Deep Learning

Lecture 2: Probability Theory

Dr. Mehrdad Maleki

Random Variable

A **random variable** is a variable whose values are depend on the outcome of a random event.

Random Variable

A **random variable** is a variable whose values are depend on the outcome of a random event. Let $\mathbf{X}$ be the outcomes of tossing a coin, then if Head appears $\mathbf{X} = 1$ and if Tails appears $\mathbf{X} = 0$.

Random Variable

A **random variable** is a variable whose values are depend on the outcome of a random event. Let $\mathbf{X}$ be the outcomes of tossing a coin, then if Head appears $\mathbf{X} = 1$ and if Tails appears $\mathbf{X} = 0$. $\mathbf{X}$ could be the outcome of tossing a dice, in this case $\mathbf{X} \in \{1, 2, 3, 4, 5, 6\}$. $\mathbf{X}$ could be the height of people comming to a bank.

Random Variable

A **random variable** is a variable whose values are depend on the outcome of a random event. Let $\mathbf{X}$ be the outcomes of tossing a coin, then if Head appears $\mathbf{X} = 1$ and if Tails appears $\mathbf{X} = 0$. $\mathbf{X}$ could be the outcome of tossing a dice, in this case $\mathbf{X} \in \{1, 2, 3, 4, 5, 6\}$. $\mathbf{X}$ could be the height of people comming to a bank. The set of all possible outputs of a random variable is called the sample space of that random variable and usually we denote it by S.

Random Variable

A **random variable** is a variable whose values are depend on the outcome of a random event. Let $\mathbf{X}$ be the outcomes of tossing a coin, then if Head appears $\mathbf{X} = 1$ and if Tails appears $\mathbf{X} = 0$. $\mathbf{X}$ could be the outcome of tossing a dice, in this case $\mathbf{X} \in \{1, 2, 3, 4, 5, 6\}$. $\mathbf{X}$ could be the height of people comming to a bank. The set of all possible outputs of a random variable is called the sample space of that random variable and usually we denote it by S. In fact any subset of the sample space is the random event.

Probability

In tossing a coin we can ask question like "How likely is that the value of $\mathbf{X}$ is equal 1?".

Probability

In tossing a coin we can ask question like "How likely is that the value of $\mathbf{X}$ is equal 1?". This is the probability of the events that for them $\mathbf{X} = 1$.

Probability

In tossing a coin we can ask question like "How likely is that the value of $\mathbf{X}$ is equal 1?". This is the probability of the events that for them $\mathbf{X} = 1$. This is a real number between 0 and 1 and we denote this by $P[\mathbf{X} = 1]$

If S be a finite set then probability of event $A \subseteq S$ is define as follow,

$$P(A) = P(\mathbf{X} \in A) = \frac{|A|}{|S|}$$

$$0 \leq P(A) \leq 1$$

$$0 \leq P(A) \leq 1$$
$$P(\emptyset) = 0$$

$$0 \leq P(A) \leq 1$$

$$P(\emptyset) = 0$$

$$P(S) = 1$$

$$0 \leq P(A) \leq 1$$

$$P(\emptyset) = 0$$

$$P(S) = 1$$

$$\textbf{if } A \cap B = \emptyset \Rightarrow P(A \cup B) = P(A) + P(B)$$

Discerete Distribution

A discrete random variable has a countable number of possible values.

Discerete Distribution

A discrete random variable has a countable number of possible values.
Probability function: describe the probability $P(\mathbf{X} \in A)$ that A is a random event from the sample space.

Discerete Distribution

A discrete random variable has a countable number of possible values.
Probability function: describe the probability $P(\mathbf{X} \in A)$ that A is a random event from the sample space.
Probability mass function (pmf): $f(i) = P[\mathbf{X} = i]$ for descrete random variables.

Continious Distribution

In general, quantities such as pressure, height, mass, weight, density, volume, temperature, and distance are examples of continuous random variables.

Continious Distribution

In general, quantities such as pressure, height, mass, weight, density, volume, temperature, and distance are examples of continuous random variables. Between any two values of a continuous random variable, there are an infinite number of other valid values.

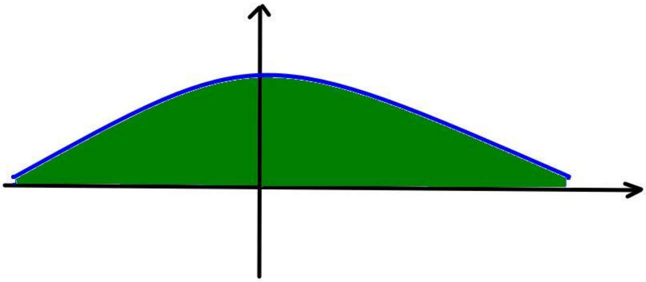
Probability density function (pdf): Is a continuous positive function $f_{\mathbf{X}}(x) : \mathbb{R} \rightarrow \mathbb{R}^{\geq 0}$ such that,

Probability density function (pdf): Is a continuous positive function $f_{\mathbf{X}}(x) : \mathbb{R} \rightarrow \mathbb{R}^{\geq 0}$ such that,

$$\int_{-\infty}^{\infty} f_{\mathbf{X}}(x) dx = 1$$

Probability density function (pdf): Is a continuous positive function $f_{\mathbf{X}}(x) : \mathbb{R} \to \mathbb{R}^{\geq 0}$ such that,

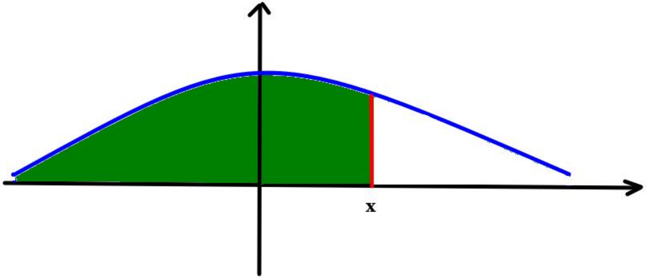
$$\int_{-\infty}^{\infty} f_{\mathbf{X}}(x)dx = 1$$



Cumulative distribution function: is the function that evaluating the probability of $\mathbf{X} \leq x$,

Cumulative distribution function: is the function that evaluating the probability of $\mathbf{X} \leq x$,i.e., $F_{\mathbf{X}}(x) = P[\mathbf{X} \leq x]$

Cumulative distribution function: is the function that evaluating the probability of $\mathbf{X} \leq x$,i.e., $F_{\mathbf{X}}(x) = P[\mathbf{X} \leq x]$



Bernoulli Trial

A **Bernoulli trial** is a random experiment with exactly two possible outcomes, "success" and "failure".

