ARTIFICIAL INTELLIGENCE

WOOHWAN JUNG

Linear Algebra & Probability Theory

Today's Lecture

- Topics
  - Linear algebra
  - Probability theory

- Not a comprehensive study on linear algebra and probability theory
- Focused on the subset that is most relevant to deep learning

Learning Proce

A lot of vector/matrix operations

Input

Model

Architecture
+
Parameters

Output

$P(Cat) = 0.9$

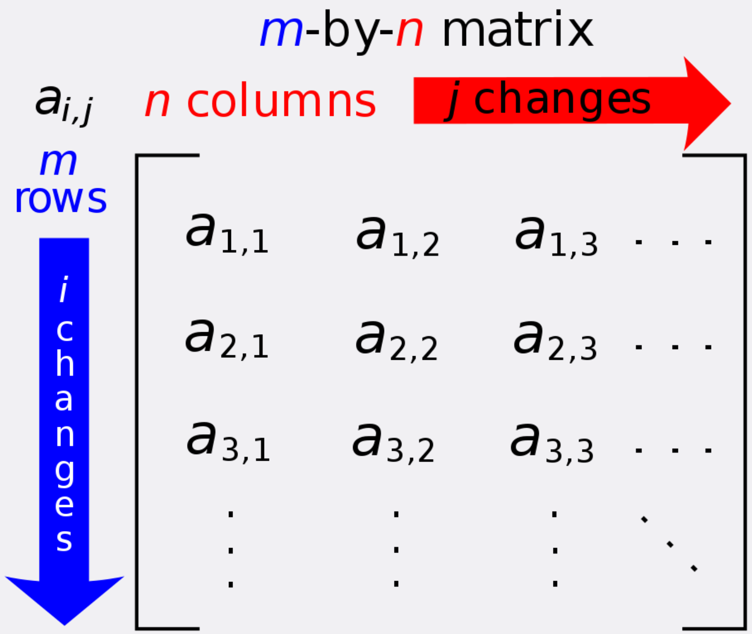
Label

$P(Cat) = 1$

Compute loss

Update parameters to minimize the loss

LINEAR ALGEBRA



Scalars and Vectors

- Scalars
  - A scalar is a single number
  - Usually denoted with italic font: a, n, x
  - Integers, real numbers, rational numbers, etc.
  - Example notation: $x \in \mathbb{R}, x \in \mathbb{Z}, x \in \mathbb{N}$
- Vectors
  - A vector is a 1-D array of numbers
  - Can be real, integer, etc.
  - Example notation for type and size:
    - $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \in \mathbb{Z}^n$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Matrices

- A matrix is a 2-D array of numbers:

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$$

- Example notation for type and shape:

$$\boldsymbol{A} \in \mathbb{R}^{m \times n}$$

Tensors

- A tensor is an (muti-dimensional) array of numbers, that may have
  - Zero dimensions, and be a scalar
  - One dimension, and be a vector
  - Two dimensions, and be a matrix
  - And more dimensions

Matrix Transpose

$$\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \\ A_{3,1} & A_{3,2} \end{bmatrix} \Rightarrow \boldsymbol{A}^\top = \begin{bmatrix} A_{1,1} & A_{2,1} & A_{3,1} \\ A_{1,2} & A_{2,2} & A_{3,2} \end{bmatrix}$$

Figure 2.1: The transpose of the matrix can be thought of as a mirror image across the main diagonal.

Matrix Addition and Subtraction

- Adding or subtracting corresponding elements
- Let $\boldsymbol{A} = \begin{bmatrix} A_{1,1} & A_{1,2} \\ A_{2,1} & A_{2,2} \end{bmatrix}$, $\boldsymbol{B} = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix}$
- $\boldsymbol{A} + \boldsymbol{B} = \begin{bmatrix} A_{1,1} + B_{1,1} & A_{1,2} + B_{1,2} \\ A_{2,1} + B_{2,1} & A_{2,2} + B_{2,2} \end{bmatrix}$
- $\boldsymbol{A} - \boldsymbol{B} = \begin{bmatrix} A_{1,1} - B_{1,1} & A_{1,2} - B_{1,2} \\ A_{2,1} - B_{2,1} & A_{2,2} - B_{2,2} \end{bmatrix}$

