Probability in Machine Learning

Three Axioms of Probability

- Given an Event E in a sample space S, $S = \bigcup_{i=1}^{N} E_i$
- First axiom
  - $P(E) \in \mathbb{R}, 0 \leq P(E) \leq 1$
- Second axiom
  - $P(S) = 1$
- Third axiom
  - Additivity, any countable sequence of mutually exclusive events E_i
  - $P(\bigcup_{i=1}^{n} E_i) = P(E_1) + P(E_2) + \cdots + P(E_n) = \sum_{i=1}^{n} P(E_i)$

Random Variables

- A random variable is a variable whose values are numerical outcomes of a random phenomenon.

- Discrete variables and Probability Mass Function (PMF)

$$\sum_{x} p_X(x) = 1$$

- Continuous Variables and Probability Density Function (PDF)

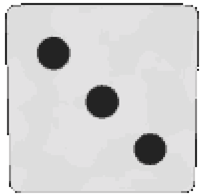
$$\Pr[a \leq X \leq b] = \int_{a}^{b} f_X(x)dx$$

Expected Value (Expectation)

- Expectation of a random variable X:

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i\, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k .$$

- Expected value of rolling one dice?



$$\mathrm{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5.$$

https://en.wikipedia.org/wiki/Expected_value

Expected Value of Playing Roulette

- Bet \$1 on single number (0 ~ 36), and get \$35 payoff if you win. What's the expected value?

$$E[Gain\ from\ \$1\ bet] = -1 \times \frac{36}{37} + 35 \times \frac{1}{37} = \frac{-1}{37}$$

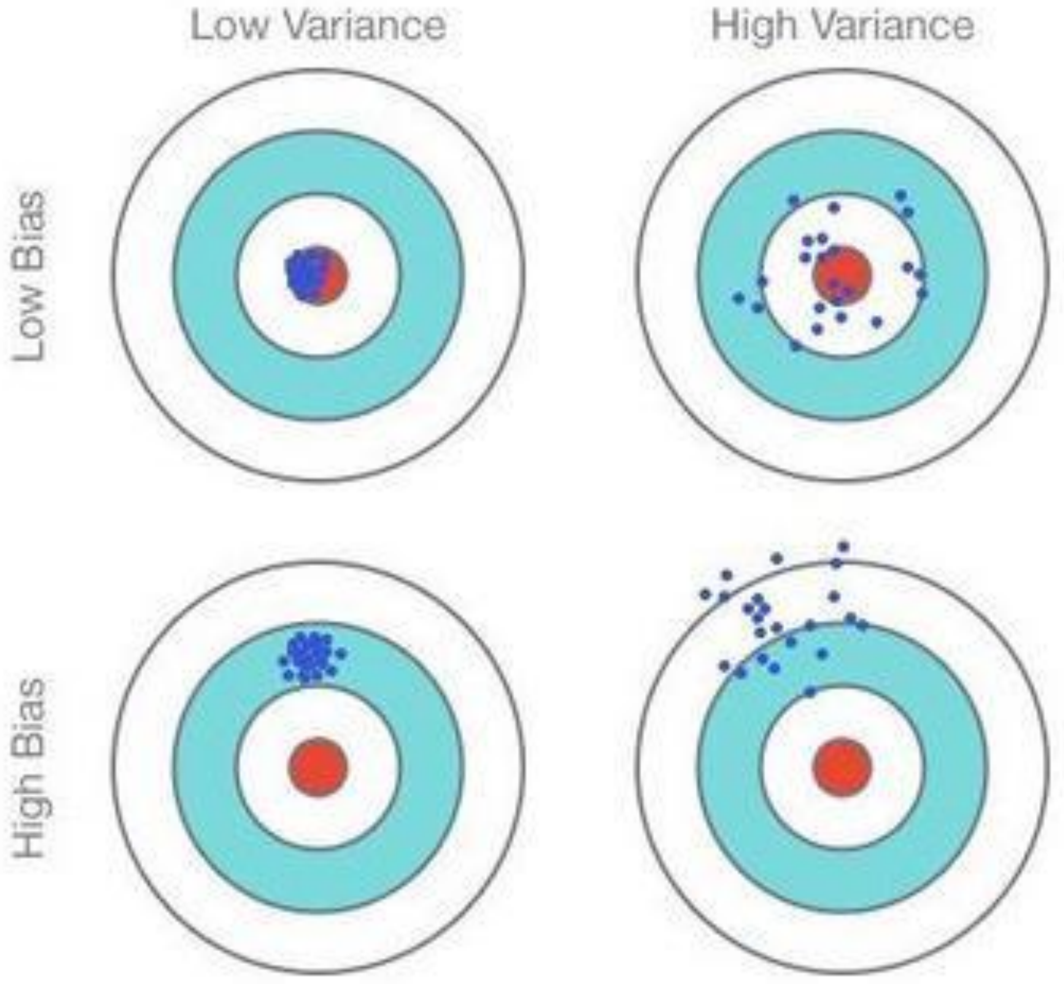
Variance

- The variance of a random variable ***X*** is the expected value of the squared deviation from the mean of ***X***

$$\mathrm{Var}(\boldsymbol{X}) = E[(\boldsymbol{X} - \mu)^2]$$

$$\begin{aligned} \mathrm{Var}(X) &= \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] \\ &= \mathrm{E}\left[X^2 - 2X\,\mathrm{E}[X] + \mathrm{E}[X]^2\right] \\ &= \mathrm{E}\left[X^2\right] - 2\,\mathrm{E}[X]\,\mathrm{E}[X] + \mathrm{E}[X]^2 \\ &= \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \end{aligned}$$

