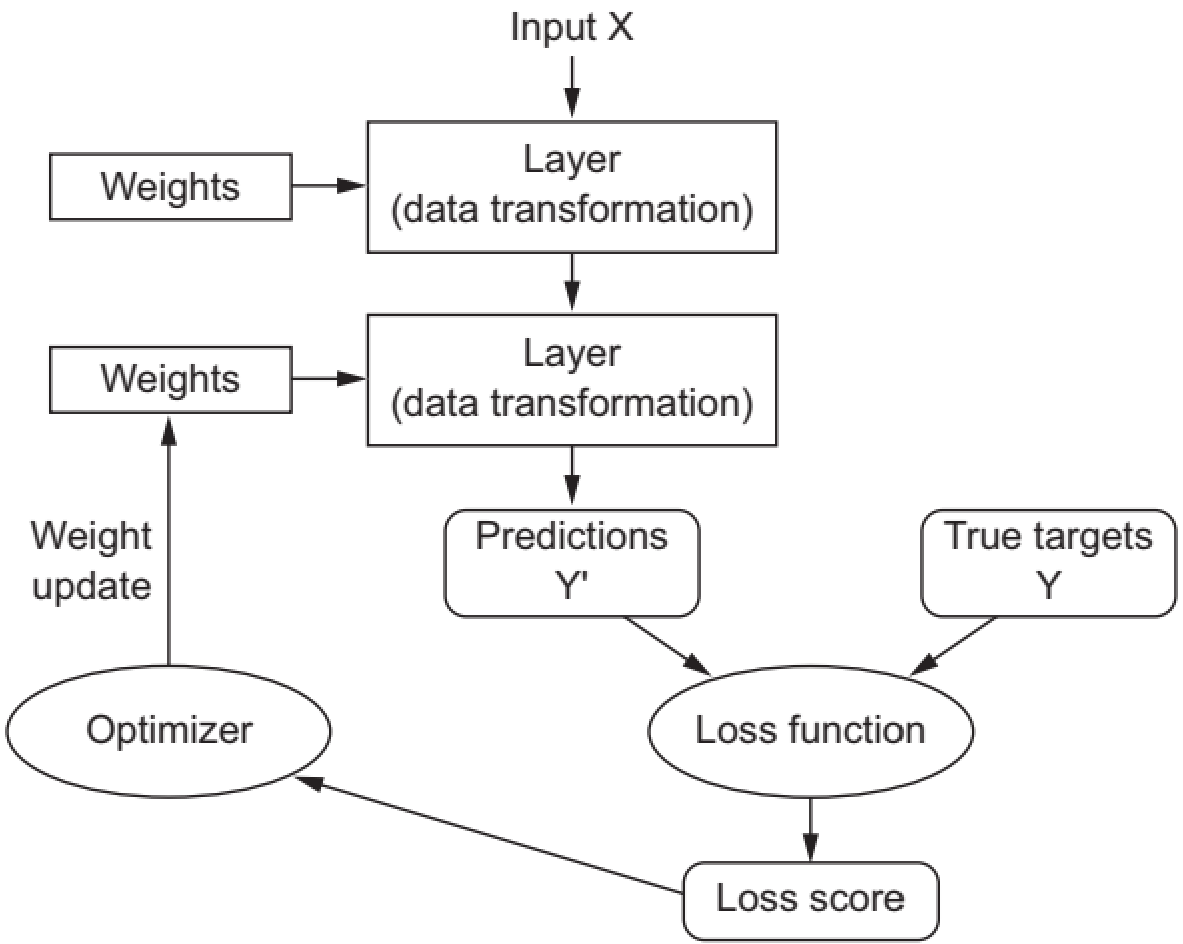
Deep Learning

Mohammad Reza Mohammadi

2021

https://quera.ir/overview/add_to_course/course/7883

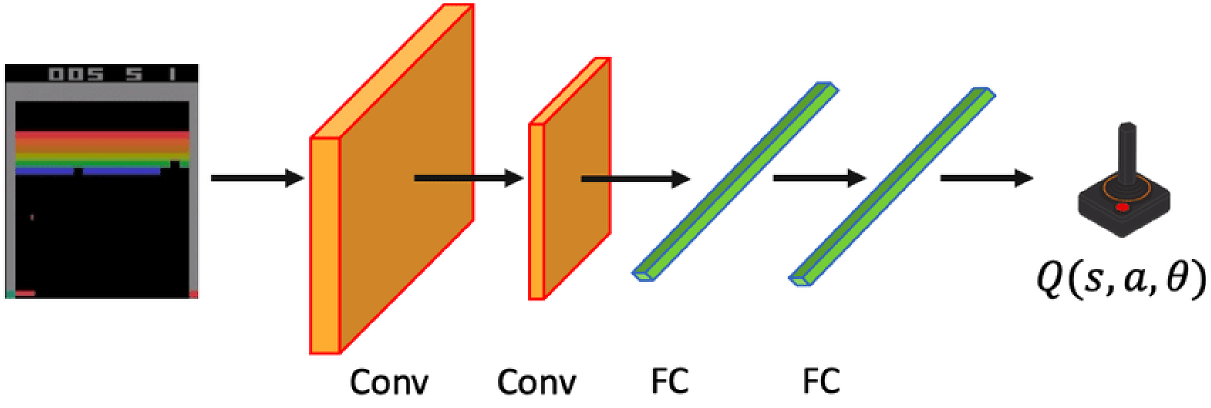
How deep learning works?