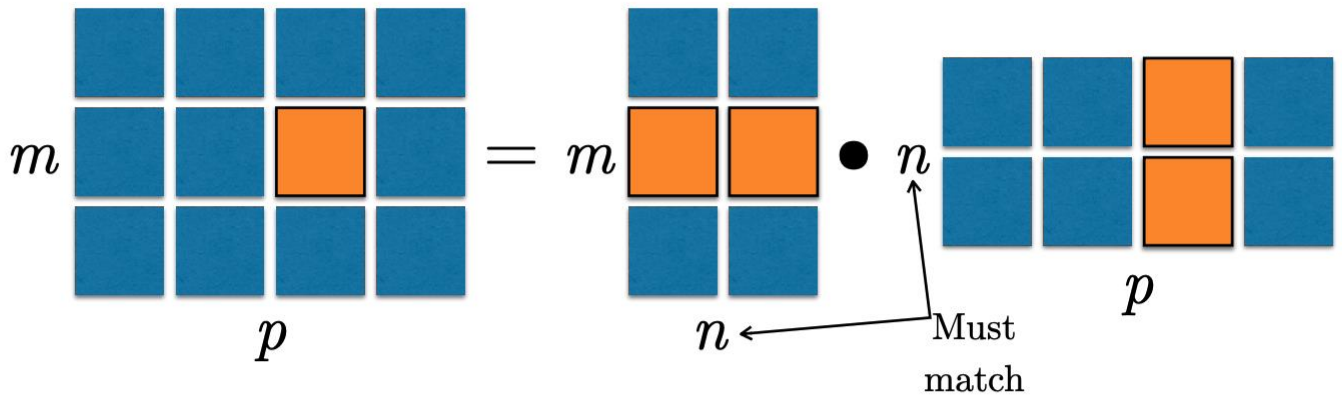
Matrix Multiplication

- For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$ and $\boldsymbol{B} \in \mathbb{R}^{n \times p}$

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}.$$

$$C_{i,j} = \sum_k A_{i,k} B_{k,j}.$$

$\mathbf{C} \in \mathbb{R}^{m \times p}$



Shape of the Result of Vector/Matrix Multiplication

- For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{B} \in \mathbb{R}^{n \times p}$, and $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B}$
  - $\boldsymbol{C} \in \mathbb{R}^{m \times p}$

- For $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, $\boldsymbol{x} \in \mathbb{R}^{n}$, and $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$
  - $\boldsymbol{y} \in \mathbb{R}^{m}$

Norms

- Functions that measure how "Large" a vector is
  - Lp norm

$$\|\boldsymbol{x}\|_p = \left(\sum_i |x_i|^p\right)^{\frac{1}{p}}$$

  - Most popular norm: L2 norm (p=2)
  - L1 norm (p=1): $\|\boldsymbol{x}\|_1 = \sum_i |x_i|$
  - Max norm ($p = \infty$): $\|\boldsymbol{x}\|_\infty = \max_i |x_i|$
- Frobenius norm of a matrix
  - $\|\boldsymbol{A}\|_p = \sqrt{\sum_i \sum_{ij} A_i^2}$

Distance Between a Pair of Vectors

- Norm of $\boldsymbol{x} - \boldsymbol{y}$
- Lp distance

$$\|\boldsymbol{x} - \boldsymbol{y}\|_p = \left(\sum_i |x_i - y_i|^p\right)^{\frac{1}{p}}$$

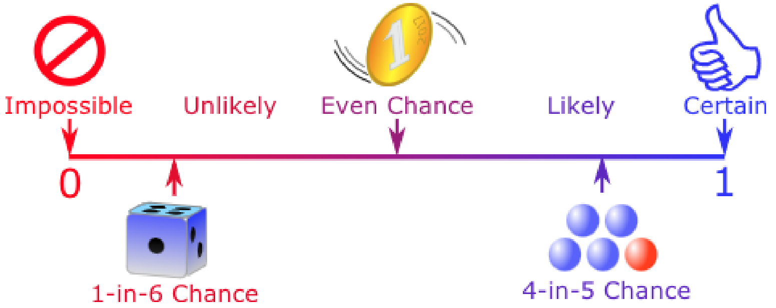
PROBABILITY THEORY

Random Variable

- A **variable** whose values depend on outcomes of a **random phenomenon**
- Discrete random variables
  - Bernoulli r.v.
  - Categorical r.v.
- Continuous random variables
  - Gaussian (Normal) r.v.
  - Laplace r.v.