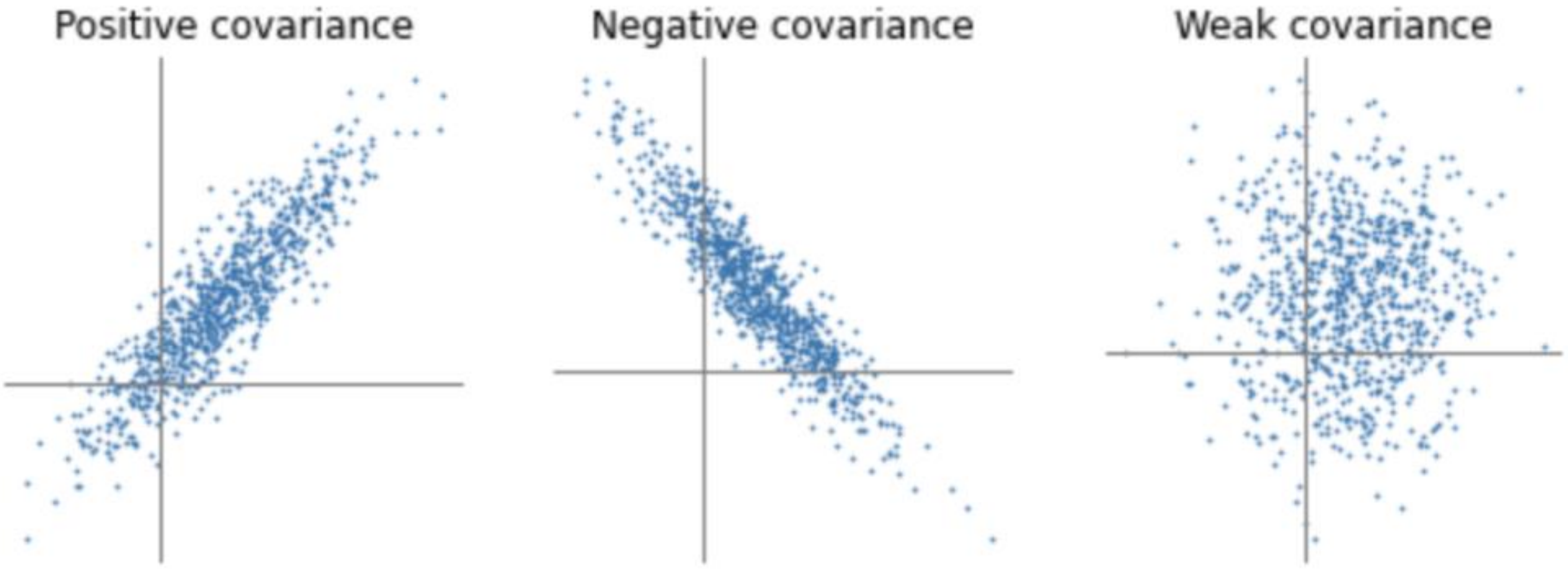
Bias and Variance



Covariance

- Covariance is a measure of the joint variability of two random variables.