Bernoulli Trial

A **Bernoulli trial** is a random experiment with exactly two possible outcomes, "success" and "failure".

$$p = P[\mathbf{X} = \text{"}success\text{"}]$$

Bernoulli Trial

A **Bernoulli trial** is a random experiment with exactly two possible outcomes, "success" and "failure".

$$p = P[\mathbf{X} = \text{"}\mathit{success}\text{"}]$$
$$q = P[\mathbf{X} = \text{"}\mathit{failure}]$$

Bernoulli Trial

A **Bernoulli trial** is a random experiment with exactly two possible outcomes, "success" and "failure".

$$p = P[\mathbf{X} = "success"]$$
$$q = P[\mathbf{X} = "failure]$$
$$p + q = 1$$

Probability mass function (PMF) of Bernouli distribution,

Probability mass function (PMF) of Bernouli distribution,

$$f(x = 0|\mathbf{p}) = q$$

Probability mass function (PMF) of Bernouli distribution,

$$f(x=0|\mathbf{p}) = q$$
$$f(x=1|\mathbf{p}) = p$$

Probability mass function (PMF) of Bernouli distribution,

$$f(x = 0|\mathbf{p}) = q$$
$$f(x = 1|\mathbf{p}) = p$$

where $\mathbf{p} = (p, q)$. So,

Probability mass function (PMF) of Bernouli distribution,

$$f(x = 0|\mathbf{p}) = q$$
$$f(x = 1|\mathbf{p}) = p$$

where $\mathbf{p} = (p, q)$. So,

$$f(x|\mathbf{p}) = p^x q^{1-x}$$

Probability mass function (PMF) of Bernouli distribution,

$$f(x = 0|\mathbf{p}) = q$$
$$f(x = 1|\mathbf{p}) = p$$

where $\mathbf{p} = (p, q)$. So,

$$\begin{aligned} f(x|\mathbf{p}) &= p^x q^{1-x} \\ &= p^x (1-p)^{1-x} \end{aligned}$$

Probability mass function (PMF) of Bernouli distribution,

$$
\begin{aligned}
f(x=0|\mathbf{p}) &= q \\
f(x=1|\mathbf{p}) &= p
\end{aligned}
$$

where $\mathbf{p} = (p, q)$. So,

$$
\begin{aligned}
f(x|\mathbf{p}) &= p^x q^{1-x} \\
&= p^x (1-p)^{1-x}
\end{aligned}
$$

for $x \in \{0, 1\}$.

Categorical (generalized Bernoulli) Distribution

Categorical distribution is a discrete probability distribution that describes the possible results of a random variable that can take on one of K possible categories, with the probability of each category separately specified. So $\mathbf{X} \in \{1, 2, \ldots, K\}$.

Probability mass function (PMF) of Categorical distribution,

Probability mass function (PMF) of Categorical distribution,

$$f(x = i|\mathbf{p}) = p_i$$

Probability mass function (PMF) of Categorical distribution,

$$f(x = i|\mathbf{p}) = p_i$$

where $\mathbf{p} = (p_1, \ldots, p_K)$. So,

Probability mass function (PMF) of Categorical distribution,

$$f(x = i|\mathbf{p}) = p_i$$

where $\mathbf{p} = (p_1, \ldots, p_K)$. So,

$$f(x|\mathbf{p}) = p_1^{\mathbf{I}_{[x=1]}} \ldots p_K^{\mathbf{I}_{[x=K]}}$$

Probability mass function (PMF) of Categorical distribution,

$$f(x = i|\mathbf{p}) = p_i$$

where $\mathbf{p} = (p_1, \ldots, p_K)$. So,

$$\begin{aligned} f(x|\mathbf{p}) &= p_1^{\mathbf{I}_{[x=1]}} \ldots p_K^{\mathbf{I}_{[x=K]}} \\ &= \prod_{i=1}^{K} p_i^{\mathbf{I}_{[x=i]}} \end{aligned}$$

Probability mass function (PMF) of Categorical distribution,

$$f(x = i|\mathbf{p}) = p_i$$

where $\mathbf{p} = (p_1, \ldots, p_K)$. So,

$$\begin{aligned} f(x|\mathbf{p}) &= p_1^{\mathbf{I}_{[x=1]}} \ldots p_K^{\mathbf{I}_{[x=K]}} \\ &= \prod_{i=1}^{K} p_i^{\mathbf{I}_{[x=i]}} \end{aligned}$$

where

$$\mathbf{I}_{[x=i]} = \begin{cases} 1 & \text{if } x = i \\ 0 & \text{else} \end{cases}$$

for $x \in \{0, 1\}$.

Conditional Probability

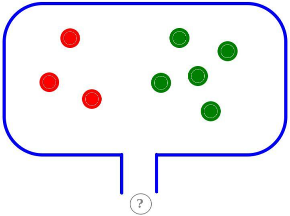
Conditional probability is a measure of the probability of an event occurring, given that another event (by assumption, presumption, assertion or evidence) has already occurred.