Probability Mass Function (PMF): $P(x)$

- A function that gives the probability that a discrete random variable is equal to some value
- For all x, $0 \leq P(X = x) \leq 1$
- $\sum P(X = x) = 1$

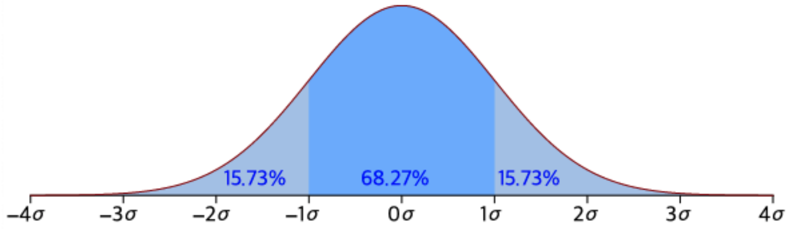


Probability is always between 0 and 1

- Example: discrete uniform distribution
  - $P(X = x) = \frac{1}{k}$ where k is the number of possible values

Probability Density Function (PDF): $p(x)$

- The PDF of a continuous random variable gives the *relative* likelihood of any outcome x
- Properties
  - $p(x) \geq 0$, (Note: we do not require $p(x) \leq 1$)
  - $\int p(x)dx = 1$



- Example: continuous uniform distribution u(a,b)

$$p(x) = \frac{1}{b-a}$$

Marginal Probability Distribution

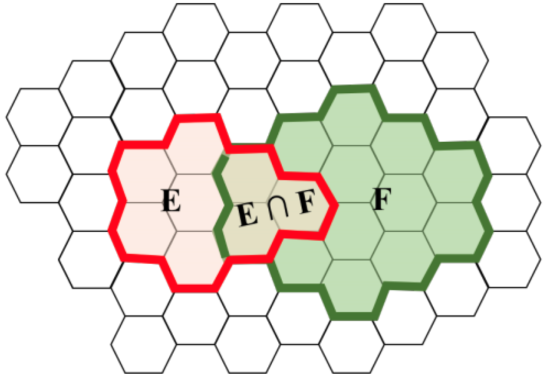
- A probability distribution over the subset of variables
- Sum rule

$$\forall x \in \mathrm{x}, P(\mathrm{x} = x) = \sum_{y} P(\mathrm{x} = x, \mathrm{y} = y).$$

$$p(x) = \int p(x, y) dy.$$

Conditional Probability

$$P(\mathrm{y} = y \mid \mathrm{x} = x) = \frac{P(\mathrm{y} = y, \mathrm{x} = x)}{P(\mathrm{x} = x)}.$$



$$P(E) = \frac{8}{50} \approx 0.16$$

$$P(E|F) = \frac{3}{14} \approx 0.21$$

Course Reading, Stanford CS 109

Chain Rule

$$P(\mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(n)}) = P(\mathrm{x}^{(1)}) \Pi_{i=2}^{n} P(\mathrm{x}^{(i)} \mid \mathrm{x}^{(1)}, \ldots, \mathrm{x}^{(i-1)}).$$

- Example (n=4)

$$P\left(X^{(1)}, X^{(2)}, X^{(3)}, X^{(4)}\right)$$
$$= P\left(X^{(1)}\right) P\left(X^{(2)} | X^{(1)}\right) P\left(X^{(3)} | X^{(1)} X^{(2)}\right)$$

Independence

- Two random variables x and y are independent if and only if

$$\forall x \in \mathrm{x}, y \in \mathrm{y},\ p(\mathrm{x}=x, \mathrm{y}=y) = p(\mathrm{x}=x)p(\mathrm{y}=y).$$