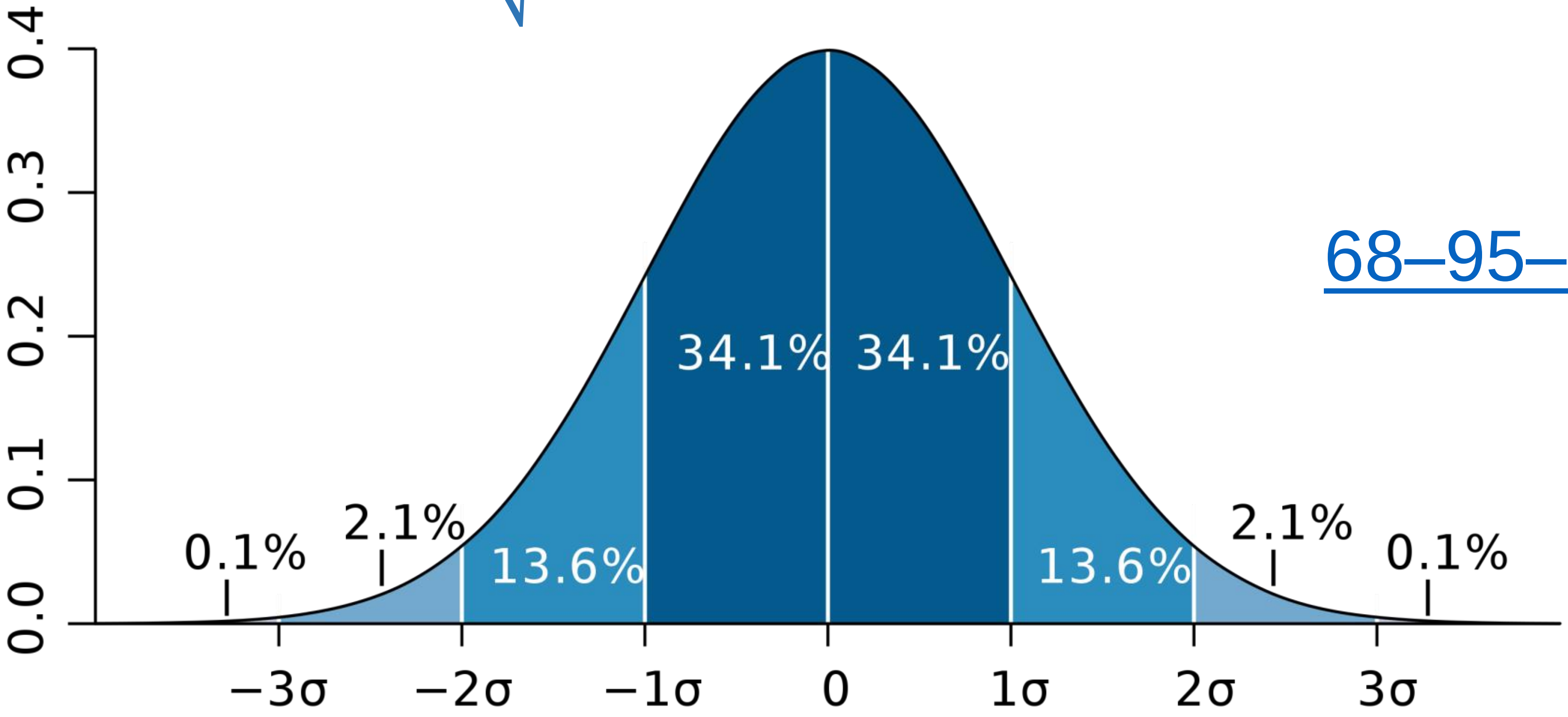
$$Cov(X,Y) = E[X - E[X]]E[Y - E[Y]] = E[XY] - E[X]E[Y]$$



https://programmathically.com/covariance-and-correlation/

Standard Deviation

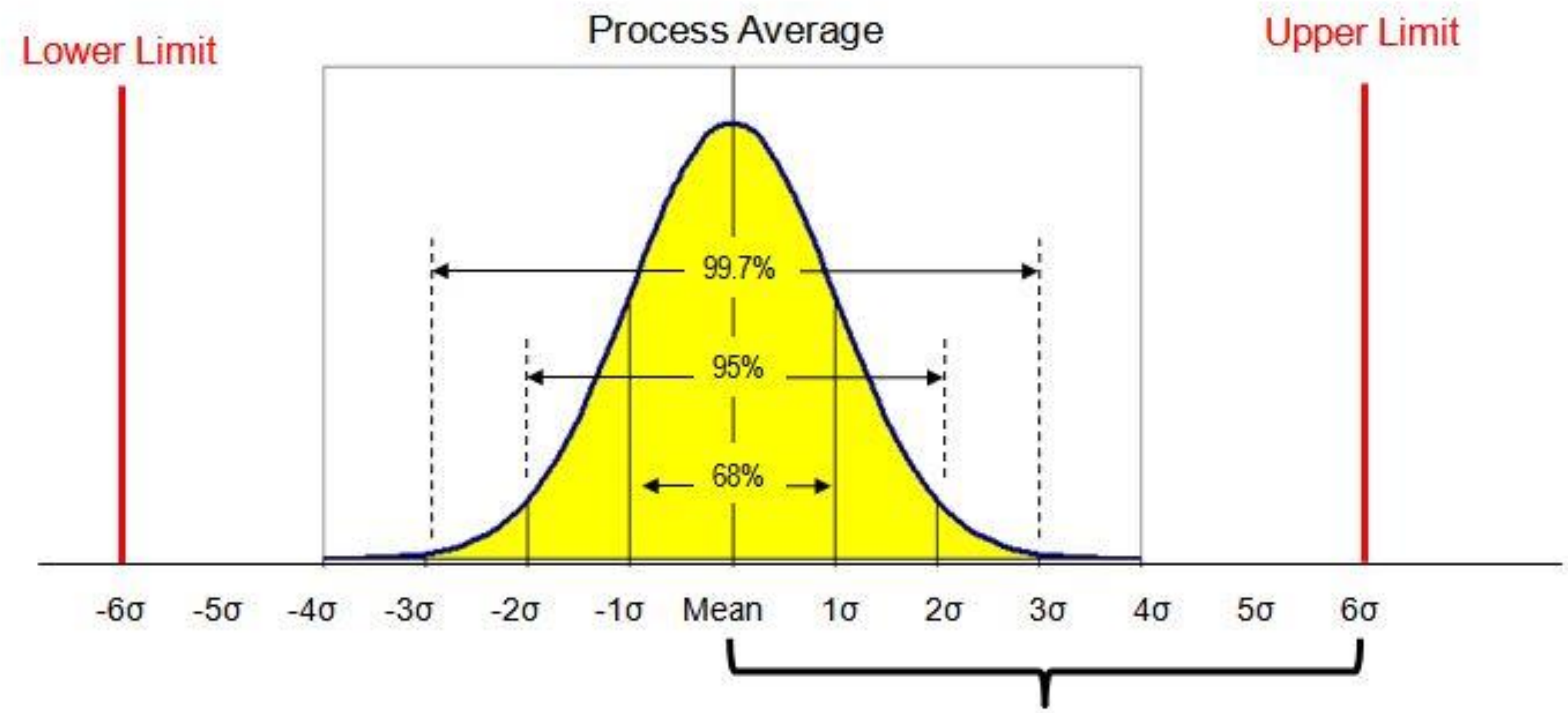
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} = \sqrt{Var(\boldsymbol{X})}$$