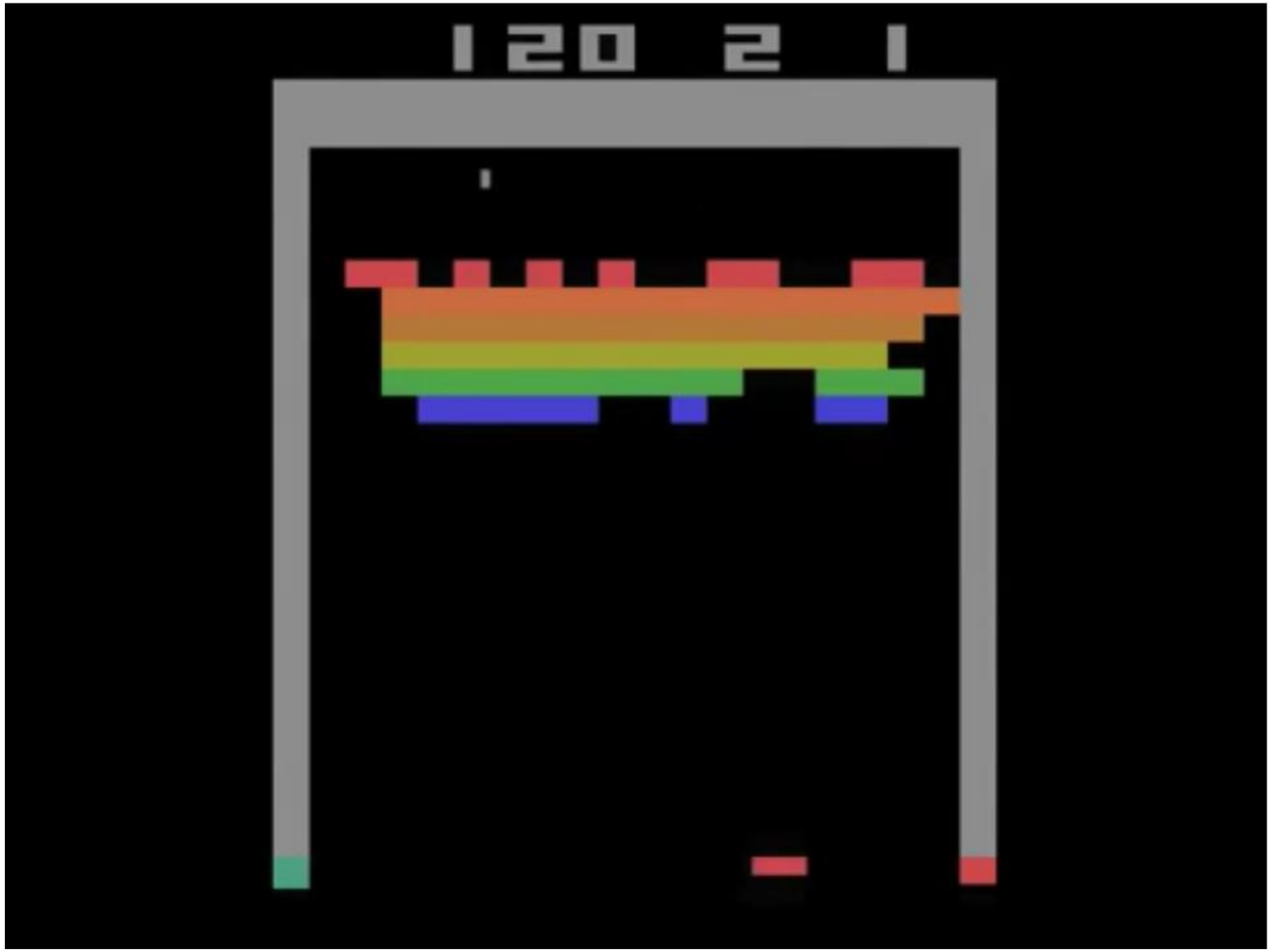


What deep learning has achieved so far?

- Deep learning rose to prominence in the early 2010s
- It has achieved remarkable results on perceptual problems such as seeing and hearing
  - Near-human-level image classification
  - Near-human-level speech recognition
  - Improved machine translation
  - Near-human-level autonomous driving
  - Ability to answer natural-language questions
  - Digital assistants such as Google Now, Amazon Alexa, and Bato

Deep Q-Network (DQN)



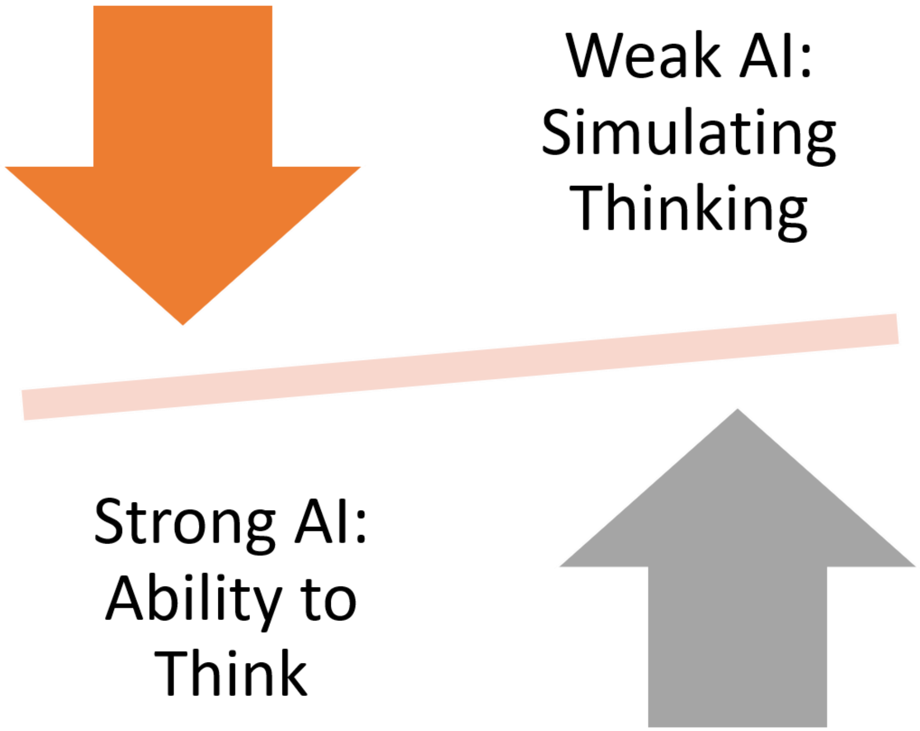


Deep learning hype

- Human-level general intelligence shouldn't be taken too seriously
  - The risk with high expectations for the short term is that, as technology fails to deliver, research investment will dry up, slowing progress for a long time
- This has happened before!
  - Marvin Minsky (1970): "In from three to eight years we will have a machine with the general intelligence of an average human being"
  - As of 2021, still far from possible!
- As these high expectations failed to materialize, researchers and government funds turned away from the field, marking the start of the first AI winter

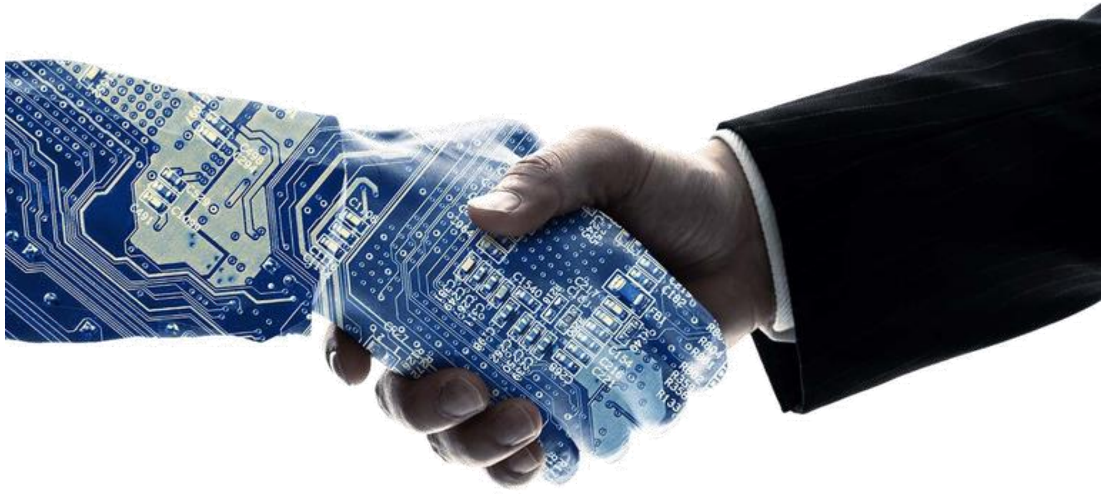
Weak AI vs. Strong AI

- Weak AI (Artificial Narrow Intelligence) :
  - is artificial intelligence that implements a limited part of mind and is focused on one narrow task
- Strong AI (Artificial General Intelligence)
  - is the hypothetical ability of an intelligent agent to understand or learn any intellectual task that a human being can