- If x and y are conditionally independent

$$\forall x \in \mathrm{x}, y \in \mathrm{y}, z \in \mathrm{z},\ p(\mathrm{x}=x, \mathrm{y}=y \mid \mathrm{z}=z) = p(\mathrm{x}=x \mid \mathrm{z}=z)p(\mathrm{y}=y \mid \mathrm{z}=z).$$

Expectation

$$\mathbb{E}_{\mathrm{x}\sim P}[f(x)] = \sum_x P(x)f(x),$$

$$\mathbb{E}_{\mathrm{x}\sim p}[f(x)] = \int p(x)f(x)dx.$$

$$\mathbb{E}_{\mathrm{x}}[\alpha f(x) + \beta g(x)] = \alpha \mathbb{E}_{\mathrm{x}}[f(x)] + \beta \mathbb{E}_{\mathrm{x}}[g(x)],$$

Variance & Covariance

- Variance

$$\mathrm{Var}(f(x)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right].$$

$$= E[f(x)^2] - E[f(x)]^2$$

- Standard deviation: square root of the variance
- Covariance

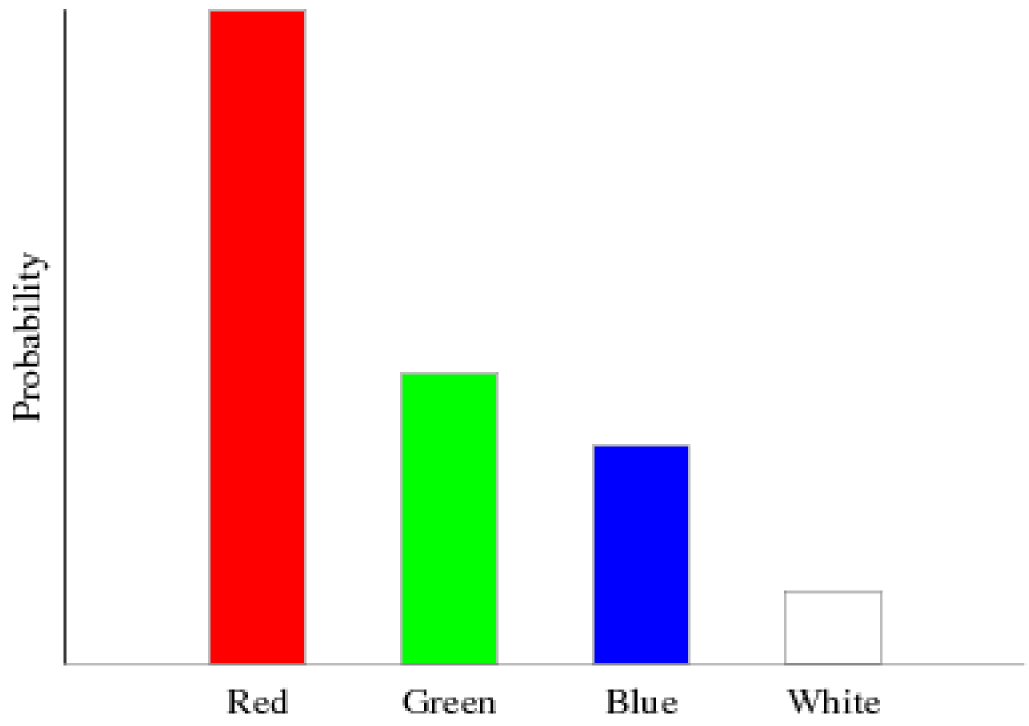
$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])(g(y) - \mathbb{E}[g(y)])\right].$$

Bernoulli Distribution

- PMF: $P(X = x) = \begin{cases} \phi & if\ x = 1 \\ 1 - \phi & if\ x = 0 \end{cases}$
- Expectation: $E[X] = \phi$
- Variance: $Var[X] = \phi(1 - \phi)$

Categorical Distribution

- A.k.a. multinoulli distribution



$$P(X = k) = p_k$$

Gaussian Distribution

- A.k.a Normal distribution
- Parameterized by variance σ^2:

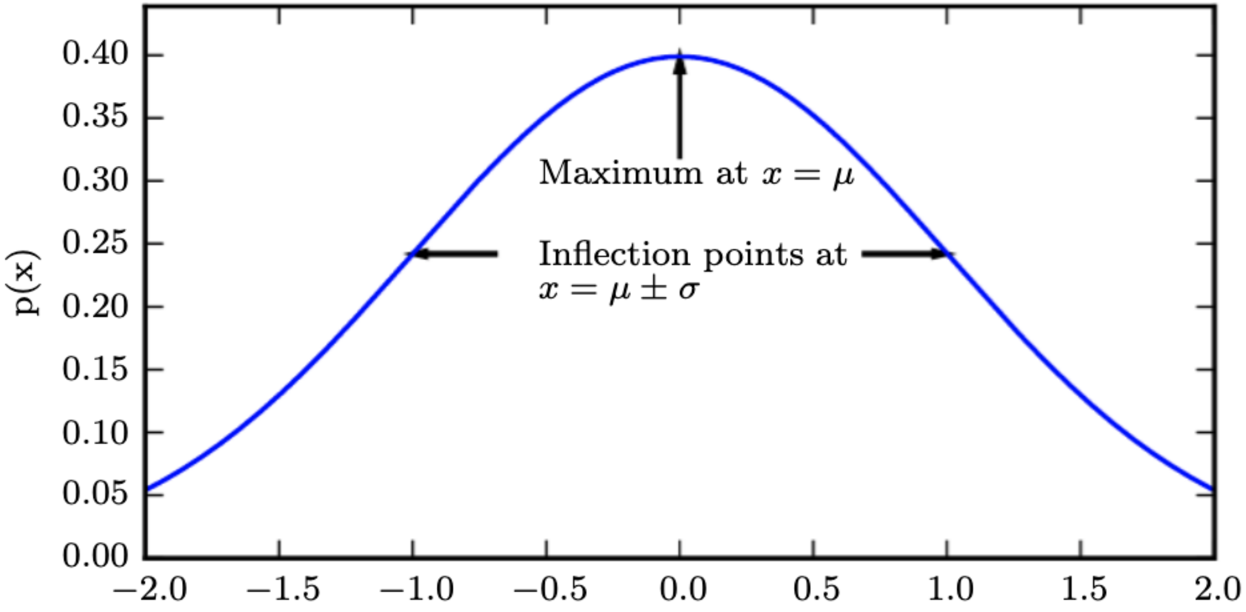
$$\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right).$$

- Parameterized by precision $\beta = \frac{1}{\sigma^2}$

$$\mathcal{N}(x;\mu,\beta^{-1}) = \sqrt{\frac{\beta}{2\pi}} \exp\left(-\frac{1}{2}\beta(x-\mu)^2\right).$$

Standard Normal Distribution

- Normal (gaussian) distribution $\mathcal{N}(x; \mu, \sigma^2)$ with $\mu = 0$ and $\sigma = 1$

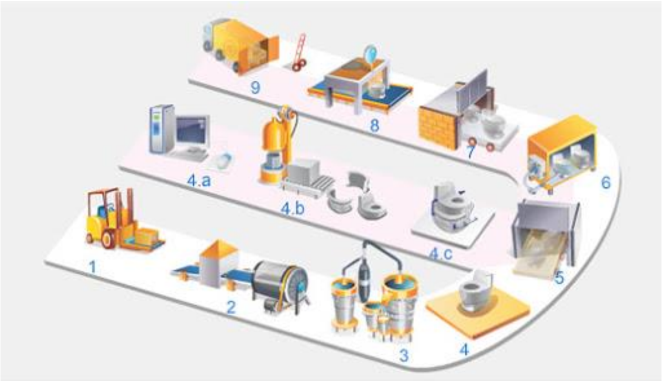


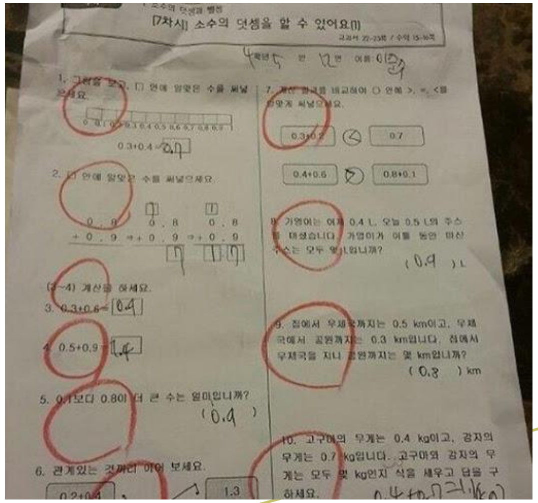
Gaussian is Indeed Normal!