68–95–99.7 rule

6 Sigma

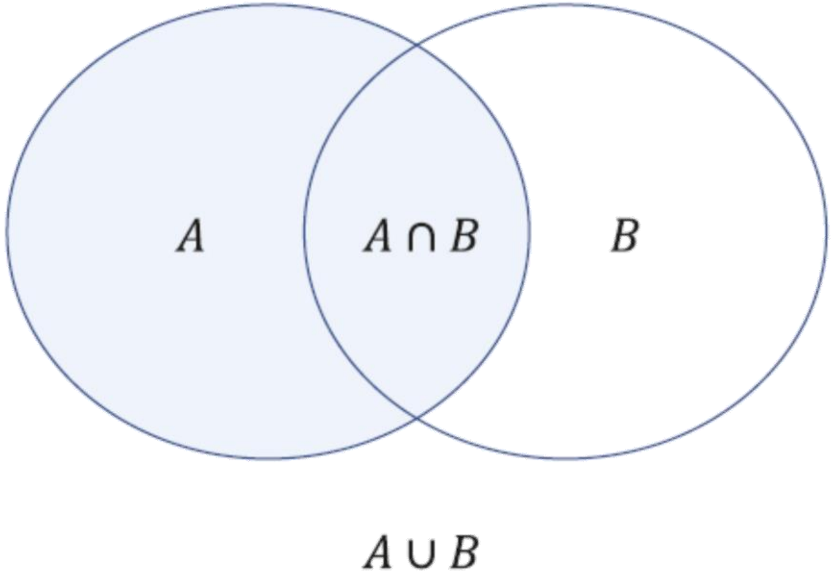
- A product has 99.99966% chance to be free of defects



https://www.leansixsigmadefinition.com/glossary/six-sigma/

Union, Intersection, and Conditional Probability

- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- $P(A \cap B)$ is simplified as $P(AB)$
- Conditional Probability $P(A|B)$, the probability of event A given B has occurred
  - $P(A|B) = P\left(\frac{AB}{B}\right)$
  - $P(AB) = P(A|B)P(B) = P(B|A)P(A)$



Chain Rule of Probability

- The joint probability can be expressed as chain rule

$$P(A_1A_2A_3\ldots A_n) = P(A_1)P(A_2/A_1)P(A_3/A_1A_2)\ldots\ldots P\left(A_n/A_1A_2..A_{(n-1)}\right)$$

$$P(A_2/A_1) = \frac{P(A_1A_2)}{P(A_1)} \qquad P(A_3/A_1A_2) = \frac{P(A_1A_2A_3)}{P(A_1A_2)}$$

Mutually Exclusive

- $P(AB) = 0$
- $P(A \cup B) = P(A) + P(B)$

Independence of Events

- Two events A and B are said to be independent if the probability of their intersection is equal to the product of their individual probabilities
  - $P(AB) = P(A)P(B)$
  - $P(A|B) = P(A)$

Bayes Rule

Training Data

likelihood

Prior

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

Evidence

Class
(Cat / Dog)

features
(Pixels of an image)

Proof:

Remember $P(A|B) = P\left(\frac{AB}{B}\right)$

So $P(AB) = P(A|B)P(B) = P(B|A)P(A)$

Then Bayes $P(A|B) = P(B|A)P(A)/P(B)$

Naïve Bayes Classifier

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

$$p(C_k \mid x_1, \dots, x_n)$$

$$\begin{aligned} p(C_k, x_1, \dots, x_n) &= p(x_1, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2, \dots, x_n, C_k) \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k)\, p(x_3, \dots, x_n, C_k) \\ &= \dots \\ &= p(x_1 \mid x_2, \dots, x_n, C_k)\, p(x_2 \mid x_3, \dots, x_n, C_k) \cdots p(x_{n-1} \mid x_n, C_k)\, p(x_n \mid C_k)\, p(C_k) \end{aligned}$$