Promise of AI

- Don’t believe the short-term hype, but do believe in the long-term vision
  - AI is coming!
- Amazing progress in the past years
- But little of this progress has made its way into the products and processes that form our world
  - Your doctor doesn’t yet use AI



Before deep learning

- Deep learning isn't always the right tool for the job
  - Sometimes there isn't enough data for deep learning to be applicable, and sometimes the problem is better solved by a different algorithm

Probabilistic modeling

- We discuss probability theory as the framework for making decisions under uncertainty
- Data comes from a process that is not completely known
- This lack of knowledge is indicated by modeling the process as a random process
- Maybe the process is actually deterministic, but because we do not have access to complete knowledge about it, we model it as random and use probability theory to analyze it

Probabilistic modeling

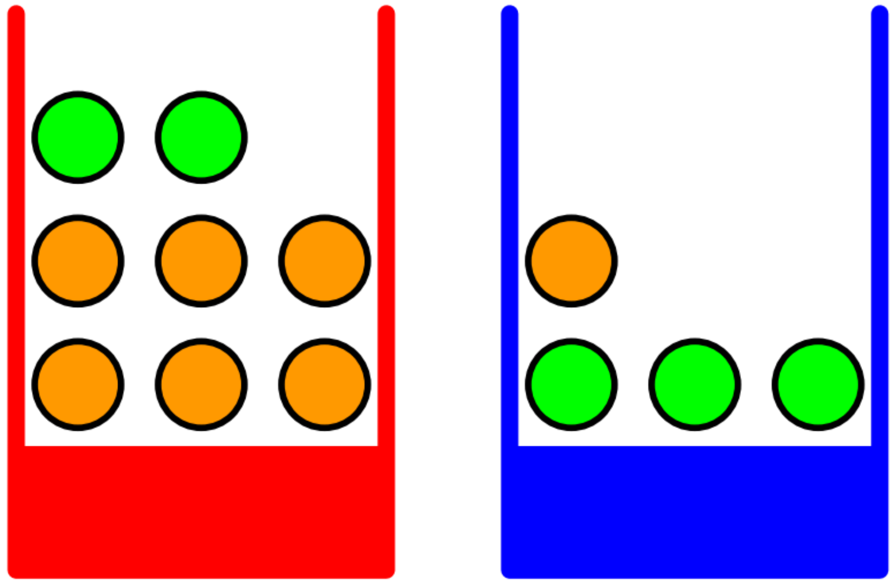
- The extra pieces of knowledge that we do not have access to are named the unobservable variables
- The outcome of tossing a coin is heads or tails, and we define a random variable that takes one of two values
- Let us say $X = 1$ denotes that the outcome of a toss is heads and $X = 0$ denotes tails

$$P(X = 1) = p_0$$

$$P(X = 0) = 1 - P(X = 1) = 1 - p_0$$

Probabilistic modeling

- Imagine we have two boxes, one red and one blue
- In the red box we have 2 green balls and 6 orange balls, and in the blue box we have 3 green balls and 1 orange ball
- Now suppose we randomly pick one of the boxes and from that box we randomly select one of the balls



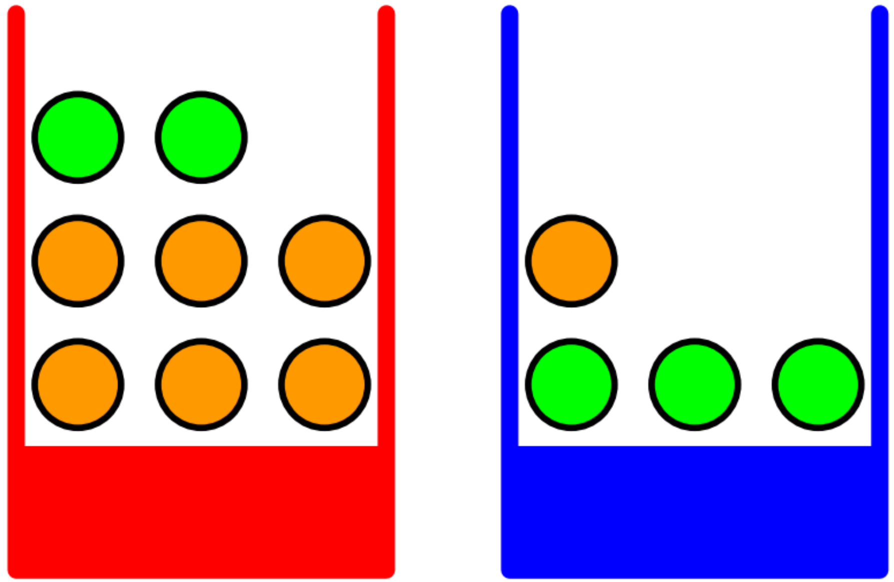
C. Bishop, "*Pattern Recognition and Machine Learning*", 2006

Probabilistic modeling

- In this example, the identity of the box that will be chosen is a random variable
- Similarly, the identity of the ball is also a random variable

$$Box = \{red, blue\}$$

$$Ball = \{green, orange\}$$



Probabilistic modeling

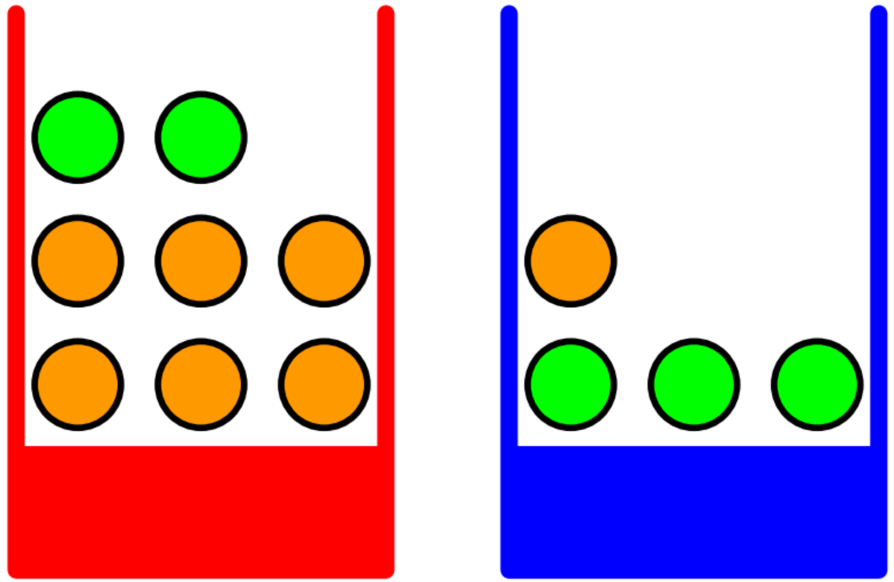
- Let us suppose that we pick the red box 40% of the time and we pick the blue box 60% of the time
- what is the probability that the selection procedure will pick a green ball?
- given that we have chosen an orange ball, what is the probability that the box we chose was the blue one?