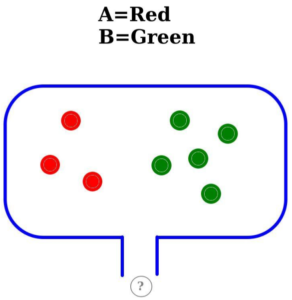
Conditional Probability

Conditional probability is a measure of the probability of an event occurring, given that another event (by assumption, presumption, assertion or evidence) has already occurred.The conditional probability of A given B is defined as follow,

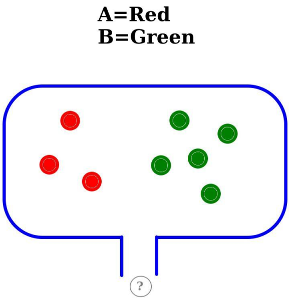
$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

A=Red
B=Green

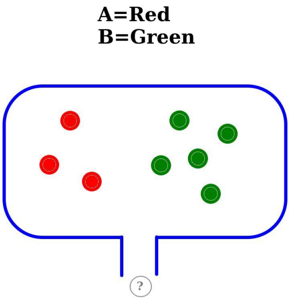




We choose one ball randomly. What is the probability that this ball is red?

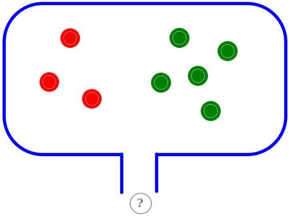


We choose one ball randomly. What is the probability that this ball is red?
$P(A) = \frac{3}{8}$,



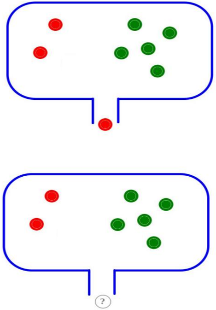
We choose one ball randomly. What is the probability that this ball is red?
$P(A) = \frac{3}{8}$, What about the probability that this ball is green?

A=Red
B=Green

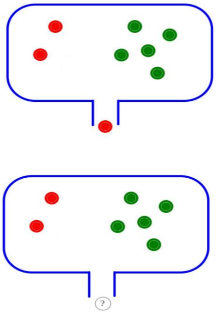


We choose one ball randomly. What is the probability that this ball is red? $P(A) = \frac{3}{8}$, What about the probability that this ball is green? $P(B) = \frac{5}{8}$.

A=second ball green
B=first ball red

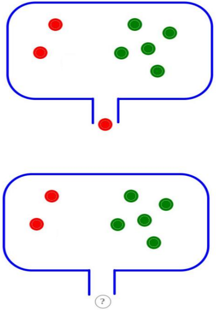


A=second ball green
B=first ball red



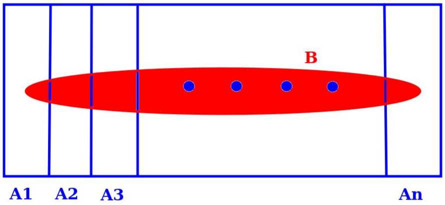
We choose one ball randomly and through it away. We choose another ball randomly. What is the probability that the second ball is green such that the fist ball is red?

A=second ball green
B=first ball red

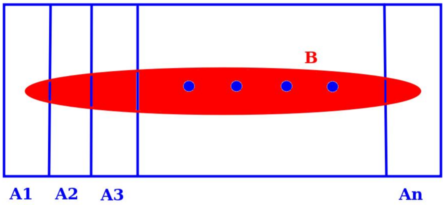


We choose one ball randomly and through it away. We choose another ball randomly. What is the probability that the second ball is green such that the fist ball is red? $P(A|B) = \frac{5}{7}$

Law of Total Probability

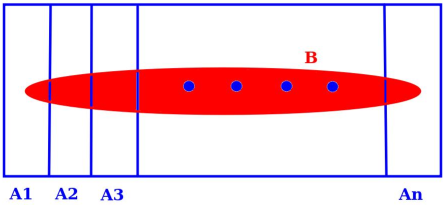


Law of Total Probability

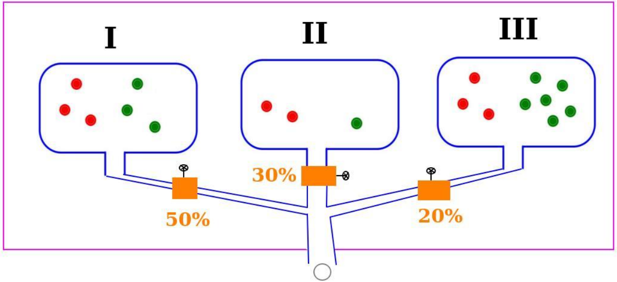


$$P(B) = P(A_1 \cap B) + \cdots + P(A_n \cap B)$$

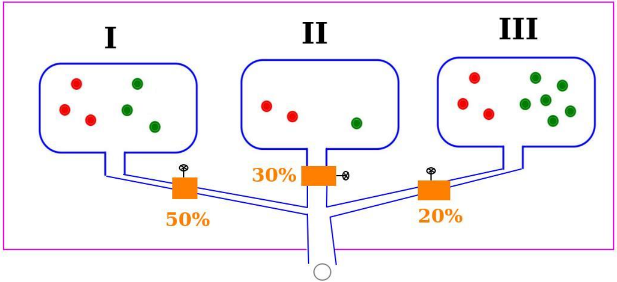
Law of Total Probability



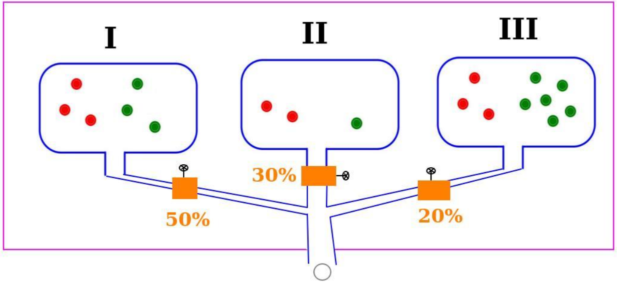
$$
\begin{aligned}
P(B) &= P(A_1 \cap B) + \cdots + P(A_n \cap B) \\
&= P(A_1)P(B|A_1) + \cdots + P(A_n)P(B|A_n)
\end{aligned}
$$