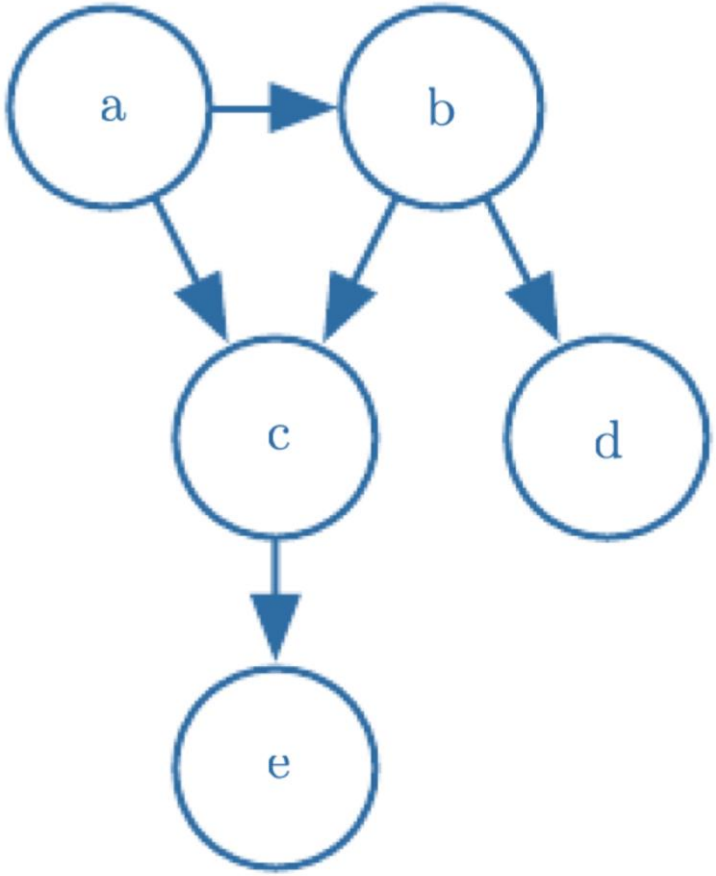
Naïve = Assume All Features Independent

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

$$\begin{aligned} p(C_k \mid x_1, \ldots, x_n) &\propto p(C_k, x_1, \ldots, x_n) \\ &= p(C_k)\, p(x_1 \mid C_k)\, p(x_2 \mid C_k)\, p(x_3 \mid C_k)\, \cdots \\ &= p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k)\,, \end{aligned}$$

Graphical Model

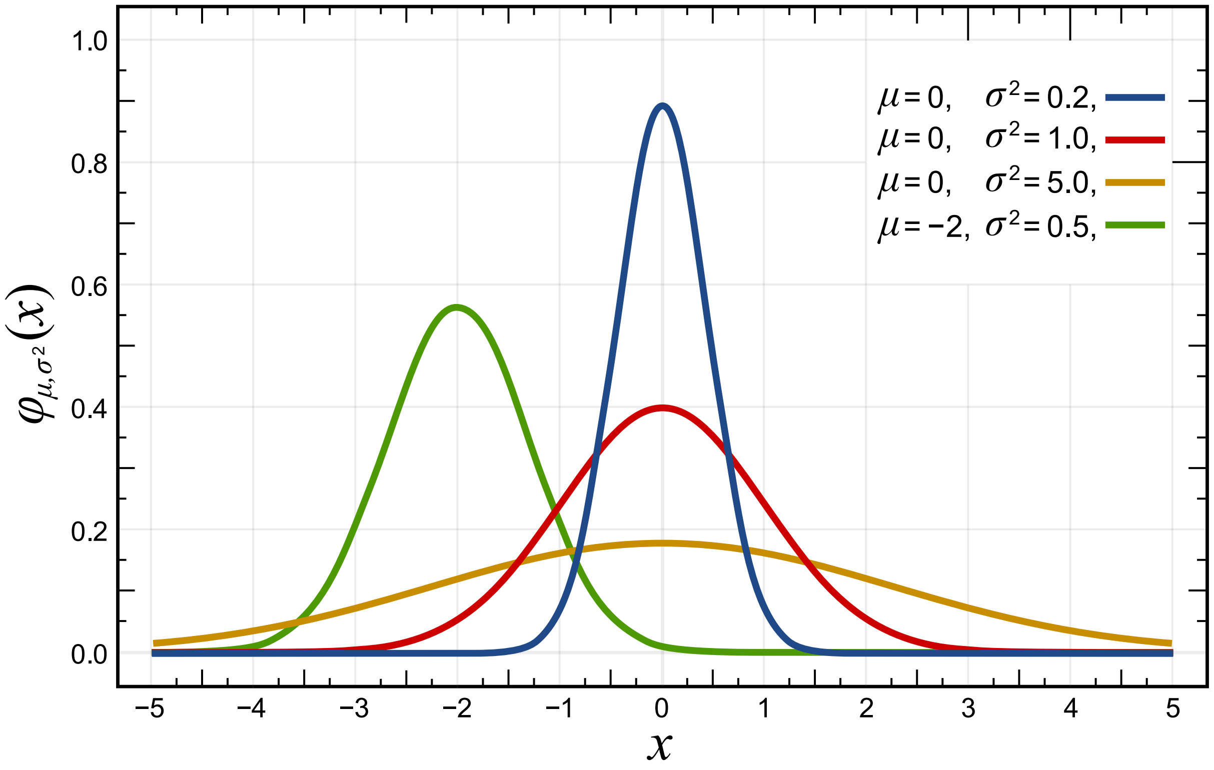
$$p(a,b,c,d,e) = p(a)p(b|a)p(c|a,b)p(d|b)p(e|c)$$