$$Box = \{red, blue\}$$

$$Ball = \{green, orange\}$$

$$P(Box = red) = 0.4$$

$$P(Box = blue) = 0.6$$

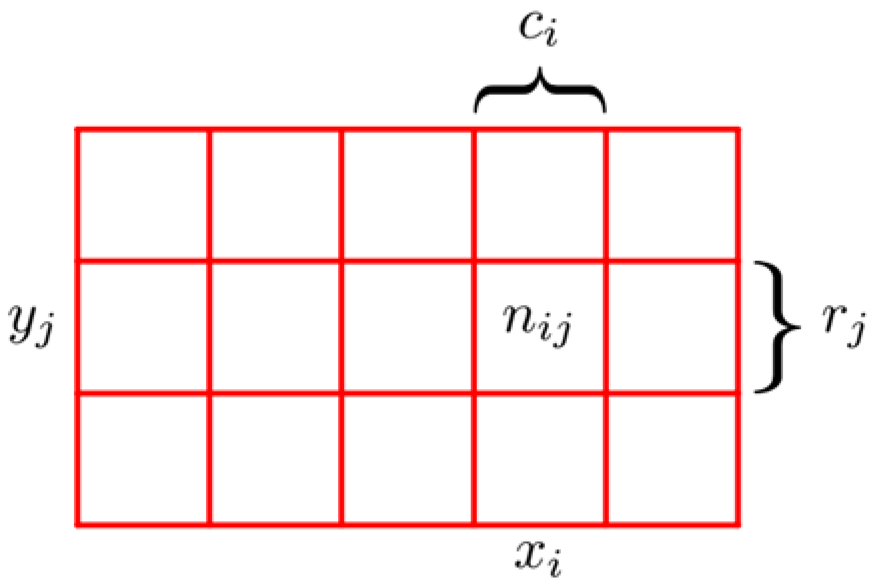


Probabilistic modeling

- In order to derive the rules of probability, consider the slightly more general example involving two random variables X and Y
- Consider a total of N trials
- The probability that X will take the value x_i and Y will take the value y_j is written $P(X = x_i, Y = y_j)$ and is called the joint probability

$$X = \{x_i \text{ for } i = 1, 2, \ldots, M\}$$

$$Y = \{y_j \text{ for } j = 1, 2, \ldots, L\}$$



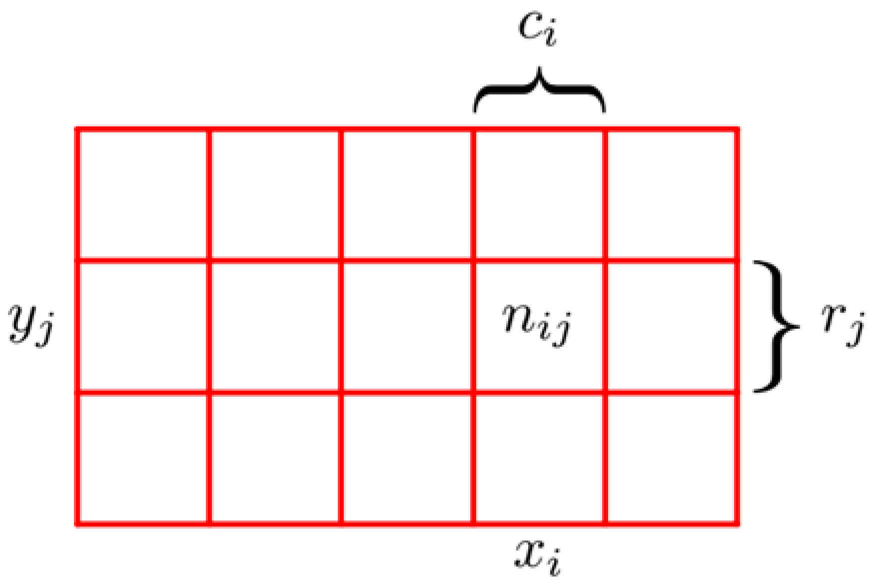
$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

Probabilistic modeling

- The probability that X takes the value x_i irrespective of the value of Y (*marginal* probability)

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N} = \sum_{j=1}^{L} P(X = x_i, Y = y_j)$$

- the sum rule of probability



Probabilistic modeling

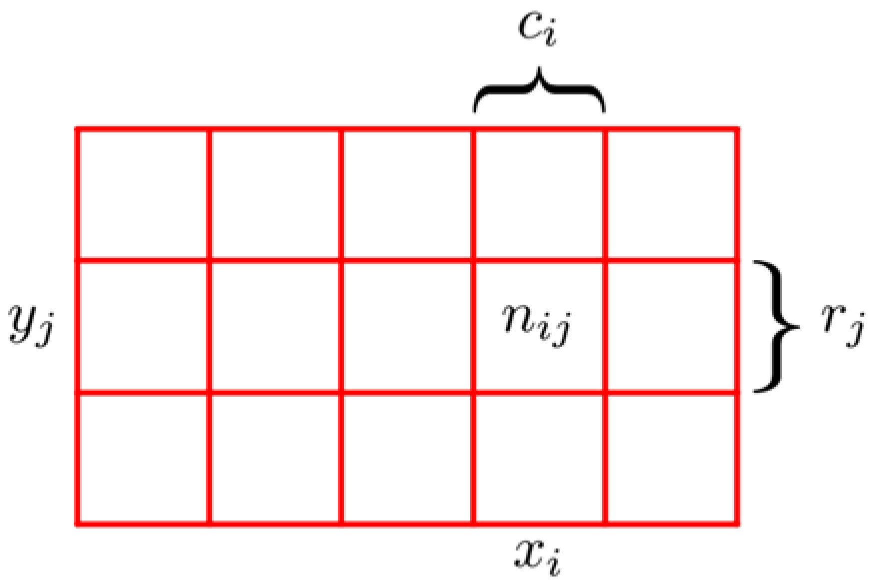
- The conditional probability of $Y = y_j$ given $X = x_i$