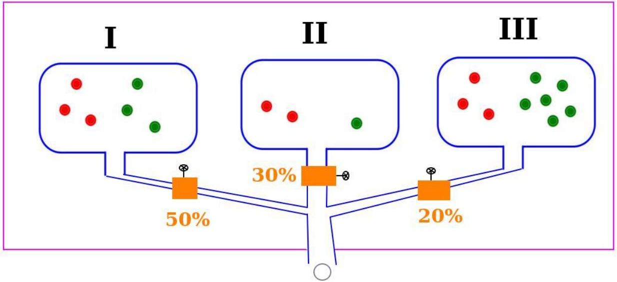
$$P(R) =$$



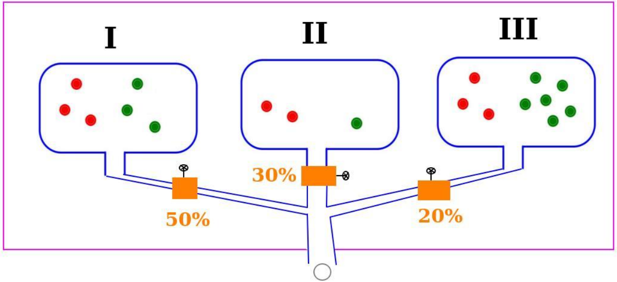
$$P(R) = P(I)P(R|I) + P(II)P(R|II) + P(III)P(R|III)$$



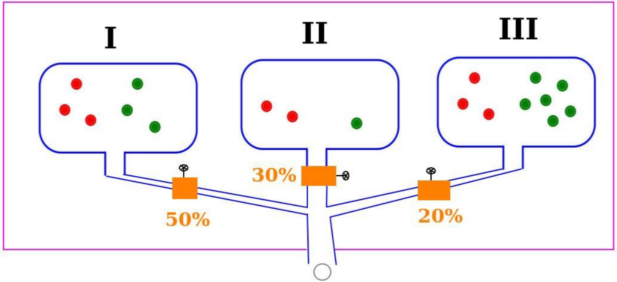
$$P(R) = P(I)P(R|I) + P(II)P(R|II) + P(III)P(R|III)$$
$$= 0.5\frac{3}{6} + 0.3\frac{2}{3} + 0.2\frac{3}{9}$$



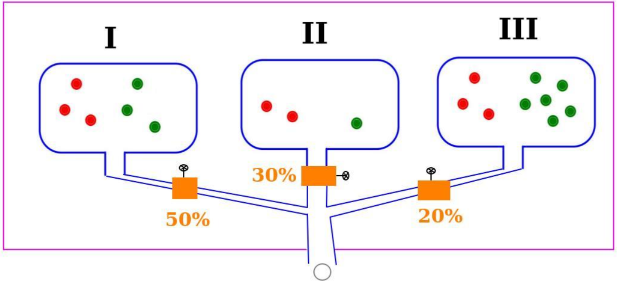
$$
\begin{aligned}
P(R) &= P(I)P(R|I) + P(II)P(R|II) + P(III)P(R|III) \\
&= 0.5\frac{3}{6} + 0.3\frac{2}{3} + 0.2\frac{3}{9} \\
&\approx 0.52
\end{aligned}
$$



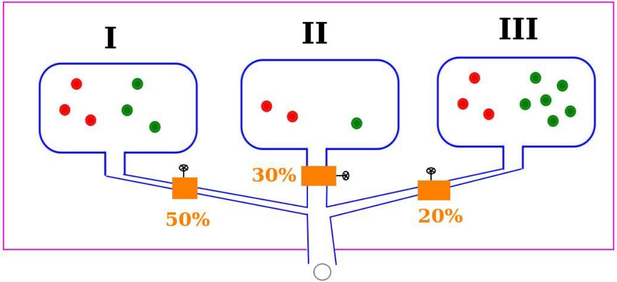
$P(G) =$



$$P(G) = P(I)P(G|I) + P(II)P(G|II) + P(III)P(G|III)$$



$$P(G) = P(I)P(G|I) + P(II)P(G|II) + P(III)P(G|III)$$
$$= 0.5\frac{3}{6} + 0.3\frac{1}{3} + 0.2\frac{6}{9}$$



$$
\begin{aligned}
P(G) &= P(I)P(G|I) + P(II)P(G|II) + P(III)P(G|III) \\
&= 0.5\frac{3}{6} + 0.3\frac{1}{3} + 0.2\frac{6}{9} \\
&\approx 0.48
\end{aligned}
$$

Motivation- Thinking Fast and Slow
(Daniel Kahneman)

Motivation- Thinking Fast and Slow

(Daniel Kahneman)

A cab was involved in a hit-and-run accident at night.

Motivation- Thinking Fast and Slow

(Daniel Kahneman)

A cab was involved in a hit-and-run accident at night.Two cab companies, the Green and the Blue, operate in the city.

Motivation- Thinking Fast and Slow

(Daniel Kahneman)

A cab was involved in a hit-and-run accident at night.Two cab companies, the Green and the Blue, operate in the city. You are given the following data:

Motivation- Thinking Fast and Slow

(Daniel Kahneman)

A cab was involved in a hit-and-run accident at night.Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 85% of the cabs in the city are Green and 15% are Blue.

Motivation- Thinking Fast and Slow

(Daniel Kahneman)

A cab was involved in a hit-and-run accident at night.Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue.

Motivation- Thinking Fast and Slow

(Daniel Kahneman)

A cab was involved in a hit-and-run accident at night.Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident

Motivation- Thinking Fast and Slow

(Daniel Kahneman)

A cab was involved in a hit-and-run accident at night. Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time.

Motivation- Thinking Fast and Slow

(Daniel Kahneman)

A cab was involved in a hit-and-run accident at night.Two cab companies, the Green and the Blue, operate in the city. You are given the following data: 85% of the cabs in the city are Green and 15% are Blue. A witness identified the cab as Blue. The court tested the reliability of the witness under the circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colors 80% of the time and failed 20% of the time. What is the probability that the cab involved in the accident was Blue rather than Green?