Normal (Gaussian) Distribution

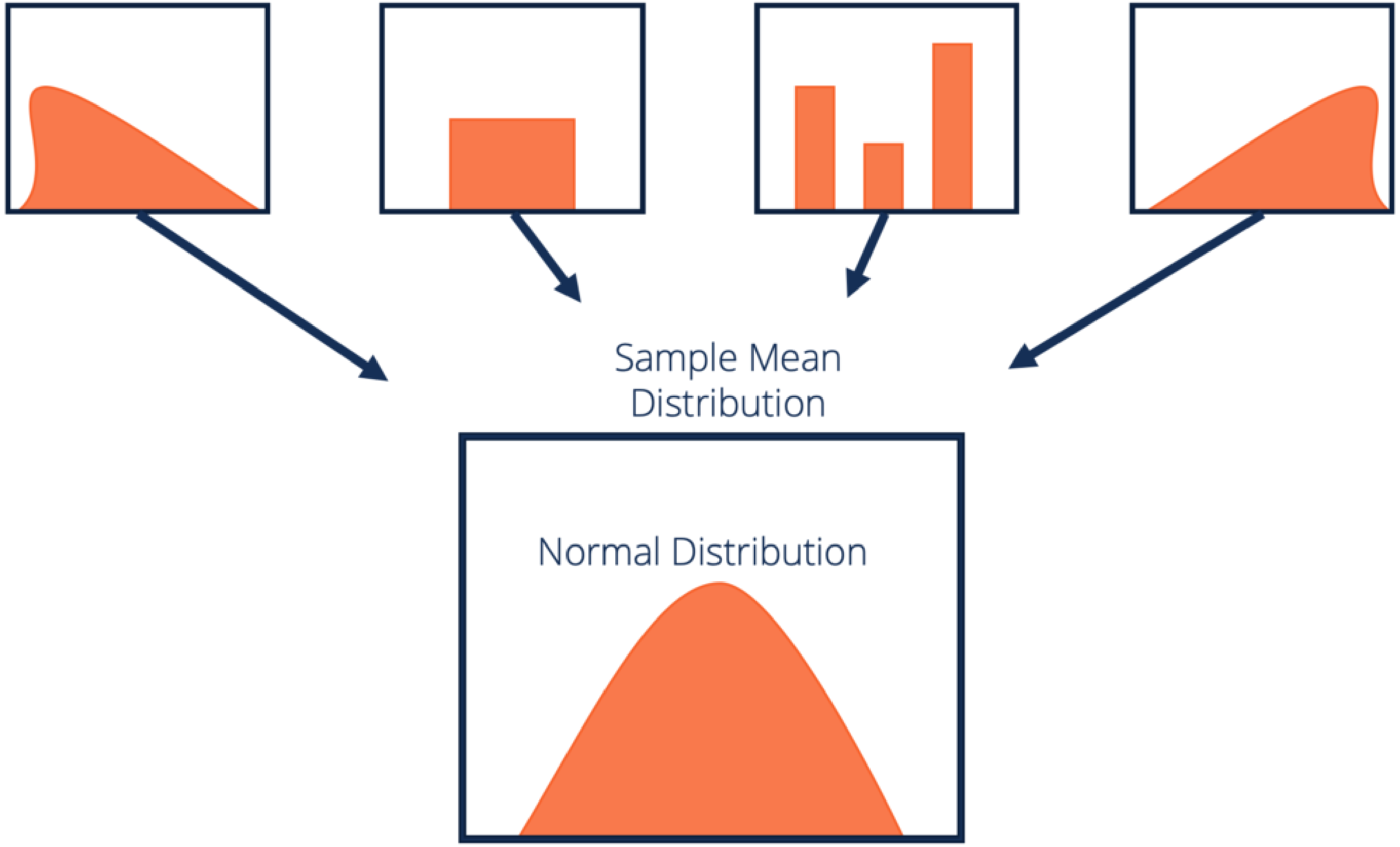
- A type of continuous probability distribution for a real-valued random variable.
- One of the most important distributions