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$

$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i}\frac{c_i}{N}$$

$$= P(Y = y_j | X = x_i) P(X = x_i)$$

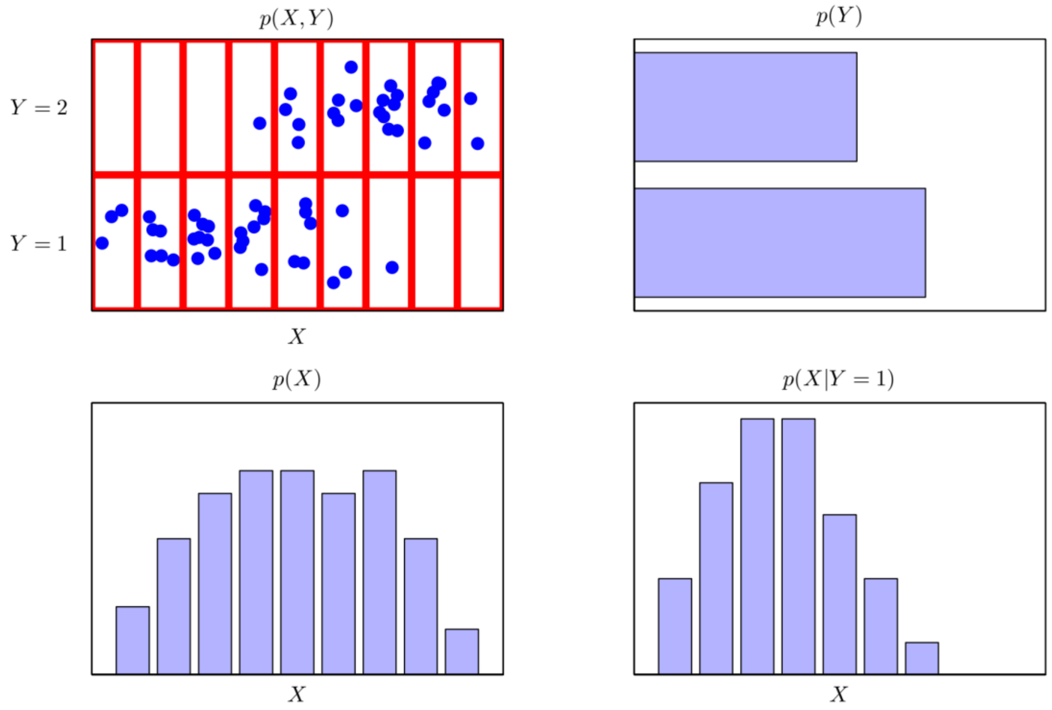
- the product rule of probability



Probabilistic modeling

- Sum rule: $P(X) = \sum_Y P(X,Y)$
- Product rule: $P(X,Y) = P(Y|X)P(X) = P(Y,X) = P(X|Y)P(Y)$
- Bayes' theorem: $P(Y|X) = \dfrac{P(X|Y)P(Y)}{P(X)}$

$$P(X) = \sum_Y P(X|Y)P(Y)$$



Probabilistic modeling

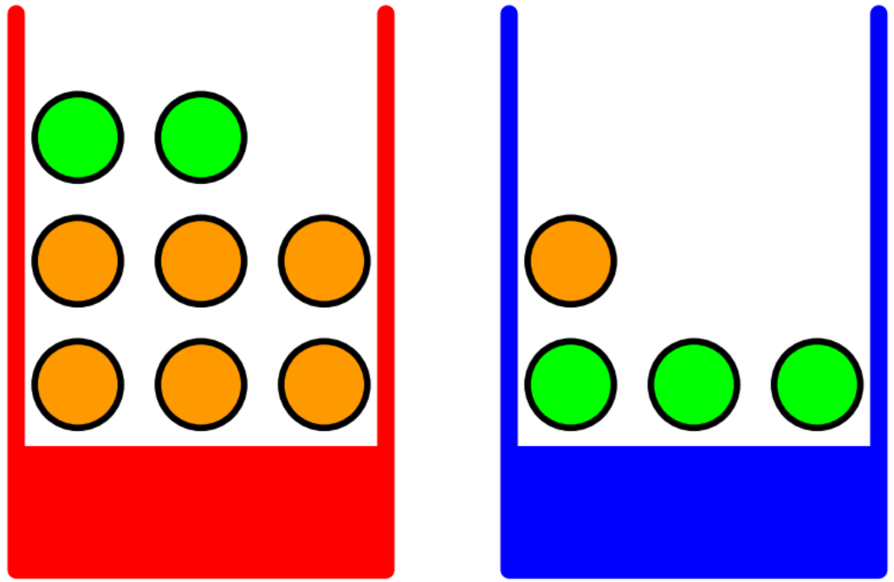
- Let us suppose that we pick the red box 40% of the time and we pick the blue box 60% of the time
- what is the probability that the selection procedure will pick a green ball?
- given that we have chosen an orange ball, what is the probability that the box we chose was the blue one?

$$Box = \{red, blue\}$$

$$Ball = \{green, orange\}$$

$$P(Box = red) = 0.4$$

$$P(Box = blue) = 0.6$$



Probabilistic modeling

- what is the probability that the selection procedure will pick a green ball?

$$P(Ball = green) = P(green|blue)P(blue) + P(green|red)P(red) = \frac{6}{10}\frac{3}{4} + \frac{4}{10}\frac{1}{4} = \frac{11}{20}$$