Bayes Rule

Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence,

Bayes Rule

Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Rule

Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A) =$ **the prior**, is the initial degree of belief in A.

Bayes Rule

Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

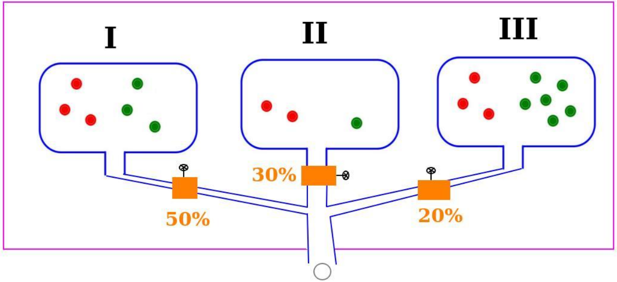
- $P(A) =$ **the prior**, is the initial degree of belief in A.
- $P(A|B) =$ **the posterior**, is the degree of belief after incorporating news that B is true.

Bayes Rule

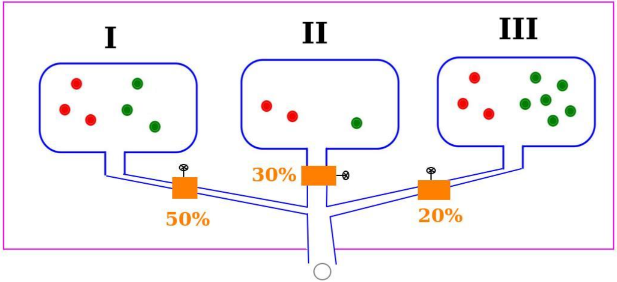
Bayes' theorem links the degree of belief in a proposition before and after accounting for evidence,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A) =$ **the prior**, is the initial degree of belief in A.
- $P(A|B) =$ **the posterior**, is the degree of belief after incorporating news that B is true.
- $P(B|A) =$ **the likelihood**, that can be estimated from the training data.



$$P(I|R) =$$



$$P(I|R) = \frac{P(R|I)P(I)}{P(R)}$$
$$= \frac{\frac{3}{6}0.5}{0.52}$$
$$\approx 0.48$$

$$P(II|R) =$$

$$
\begin{aligned}
P(II|R) &= \frac{P(R|II)P(II)}{P(R)} \\
&= \frac{\frac{2}{3}0.3}{0.52} \\
&\approx 0.38
\end{aligned}
$$

$$
P(III|R) =
$$

$$
\begin{aligned}
P(II|R) &= \frac{P(R|II)P(II)}{P(R)} \\
&= \frac{\frac{2}{3}0.3}{0.52} \\
&\approx 0.38
\end{aligned}
$$

$$
\begin{aligned}
P(III|R) &= \frac{P(R|III)P(III)}{P(R)} \\
&= \frac{\frac{3}{9}0.2}{0.52} \\
&\approx 0.12
\end{aligned}
$$

$$P(I|G) =$$

$$
\begin{aligned}
P(I|G) &= \frac{P(G|I)P(I)}{P(G)} \\
&= \frac{\frac{3}{6}0.5}{0.48} \\
&\approx 0.52
\end{aligned}
$$

$$
P(II|G) =
$$

$$
\begin{aligned}
P(I|G) &= \frac{P(G|I)P(I)}{P(G)} \\
&= \frac{\frac{3}{6}0.5}{0.48} \\
&\approx 0.52
\end{aligned}
$$

$$
\begin{aligned}
P(II|G) &= \frac{P(G|II)P(II)}{P(G)} \\
&= \frac{\frac{1}{3}0.3}{0.48} \\
&\approx 0.2
\end{aligned}
$$

$$
P(III|G) =
$$

$$
\begin{aligned}
P(I|G) &= \frac{P(G|I)P(I)}{P(G)} \\
&= \frac{\frac{3}{6}0.5}{0.48} \\
&\approx 0.52 \\
P(II|G) &= \frac{P(G|II)P(II)}{P(G)} \\
&= \frac{\frac{1}{3}0.3}{0.48} \\
&\approx 0.2 \\
P(III|G) &= \frac{P(G|III)P(III)}{P(G)} \\
&= \frac{\frac{6}{9}0.2}{0.48} \\
&\approx 0.27
\end{aligned}
$$

Chain rule for conditional probability

$$P(A_n \cap \cdots \cap A_1) =$$

Chain rule for conditional probability

$$P(A_n \cap \cdots \cap A_1) = P(A_n|A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1}|A_{n-2} \cap \cdots \cap A_1) \cdot \cdots \cdot P(A_1)$$

Chain rule for conditional probability

$$\begin{aligned}P(A_n \cap \cdots \cap A_1) &= P(A_n | A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) \cdot \cdots \cdot P(A_1) \\ &= \prod_{k=1}^{n} P(A_k | A_{k-1} \cap \cdots \cap A_1)\end{aligned}$$

Chain rule for conditional probability

$$
\begin{aligned}
P(A_n \cap \cdots \cap A_1) &= P(A_n | A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) \cdot \cdots \cdot P(A_1) \\
&= \prod_{k=1}^{n} P(A_k | A_{k-1} \cap \cdots \cap A_1)
\end{aligned}
$$

$$
P(\mathbf{X}_n, \ldots, \mathbf{X}_1) =
$$

Chain rule for conditional probability

$$
\begin{aligned}
P(A_n \cap \cdots \cap A_1) &= P(A_n | A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) \cdot \cdots \cdot P(A_1) \\
&= \prod_{k=1}^{n} P(A_k | A_{k-1} \cap \cdots \cap A_1)
\end{aligned}
$$

$$
P(\mathbf{X}_n, \ldots, \mathbf{X}_1) = P(\mathbf{X}_n | \mathbf{X}_{n-1}, \ldots, \mathbf{X}_1) P(\mathbf{X}_{n-1} | \mathbf{X}_{n-2}, \ldots, \mathbf{X}_1) \ldots P(\mathbf{X}_1)
$$