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



Central Limit Theory

- Averages of samples of observations of random variables independently drawn from independent distributions converge to the normal distribution



https://corporatefinanceinstitute.com/resources/knowledge/other/central-limit-theorem/

Bernoulli Distribution

- Definition

$$\Pr(X=1) = p = 1 - \Pr(X=0) = 1 - q.$$

- PMF

$$f(k;p) = \begin{cases} p & \text{if } k = 1, \\ q = 1 - p & \text{if } k = 0. \end{cases}$$

- $E[X] = p$
- $\mathrm{Var}(X) = pq$

https://acegif.com/flipping-coin-gifs/

https://en.wikipedia.org/wiki/Bernoulli_distribution

Information Theory

- Self-information:

$$I(x) = -\log P(x)$$

$$\log \frac{1}{P(x)} = -\log P(x)$$

- Shannon Entropy:

$$H = -\sum_i p_i \log_2 p_i$$

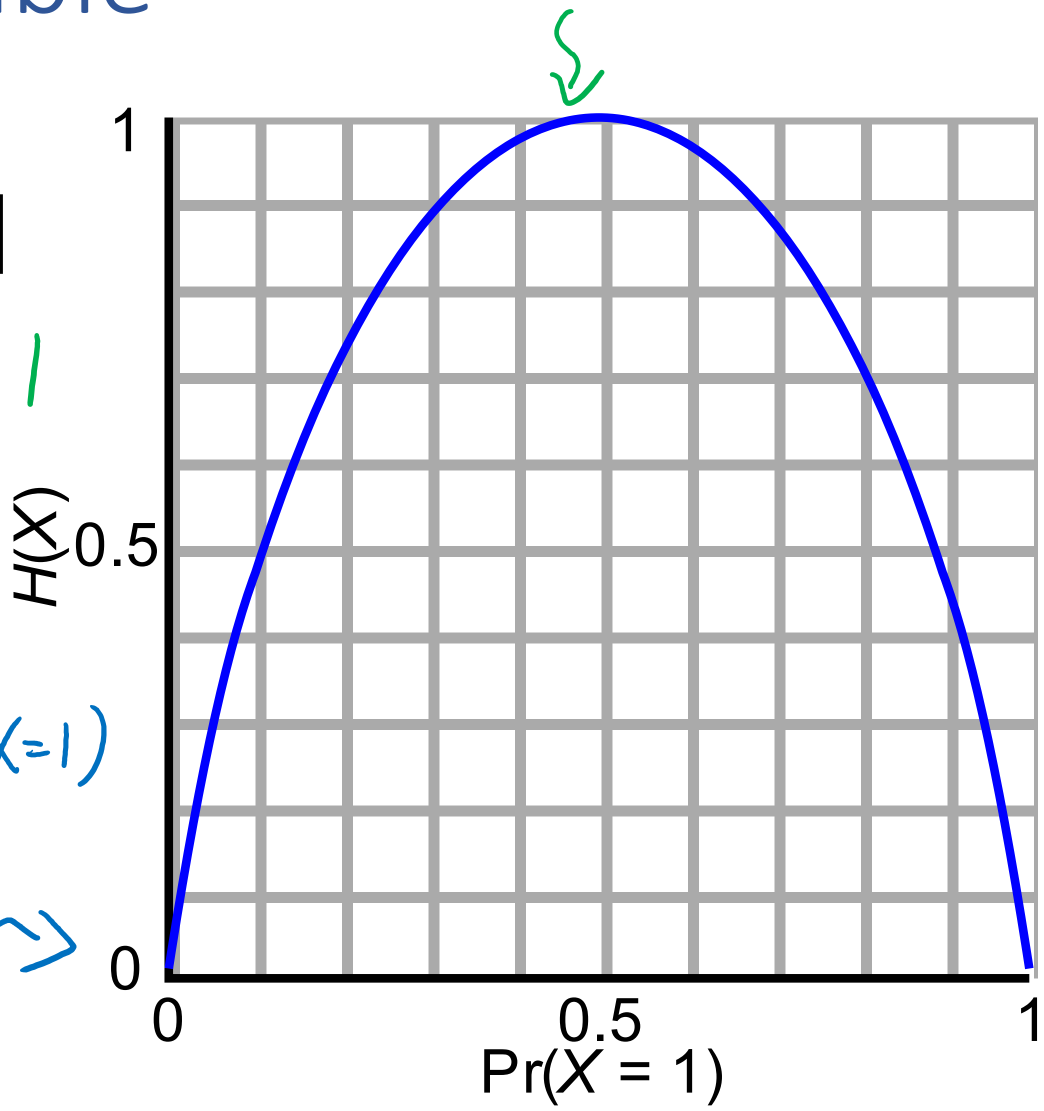
Entropy of Bernoulli Variable

- $H(x) = E[I(x)] = -E[\log P(x)]$

$H(x) = -0.5 \times \log_2 \frac{1}{2} - 0.5 \times \log_2 \frac{1}{2} = 1$

$H(X) = -1 \times \underbrace{\log_2 1}_{=0} - 0 \times \log_2 0$

Pr(X=1)