- Gaussian distributions are sensible choice for many application
  - Especially in the absence of prior knowledge

Central limit theorem

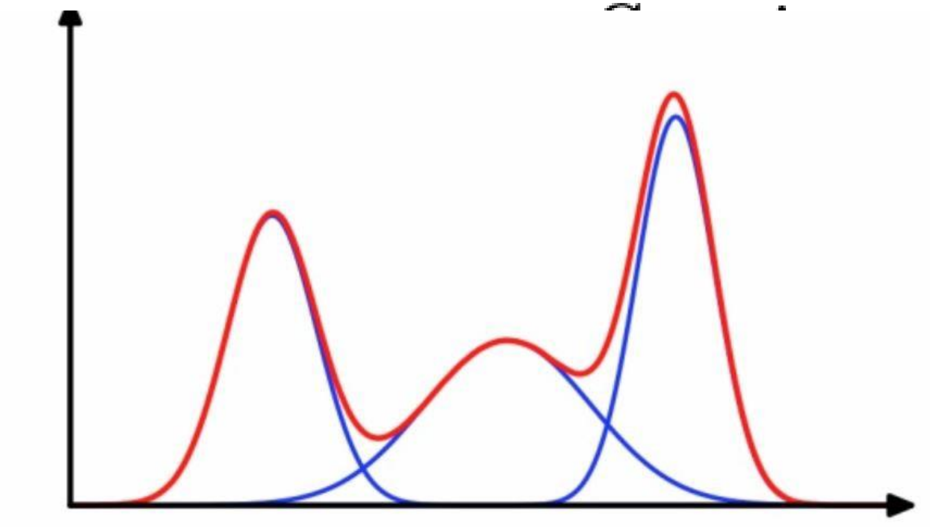
"The sum of many independent random variables is approximately normally distributed"





Mixture of Distributions

$$P(\mathrm{x}) = \sum_i P(\mathrm{c} = i) P(\mathrm{x} \mid \mathrm{c} = i)$$



Gaussian mixture

Bayes' Rule

$$\overbrace{P(\mathrm{x} \mid \mathrm{y})}^{\text{Posterior}} = \frac{\overbrace{P(\mathrm{x})}^{\text{Prior}}\overbrace{P(\mathrm{y} \mid \mathrm{x})}^{\text{Likelihood}}}{P(\mathrm{y})}.$$

Learning Proce

A lot of vector/matrix operations

Input

Model

Architecture
+
Parameters

Output

$P(Cat) = 0.9$

Label

$P(Cat) = 1$

Compute loss

Update parameters to minimize the loss

Q1: How to output probability distributions (from the results of vector/matrix operations)?

Q2: How to measure the distance between two probability distributions?

Q1: How to output probability distributions?

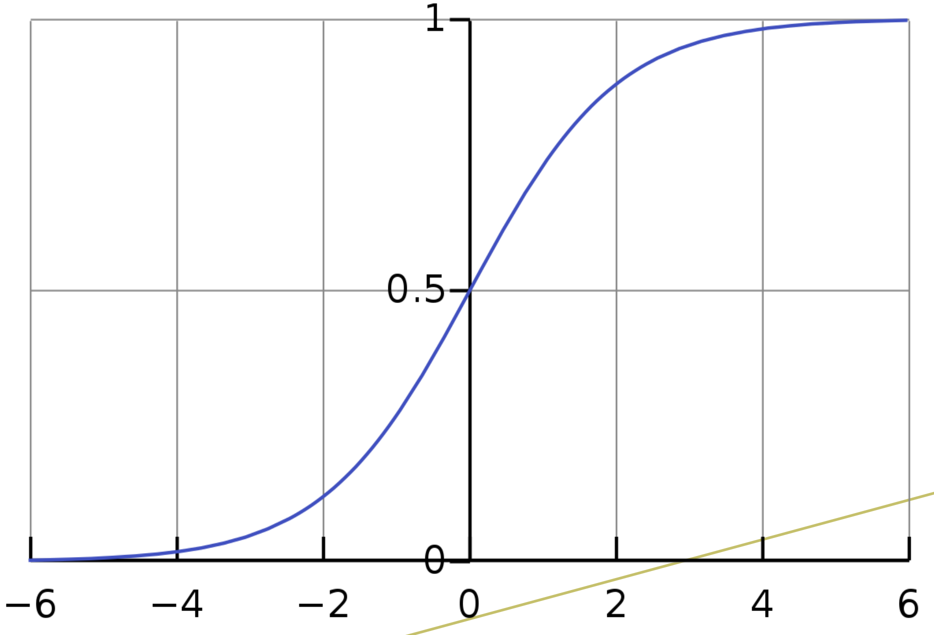
- Range of vector/matrix operations: $(-\infty, +\infty)$
- Range of $P(X)$ in Bernoulli distribution [0,1]