Chain rule for conditional probability

$$P(A_n \cap \cdots \cap A_1) = P(A_n | A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) \cdot \cdots \cdot P(A_1)$$
$$= \prod_{k=1}^{n} P(A_k | A_{k-1} \cap \cdots \cap A_1)$$

$$P(\mathbf{X}_n, \ldots, \mathbf{X}_1) = P(\mathbf{X}_n | \mathbf{X}_{n-1}, \ldots, \mathbf{X}_1) P(\mathbf{X}_{n-1} | \mathbf{X}_{n-2}, \ldots, \mathbf{X}_1) \ldots P(\mathbf{X}_1)$$
$$= \prod_{k=1}^{n} P(\mathbf{X}_k | \mathbf{X}_{k-1}, \ldots, \mathbf{X}_1)$$

Chain rule for conditional probability

$$P(A_n \cap \cdots \cap A_1) = P(A_n | A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) \cdot \cdots \cdot P(A_1)$$
$$= \prod_{k=1}^{n} P(A_k | A_{k-1} \cap \cdots \cap A_1)$$

$$P(\mathbf{X}_n, \ldots, \mathbf{X}_1) = P(\mathbf{X}_n | \mathbf{X}_{n-1}, \ldots, \mathbf{X}_1) P(\mathbf{X}_{n-1} | \mathbf{X}_{n-2}, \ldots, \mathbf{X}_1) \ldots P(\mathbf{X}_1)$$
$$= \prod_{k=1}^{n} P(\mathbf{X}_k | \mathbf{X}_{k-1}, \ldots, \mathbf{X}_1)$$

$$P(I\ like\ you) =$$

Chain rule for conditional probability

$$
\begin{aligned}
P(A_n \cap \cdots \cap A_1) &= P(A_n | A_{n-1} \cap \cdots \cap A_1) \cdot P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) \cdot \cdots \cdot P(A_1) \\
&= \prod_{k=1}^{n} P(A_k | A_{k-1} \cap \cdots \cap A_1)
\end{aligned}
$$

$$
\begin{aligned}
P(\mathbf{X}_n, \ldots, \mathbf{X}_1) &= P(\mathbf{X}_n | \mathbf{X}_{n-1}, \ldots, \mathbf{X}_1) P(\mathbf{X}_{n-1} | \mathbf{X}_{n-2}, \ldots, \mathbf{X}_1) \ldots P(\mathbf{X}_1) \\
&= \prod_{k=1}^{n} P(\mathbf{X}_k | \mathbf{X}_{k-1}, \ldots, \mathbf{X}_1)
\end{aligned}
$$

$$
P(I\ like\ you) = P(you | I\ ,\ like) \cdot P(like | I) \cdot P(I)
$$

Naive Bayes Classifier

- Suppose you have a binary classifier with two classes, C_1, C_2.

Naive Bayes Classifier

- Suppose you have a binary classifier with two classes, C_1, C_2.
- Given a problem instance to be classified, represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$.

Naive Bayes Classifier

- Suppose you have a binary classifier with two classes, C_1, C_2.
- Given a problem instance to be classified, represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$.
- We need to calculate the following conditional probabilities,

$$p(C_1|x_1, \ldots, x_n), \quad p(C_2|x_1, \ldots, x_n)$$

Naive Bayes Classifier

- Suppose you have a binary classifier with two classes, C_1, C_2.
- Given a problem instance to be classified, represented by a vector $\mathbf{x} = (x_1, \ldots, x_n)$.
- We need to calculate the following conditional probabilities,

$$p(C_1 | x_1, \ldots, x_n), \;\; p(C_2 | x_1, \ldots, x_n)$$

- The bigger probability determine the class of $\mathbf{x}$.

- By the Bayes formula,

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

- By the Bayes formula,

$$p(C_k|\mathbf{x}) = \frac{p(C_k)p(\mathbf{x}|C_k)}{p(\mathbf{x})}$$

- **Naive Bayes Assumption**: all features in $\mathbf{x}$ are mutually independent, i.e.,

$$p(x_i|x_{i+1}, \ldots, x_n, C_k) = p(x_i|C_k)$$

- So,

$$p(C_k|x_1, \ldots, x_n) = \frac{p(C_k)}{p(\mathbf{x})} p(x_1|C_k) \times \cdots \times p(x_n|C_k)$$

Example

- If $C_1 = 0$, $C_2 = 4$ and $\mathbf{x} = (I, like, you)$ then the probability that this sentence has the tag 0 is,

$$p(C_1|I, like, you) = \frac{p(C_1)}{p(\mathbf{x})} p(I|C_1) \times p(like|C_2) \times p(you|C_1)$$

- $p(C_1) = \frac{\text{number of sentence with tag 0}}{\text{total number of sentences}}$

Example

- If $C_1 = 0$, $C_2 = 4$ and $\mathbf{x} = (I, like, you)$ then the probability that this sentence has the tag 0 is,

$$p(C_1|I, like, you) = \frac{p(C_1)}{p(\mathbf{x})} p(I|C_1) \times p(like|C_2) \times p(you|C_1)$$

- $p(C_1) = \frac{\text{number of sentence with tag 0}}{\text{total number of sentences}}$
- $p(\mathbf{x})$ is constant.
- $p(like|C_1) = \frac{\text{(how many times "like" appears in sentences with tag 0)+1}}{\text{number of words in the dictionary}}$

Expectation

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} i\, P[\mathbf{X} = i]$$

Expectation

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} i\, P[\mathbf{X} = i]$$

If $\mathbf{X} \in \{0, 1\}$ be the random variable of tossing a fair coin then,

Expectation

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} i \, P[\mathbf{X} = i]$$

If $\mathbf{X} \in \{0, 1\}$ be the random variable of tossing a fair coin then,

$$\mathbb{E}[\mathbf{X}] = 0 \, P[\mathbf{X} = 0] + 1 \, P[\mathbf{X} = 1]$$

Expectation

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} i \, P[\mathbf{X} = i]$$

If $\mathbf{X} \in \{0, 1\}$ be the random variable of tossing a fair coin then,

$$\begin{aligned} \mathbb{E}[\mathbf{X}] &= 0 \, P[\mathbf{X} = 0] + 1 \, P[\mathbf{X} = 1] \\ &= \frac{1}{2} \end{aligned}$$