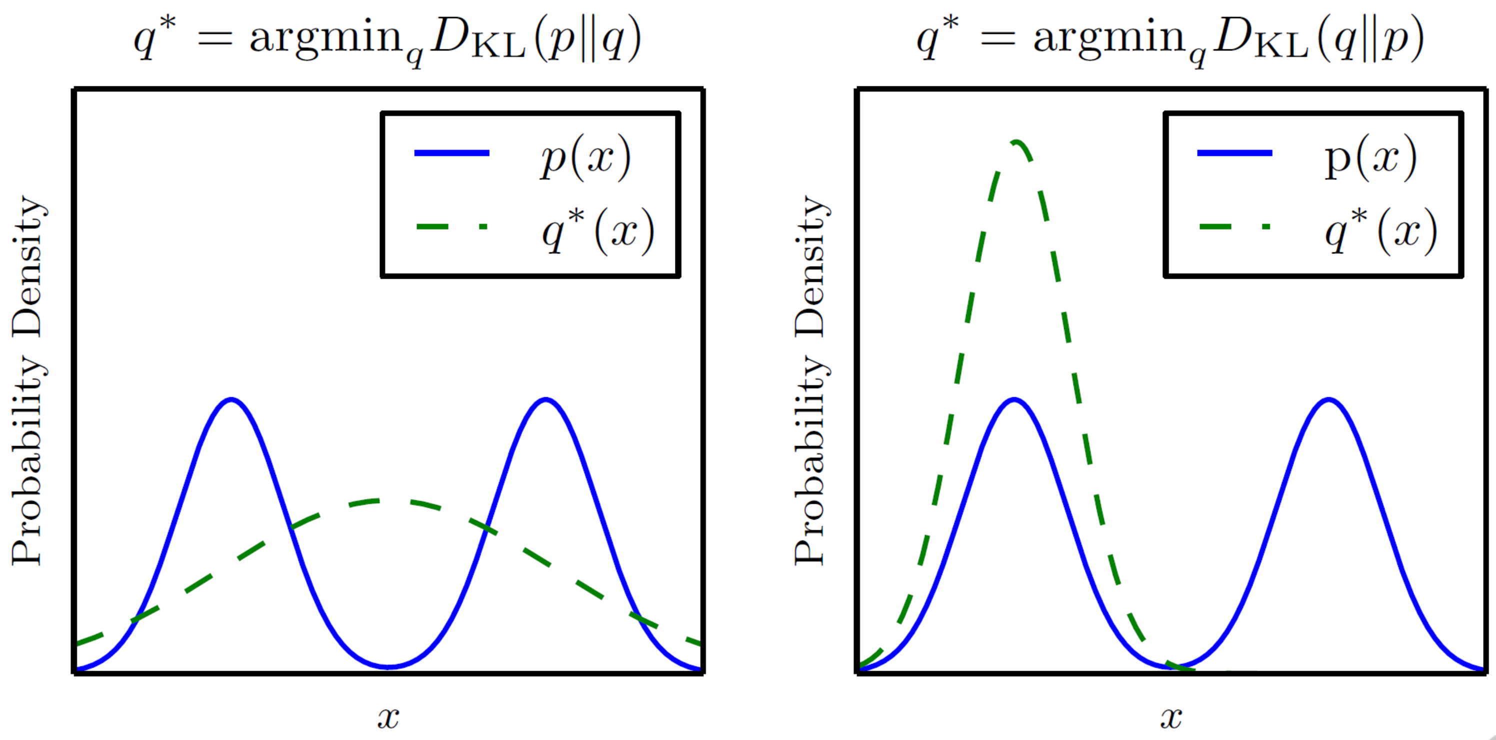
https://en.wikipedia.org/wiki/Information_theory

Kullback-Leibler (KL) Divergence

- $D_{KL}(p||q) = E[\log P(X) - \log Q(X)] = E\left[\log \frac{P(x)}{Q(x)}\right]$

KL Divergence is Asymmetric!

$D_{KL}(p||q) \mathrel{!=} D_{KL}(q||p)$



https://www.deeplearningbook.org/slides/03_prob.pdf

Key Takeaways

- Expected value (expectation) is mean (weighted average) of a random variable
- Event A and B are independent if $P(AB) = P(A)P(B)$
- Event A and B are mutually exclusive if $P(AB) = 0$
- Central limit theorem tells us that Normal distribution is the one, if the data probability distribution is unknown
- Entropy is expected value of self information $-E[\log P(x)]$
- KL divergence can measure the difference of two probability distributions and is asymmetric