- Logistic sigmoid

It's not standard deviation

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

  - Domain: $-\infty \sim +\infty$
  - Range: 0~1



Q1: How to output probability distributions?

- Softplus function

$$\zeta(x) = \log\left(1 + \exp(x)\right).$$

  - Softened version of

$$x^+ = \max(0, x).$$

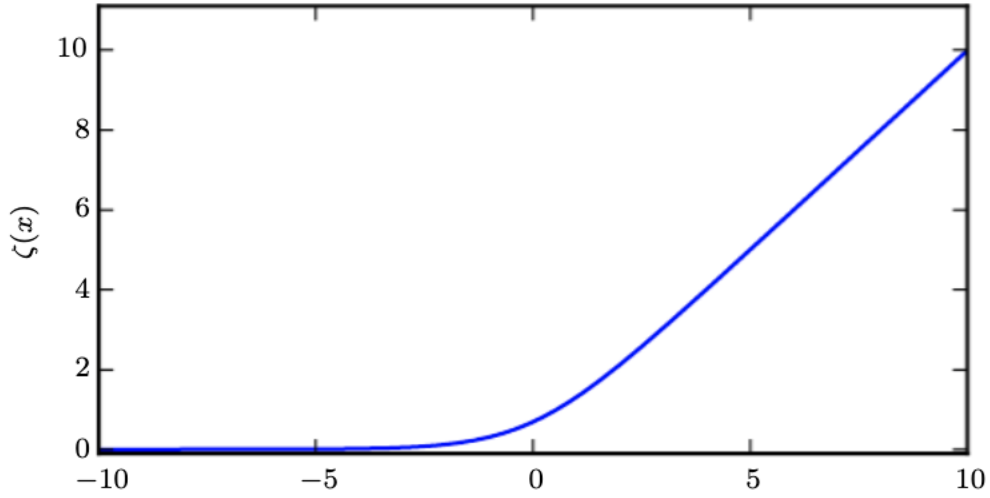


Figure 3.4: The softplus function.

  - Can be used to represent standard deviations

Q2: How to Measure the Distance Between Two Probability Distributions?

A: Information Theory

INFORMATION THEORY (OPTIONAL)

Information

- Quantity of information

1000 bits	Same quantity?	1000 bits
0000000…000000000	⟺	0010001…111001001
	Same quantity?	
0 * 1000	⟺	0*2,1*1,0*3…1*3,0*2,1*1,0*2,1*1