Expectation

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} i\, P[\mathbf{X} = i]$$

If $\mathbf{X} \in \{0, 1\}$ be the random variable of tossing a fair coin then,

$$\begin{aligned}\mathbb{E}[\mathbf{X}] &= 0\, P[\mathbf{X} = 0] + 1\, P[\mathbf{X} = 1] \\ &= \frac{1}{2}\end{aligned}$$

If $\mathbf{X} \in \{1, 2, 3, 4, 5, 6\}$ be the random variable of tossing a fair dice then,

Expectation

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} i \, P[\mathbf{X} = i]$$

If $\mathbf{X} \in \{0, 1\}$ be the random variable of tossing a fair coin then,

$$\begin{aligned}\mathbb{E}[\mathbf{X}] &= 0 \, P[\mathbf{X} = 0] + 1 \, P[\mathbf{X} = 1] \\ &= \frac{1}{2}\end{aligned}$$

If $\mathbf{X} \in \{1, 2, 3, 4, 5, 6\}$ be the random variable of tossing a fair dice then,

$$\mathbb{E}[\mathbf{X}] = 1 \, P[\mathbf{X} = 1] + 2 \, P[\mathbf{X} = 2] + \cdots + 6 \, P[\mathbf{X} = 6]$$

Expectation

$$\mathbb{E}[\mathbf{X}] = \sum_{i=1}^{n} i\, P[\mathbf{X} = i]$$

If $\mathbf{X} \in \{0, 1\}$ be the random variable of tossing a fair coin then,

$$\begin{aligned}\mathbb{E}[\mathbf{X}] &= 0\, P[\mathbf{X} = 0] + 1\, P[\mathbf{X} = 1] \\ &= \frac{1}{2}\end{aligned}$$

If $\mathbf{X} \in \{1, 2, 3, 4, 5, 6\}$ be the random variable of tossing a fair dice then,

$$\begin{aligned}\mathbb{E}[\mathbf{X}] &= 1\, P[\mathbf{X} = 1] + 2\, P[\mathbf{X} = 2] + \cdots + 6\, P[\mathbf{X} = 6] \\ &= \frac{1+2+3+4+5+6}{6} \\ &= 3.5\end{aligned}$$

Variance

$$\mathbb{V}[\mathbf{X}] =$$

Variance

$$\mathbb{V}[\mathbf{X}] = \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2]$$

Variance

$$
\begin{aligned}
\mathbb{V}[\mathbf{X}] &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] \\
&= \mathbb{E}[\mathbf{X}^2] - (\mathbb{E}[\mathbf{X}])^2
\end{aligned}
$$

Variance

$$
\begin{aligned}
\mathbb{V}[\mathbf{X}] &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] \\
&= \mathbb{E}[\mathbf{X}^2] - (\mathbb{E}[\mathbf{X}])^2 \\
&= \sum_{i=1}^{n} i^2\, P[\mathbf{X} = i] - (\sum_{i=1}^{n} i\, P[\mathbf{X} = i])^2
\end{aligned}
$$

Variance

$$
\begin{aligned}
\mathbb{V}[\mathbf{X}] &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] \\
&= \mathbb{E}[\mathbf{X}^2] - (\mathbb{E}[\mathbf{X}])^2 \\
&= \sum_{i=1}^{n} i^2 \, P[\mathbf{X} = i] - (\sum_{i=1}^{n} i \, P[\mathbf{X} = i])^2
\end{aligned}
$$

Tossing a coin,

Variance

$$\begin{aligned}
\mathbb{V}[\mathbf{X}] &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] \\
&= \mathbb{E}[\mathbf{X}^2] - (\mathbb{E}[\mathbf{X}])^2 \\
&= \sum_{i=1}^{n} i^2 \, P[\mathbf{X} = i] - (\sum_{i=1}^{n} i \, P[\mathbf{X} = i])^2
\end{aligned}$$

Tossing a coin,

$$\mathbb{V}[\mathbf{X}] =$$

Variance

$$\begin{aligned}
\mathbb{V}[\mathbf{X}] &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] \\
&= \mathbb{E}[\mathbf{X}^2] - (\mathbb{E}[\mathbf{X}])^2 \\
&= \sum_{i=1}^{n} i^2\, P[\mathbf{X} = i] - (\sum_{i=1}^{n} i\, P[\mathbf{X} = i])^2
\end{aligned}$$

Tossing a coin,

$$\mathbb{V}[\mathbf{X}] = (0^2\, P[\mathbf{X} = 0] + 1^2\, P[\mathbf{X} = 1]) - (0\, P[\mathbf{X} = 0] + 1\, P[\mathbf{X} = 1])^2$$

Variance

$$\begin{aligned}\mathbb{V}[\mathbf{X}] &= \mathbb{E}[(\mathbf{X}-\mathbb{E}[\mathbf{X}])^2] \\ &= \mathbb{E}[\mathbf{X}^2] - (\mathbb{E}[\mathbf{X}])^2 \\ &= \sum_{i=1}^{n} i^2\, P[\mathbf{X}=i] - (\sum_{i=1}^{n} i\, P[\mathbf{X}=i])^2\end{aligned}$$

Tossing a coin,

$$\begin{aligned}\mathbb{V}[\mathbf{X}] &= (0^2\, P[\mathbf{X}=0] + 1^2\, P[\mathbf{X}=1]) - (0\, P[\mathbf{X}=0] + 1\, P[\mathbf{X}=1])^2 \\ &= \frac{1}{2} - (\frac{1}{2})^2\end{aligned}$$

Variance

$$\begin{aligned}\mathbb{V}[\mathbf{X}] &= \mathbb{E}[(\mathbf{X} - \mathbb{E}[\mathbf{X}])^2] \\ &= \mathbb{E}[\mathbf{X}^2] - (\mathbb{E}[\mathbf{X}])^2 \\ &= \sum_{i=1}^{n} i^2\, P[\mathbf{X} = i] - (\sum_{i=1}^{n} i\, P[\mathbf{X} = i])^2\end{aligned}$$

Tossing a coin,

$$\begin{aligned}\mathbb{V}[\mathbf{X}] &= (0^2\, P[\mathbf{X} = 0] + 1^2\, P[\mathbf{X} = 1]) - (0\, P[\mathbf{X} = 0] + 1\, P[\mathbf{X} = 1])^2 \\ &= \frac{1}{2} - (\frac{1}{2})^2 \\ &= \frac{1}{4}\end{aligned}$$

Normal (Gaussian) Distribution

Why Should I care about Normal Distribution?