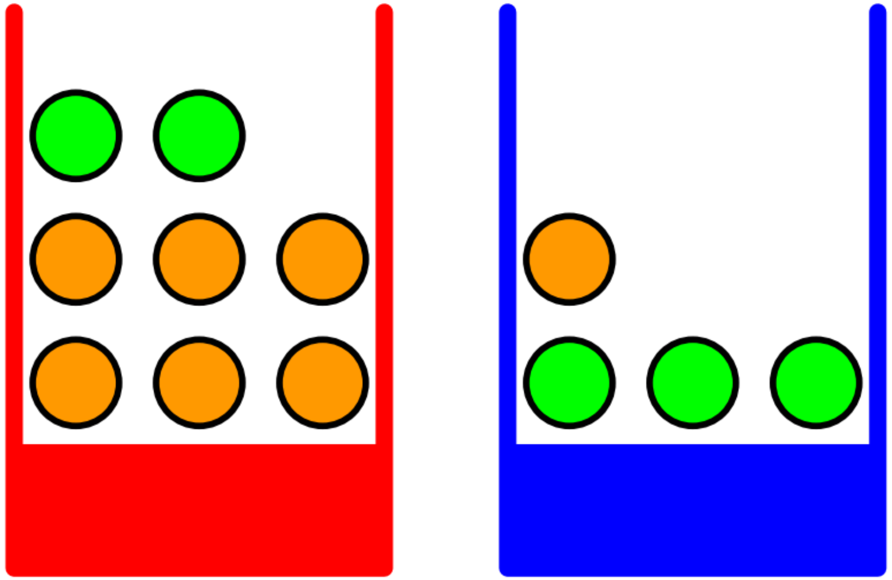
$$P(Ball = orange) = \frac{9}{20}$$

$$P(Ball = green|Box = red) = \frac{2}{8}$$

$$P(Ball = green|Box = blue) = \frac{3}{4}$$

$$P(Box = red) = 0.4$$

$$P(Box = blue) = 0.6$$



Probabilistic modeling

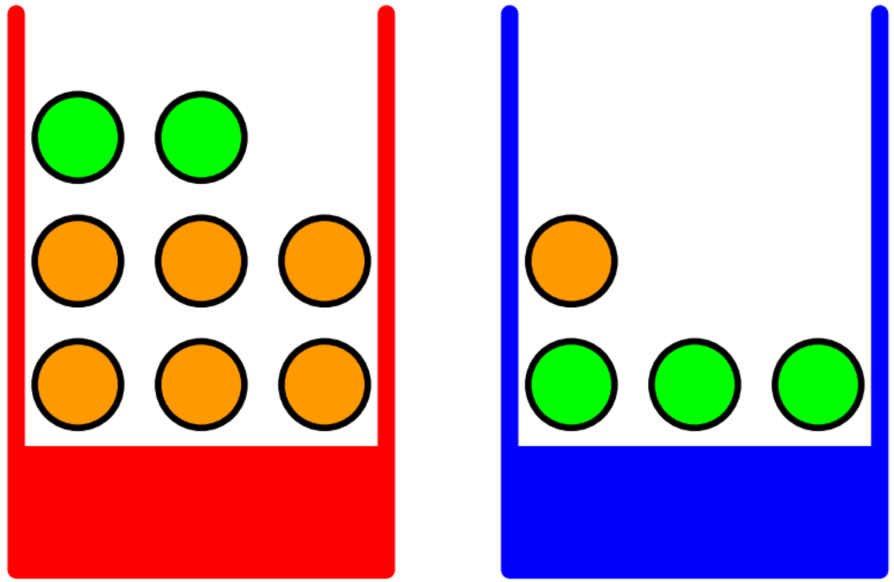
- given that we have chosen an orange ball, what is the probability that the box we chose was the blue one?

$$P(Box = blue|Ball = orange) = \frac{P(orange|blue)P(blue)}{P(orange)} = \frac{\frac{1}{4}\frac{6}{10}}{\frac{6}{8}\frac{4}{10} + \frac{1}{4}\frac{6}{10}} = \frac{1}{3}$$

$$P(Box = red|Ball = orange) = \frac{2}{3}$$

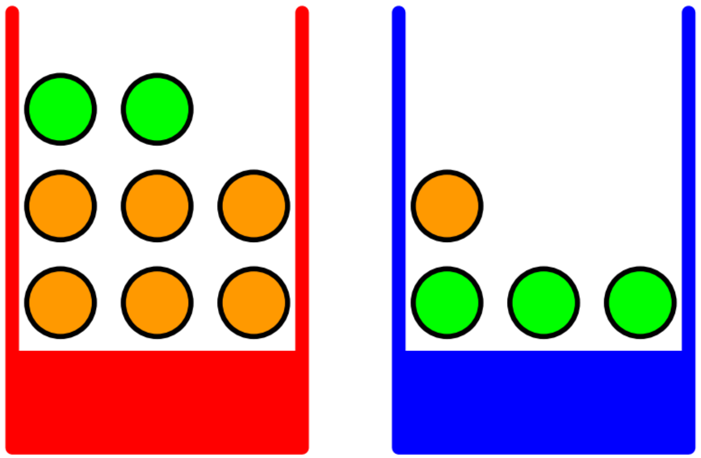
$$P(Box = red) = 0.4$$

$$P(Box = blue) = 0.6$$



Probabilistic modeling

- We can provide an important interpretation of Bayes' theorem as follows:
  - If we had been asked which box had been chosen before being told the identity of the selected ball, then the most complete information we have available is provided by the probability $P(Box)$, which we call this the prior probability
  - Once we are told that the ball is an orange, we can then use Bayes' theorem to compute the probability $P(Box|Ball)$, which we shall call the posterior probability



$P(red|orange) = 2/3$

$P(blue|orange) = 1/3$

$P(Box = red) = 0.4$

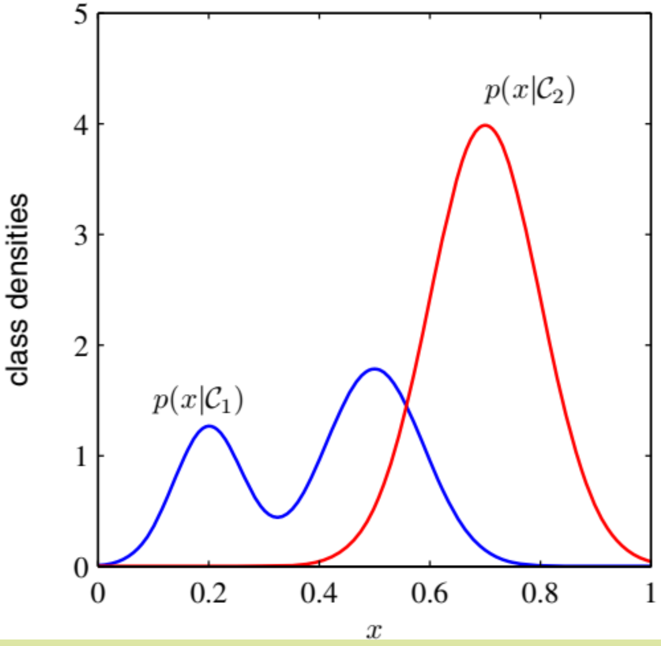
$P(Box = blue) = 0.6$

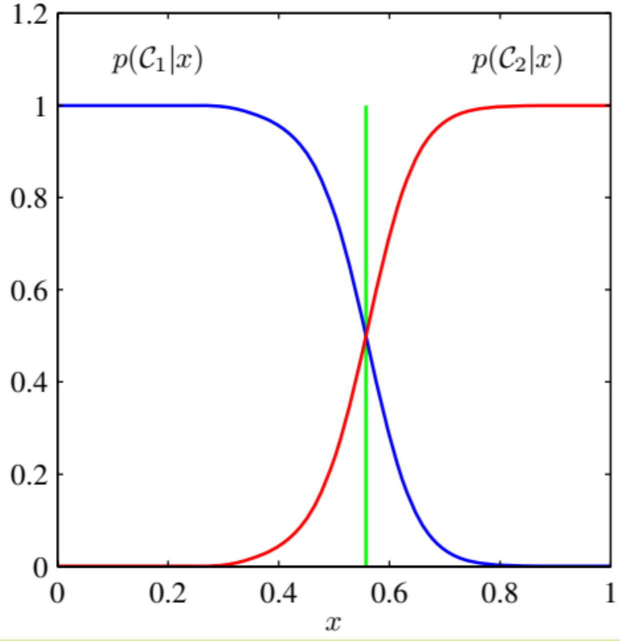
Bayes' classifier

- Bayes' classifier chooses the class with the highest posterior probability

$$P(C_i|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|C_i)P(C_i)}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|C_i)P(C_i)}{\sum_{k=1}^{K} P(\boldsymbol{x}|C_k)P(C_k)}$$

$$\text{choose } \ C_i \ \text{ if } \ P(C_i|\boldsymbol{x}) = \max_k P(C_k|\boldsymbol{x})$$