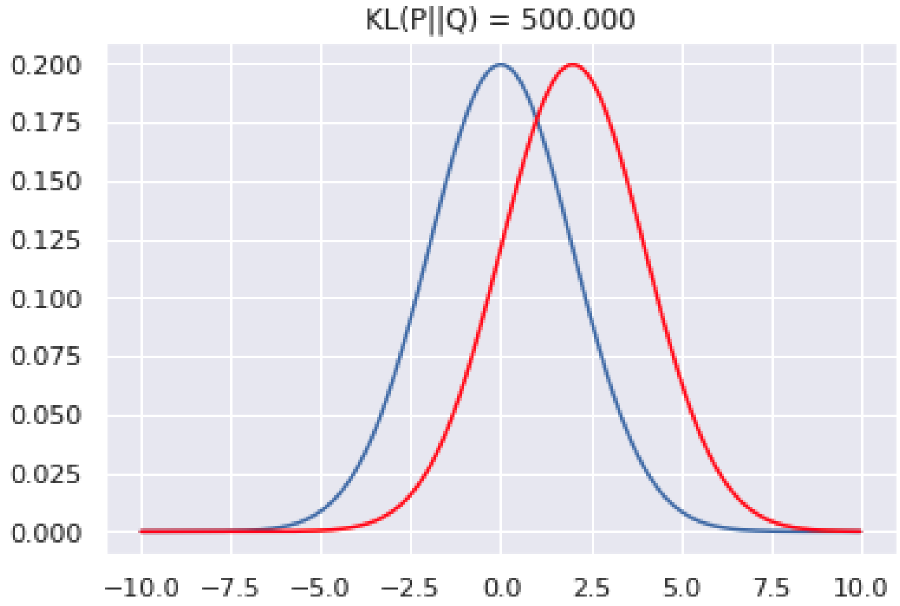
(Self) Information $I(x)$

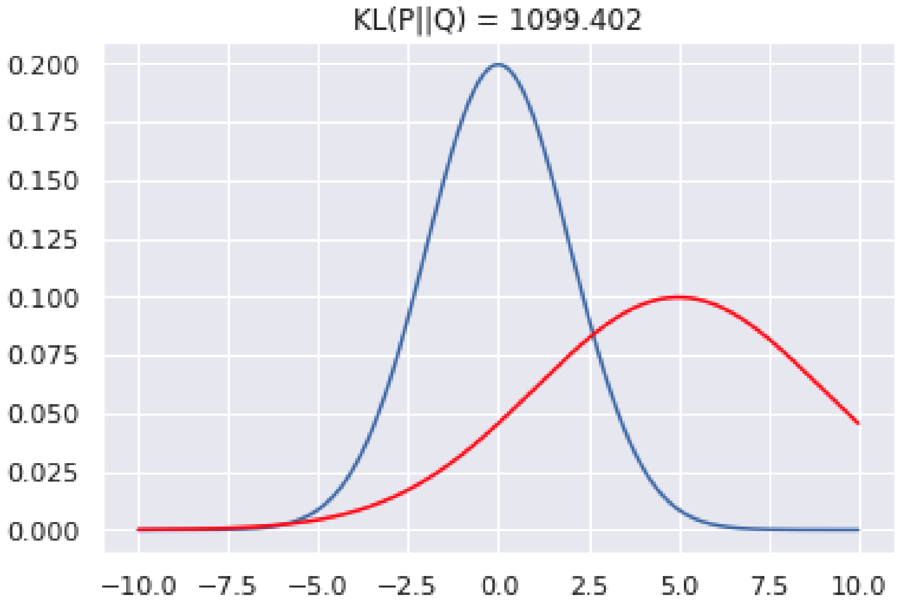
- Roughly speaking, the minimum number of bits to encode a signal x
- Definition
  - $I(x) = -\log P(x)$
- Intuition
  - If a pattern is frequent, it can be simply and efficiently encoded/compressed
  - Example: 0000000…000000000

Information Theory

- Self information of an event $\mathrm{x} = x$
  - $I(x) = -\log P(x)$
- Entropy
  - $H(X) = E_{X\sim P}[I(x)] = -E_{X\sim P}[\log P(x)]$
  - Computation: $H(X) = -\sum_x P(x) \log P(x)$
- Kullback-Leibler(KL) divergence
  - $D_{KL}(P||Q) = E_{X\sim P}\left[\log \frac{P(x)}{Q(x)}\right]$
    - P,Q: probability distributions
  - Computation: $D_{KL}(P||Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$

Kullback-Leibler(KL) Divergence





Appendix: Why L2 Norm?

- Likelihood: P(Observation|Model)
- Log-likelihood: logP(Observation|Model)
- Pdf of gaussian distribution f $\mathcal{N}(x;\mu,\sigma^2) = \sqrt{\frac{1}{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$

  - $\log P(X) = -\frac{1}{2\sigma^2}(x-\mu)^2 + C$

Minimizing L2 distance $\approx$ Maximizing logP(X) where X~N