- It is the most important distribution in nature.

- Examples: weight, height, blood pressure, etc.

- It is a fundamental concept in data science.

Probability Density Function (PDF)

```python
import numpy as np
import matplotlib.pyplot as plt
```

```python
def normal(x,mu,sigma):
    return (1/(np.sqrt(2*np.pi*sigma**2))*np.exp(-(x-mu)**2/(2*sigma**2)))
```

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

mean

standard deviation

$$X \sim \mathcal{N}(\mu, \sigma^2)$$

If $\mu = 0$, $\sigma = 1$ it is called standard normal distribution.

Code

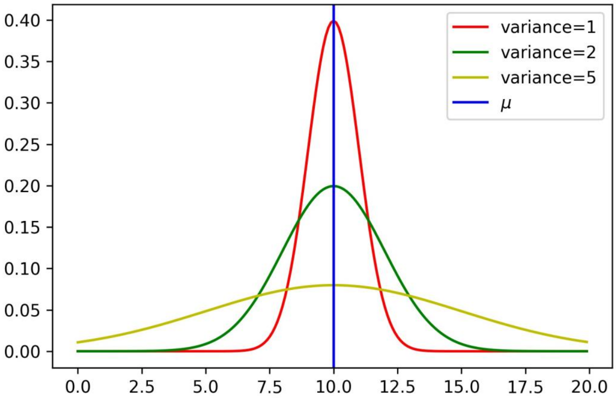
```
x=np.arange(0,20,0.1)
v1=1
v2=2
v3=5
y1=normal(x,10,v1)
y2=normal(x,10,v2)
y3=normal(x,10,v3)
```

Generating three different normal distribution with same mean but different variance

```
plt.plot(x,y1,'r',label='variance={}'.format(v1))
plt.plot(x,y2,'g',label='variance={}'.format(v2))
plt.plot(x,y3,'y',label='variance={}'.format(v3))
plt.axvline(x=10, c='b',label='$\mu$')
plt.legend()
```

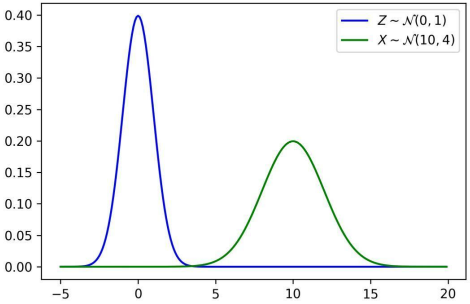
Draw vertical line x=10

Increasing Variance Reduce the Height of Curve

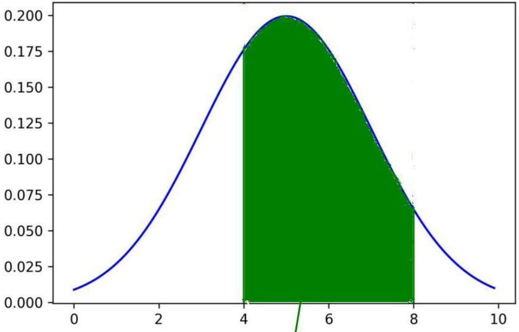


Converting to Standard Normal Distribution

If $X \sim \mathcal{N}(\mu, \sigma^2) \longrightarrow$ then $Z = \frac{X-\mu}{\sigma}$ is a standard random variable, i.e., $Z \sim \mathcal{N}(0, 1)$



Area Under Normal Distribution



$\mathbb{P}[4 < X < 8]$

Probability of X to be between 4 and 8

scipy

```
import math
from scipy import stats
A = stats.norm(3, math.sqrt(16)) # Declare A to be a normal random variable
print A.pdf(4)       # f(3), the probability density at 3
print A.cdf(2)       # F(2), which is also P(Y < 2)
print A.rvs()        # Get a random sample from A
```

What is N-day moving average?

In a time series it compute the average of past N days as the present value.

If f represent the time series of Bitcoin then a 7-day moving average of f is another time series, say g, such that today's value of g is the average price of past 7 days.

How to compute the moving average?

Use Convolution of original series with a kernel of average weight to obtain moving average.

$$(f \star g)[n] = \sum_{i=-\infty}^{\infty} f[i] \cdot g[n-i]$$

Don't worry about the formula of the convolution, Numpy will compute it for you!

Example

Let f be a time series and we want to compute the 10-day moving average.

Let g=$[0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1,0.1]$ be the kernel

np.convolve(f,g)=convolution of f and g by numpy

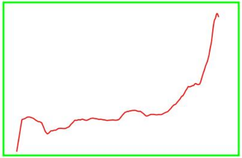
Convolution is act like a filter.

original series $\longrightarrow$ $\boxed{\star\, g}$ $\longrightarrow$ smooth time series

$$\boxed{f} \xrightarrow{g = [0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]} \boxed{f \star g}$$

input

output



Python Code

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
```

```python
bitcoin=pd.read_csv("BTC-USD.csv", index_col="Date")
```

```python
kernel=[0.1 for i in range(10)]
```

```python
kernel
```

```
[0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1, 0.1]
```

```python
smooth=np.convolve(bitcoin['Open'],kernel)
```

convolution of bitcoin['Open'] and kernel

Thank You