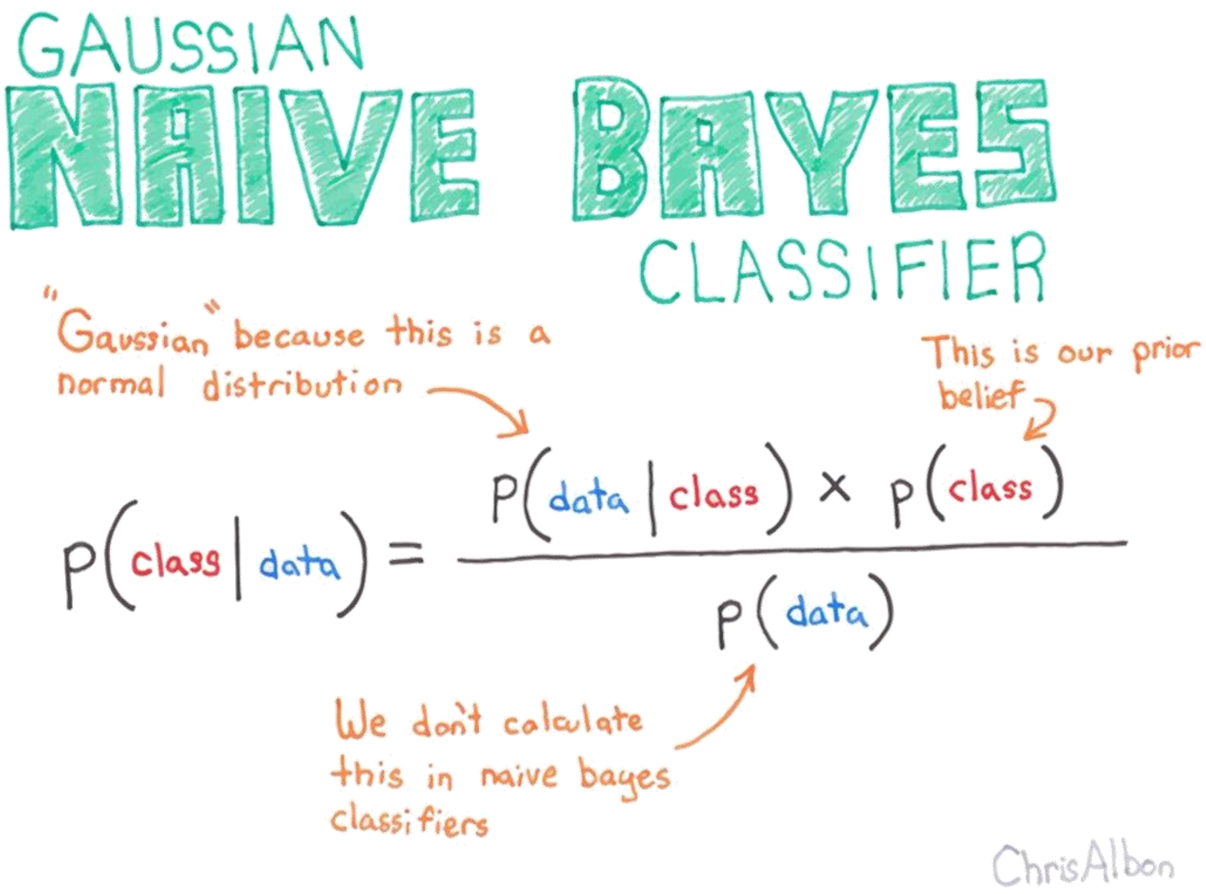




Naïve Bayes

- A type of machine-learning classifier based on applying Bayes' theorem while assuming that the features in the input data are all independent

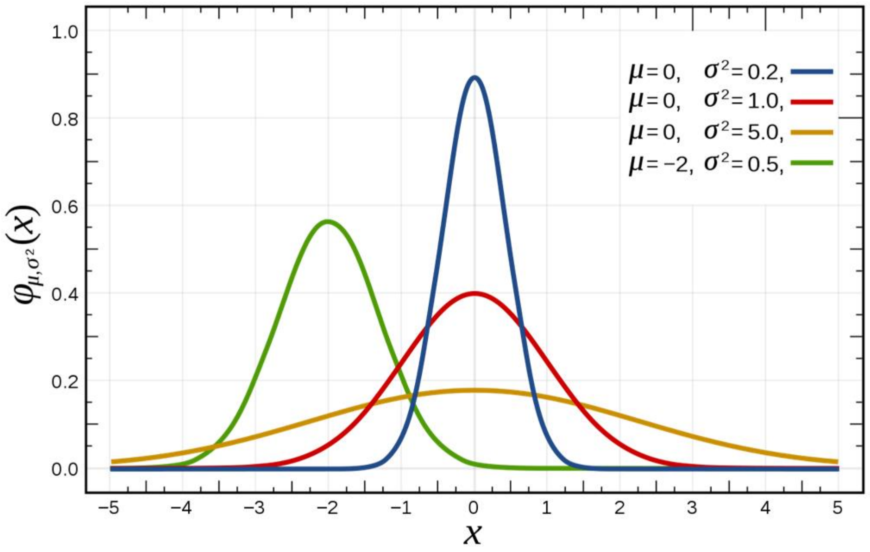
$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$



Naïve Bayes: Example

- Problem: classify whether a given person is a male or a female
- The features include height, weight, and foot size.

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9



Naïve Bayes: Example

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

Person	height (feet)	weight (lbs)	foot size(inches)
male	6	180	12
male	5.92 (5'11")	190	11
male	5.58 (5'7")	170	12
male	5.92 (5'11")	165	10
female	5	100	6
female	5.5 (5'6")	150	8
female	5.42 (5'5")	130	7
female	5.75 (5'9")	150	9

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033×10^{-2}	176.25	1.2292×10^{2}	11.25	9.1667×10^{-1}
female	5.4175	9.7225×10^{-2}	132.5	5.5833×10^{2}	7.5	1.6667

Naïve Bayes: Example

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033×10^{-2}	176.25	1.2292×10^{2}	11.25	9.1667×10^{-1}
female	5.4175	9.7225×10^{-2}	132.5	5.5833×10^{2}	7.5	1.6667

Naïve Bayes: Example

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

$$P(\text{male}) = 0.5$$

$$p(\text{height} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789,$$

$$p(\text{weight} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(130-\mu)^2}{2\sigma^2}\right) = 5.9881 \cdot 10^{-6}$$

$$p(\text{foot size} \mid \text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(8-\mu)^2}{2\sigma^2}\right) = 1.3112 \cdot 10^{-3}$$

Naïve Bayes: Example

Person	height (feet)	weight (lbs)	foot size(inches)
sample	6	130	8

$$\text{posterior (male)} = \frac{P(\text{male})\, p(\text{height} \mid \text{male})\, p(\text{weight} \mid \text{male})\, p(\text{foot size} \mid \text{male})}{evidence}$$

$$\text{posterior (female)} = \frac{P(\text{female})\, p(\text{height} \mid \text{female})\, p(\text{weight} \mid \text{female})\, p(\text{foot size} \mid \text{female})}{evidence}$$

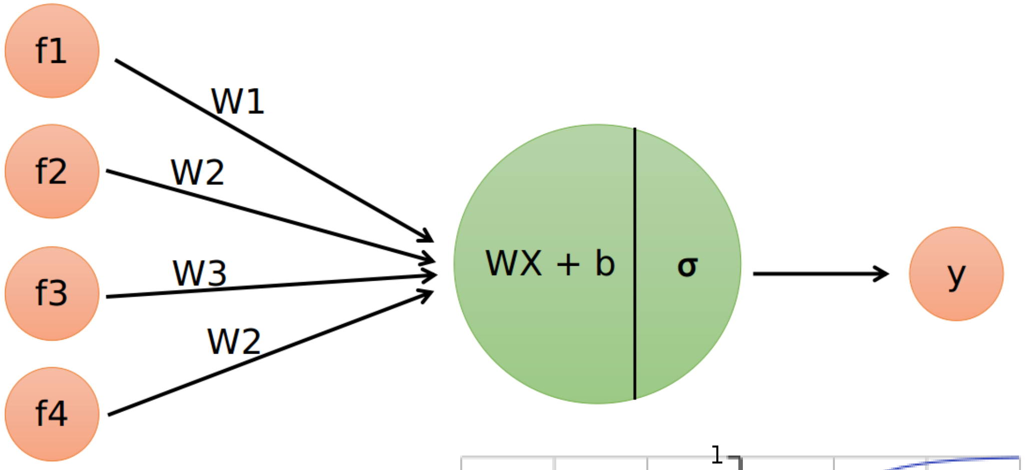
$$\text{posterior numerator (male)} = \text{their product} = 6.1984 \cdot 10^{-9}$$

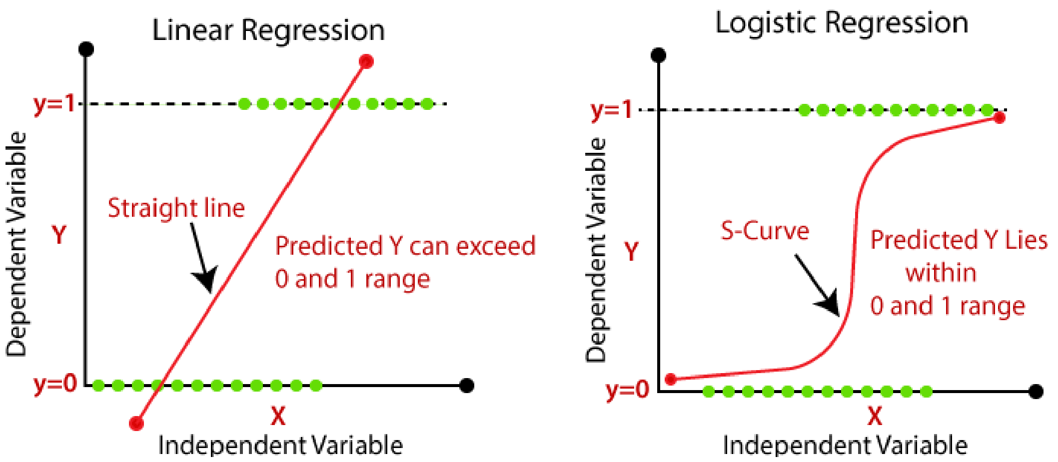
$$\text{posterior numerator (female)} = \text{their product} = 5.3778 \cdot 10^{-4}$$

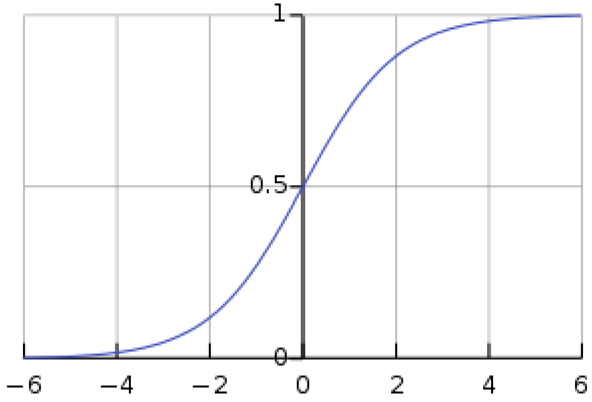
Person	mean (height)	variance (height)	mean (weight)	variance (weight)	mean (foot size)	variance (foot size)
male	5.855	3.5033×10^{-2}	176.25	1.2292×10^{2}	11.25	9.1667×10^{-1}
female	5.4175	9.7225×10^{-2}	132.5	5.5833×10^{2}	7.5	1.6667

Logistic Regression

- Sometimes considered to be the “hello world” of modern machine learning



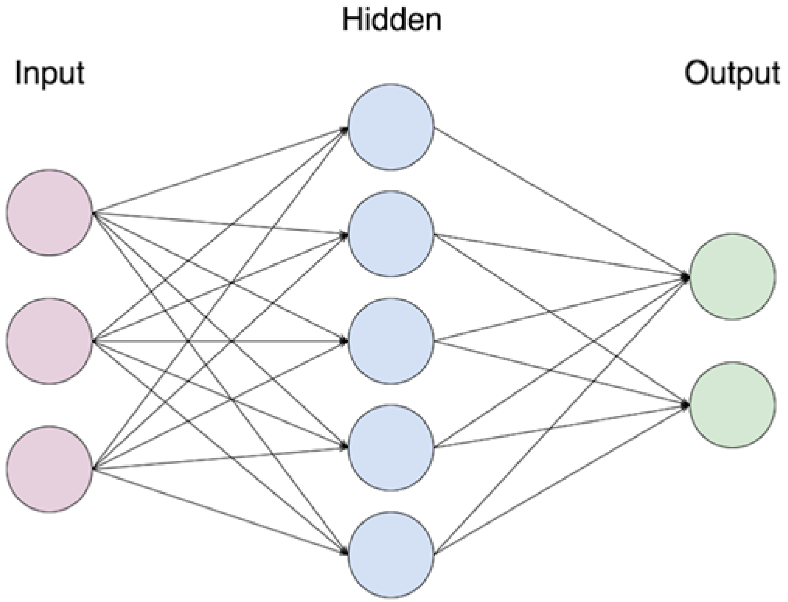




$$S(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{e^x+1} = 1 - S(-x)$$

Early Neural Networks

- Although the core ideas of neural networks were investigated in toy forms as early as the 1950s, the approach took decades to get started
- For a long time, the missing piece was an efficient way to train large neural networks
- This changed in the mid-1980s, when multiple people independently rediscovered the Backpropagation algorithm



Return of Neural Networks

- Around 2010, although neural networks were almost completely shunned by the scientific community at large, a number of people still working on neural networks started to make important breakthroughs
- ImageNet: a very difficult problem

Return of Neural Networks

- ImageNet: a very difficult problem
  - classifying high resolution color images into 1,000 different categories after training on 1.4 million images
  - 2011: classical approaches, 25.8%
  - 2012: deep learning, 15.3% (huge breakthorough)
    - Since then, dominated by CNNs