SVD & Information Theory

Fartash Faghri

University of Toronto

CSC2515, Fall 2019

Eigen-values & Eigen-vectors

Vectors v and scalars lambda that s.t. $$\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$$

Col space, $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$

Row space, $\boldsymbol{u}^\top\boldsymbol{A} = \lambda\boldsymbol{u}^\top$

What are the e-values and e-vectors of A, B:

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$$

Singular Value Decomposition (SVD)

Any real matrix $\boldsymbol{A} \in \mathbb{R}^{n \times m}$ can be decomposed as

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$$

where $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{V} \in \mathbb{R}^{m \times m}$ are orthonormal

and $\boldsymbol{S} \in \mathbb{R}^{n \times m}$ is diagonal.

$$A = USV^\top$$

$U \in \mathbb{R}^{n \times n}$ e-vectors of $AA^\top = US^2U^\top$

$$= USV^\top (USV^\top)^\top = USV^\top VSU^\top$$

$V \in \mathbb{R}^{m \times m}$ e-vectors of $A^\top A = VS^2V^\top$

$S \in \mathbb{R}^{n \times m}$ Singular-values (non-negative)

*Eigen-value decomposition and SVD are not the same.
*Even if eigen-value decomposition is defined, eigen-values and singular-values are not generally the same.

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$$

What is the SVD of A, B?

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$$

Positive Semi-Definite (PSD)

$$\boldsymbol{A}\boldsymbol{v} = \lambda \boldsymbol{v}$$

Definition: $\forall \boldsymbol{x} \in \mathbb{R}^n, \quad \boldsymbol{x}\boldsymbol{A}\boldsymbol{x}^\top \geq 0$

Also means, all e-values are non-negative: $\lambda \geq 0$

E.g. in the normal distribution, the probability is non-negative:

(1D)

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

(nD)

$$\frac{1}{(2\pi)^{\frac{k}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})}$$

$$A = USV^\top$$

Are these PSD?

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad B = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$$

Matrix Inverse

$$\boldsymbol{A}\boldsymbol{A}^{-1} = \mathbb{I}_{n \times n}$$

If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is full rank, $\boldsymbol{A}^{-1} \in \mathbb{R}^{n \times n}$

If not full rank, inverse doesn’t exist (pseudoinverse)

If not square, inverse is not defined:

$\boldsymbol{A}^{-1}\boldsymbol{A} = \mathbb{I}_{m \times m}$ or $\boldsymbol{A}\boldsymbol{A}^{-1} = \mathbb{I}_{n \times n}$?

Matrix inverse using SVD $\boldsymbol{A}\boldsymbol{A}^{-1} = \mathbb{I}_{n \times n}$ $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\top}$

If $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is full rank $\boldsymbol{A}^{-1} = \boldsymbol{U}\boldsymbol{S}^{-1}\boldsymbol{U}^{-1}$

If not full rank, $\boldsymbol{A}^{+} = \boldsymbol{V}\boldsymbol{S}^{+}\boldsymbol{U}^{\top}$

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

Moore-Penrose inverse is the most common pseudoinverse.

Common in practice: $\boldsymbol{A}^{+} = \boldsymbol{V}(\boldsymbol{S} + \lambda\mathbb{I})^{-1}\boldsymbol{U}^{\top}$ (not MP-inv)

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$$

What is the inverse?

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \qquad \boldsymbol{B} = \begin{bmatrix} -1 & 0 \\ 0 & 0 \end{bmatrix}$$

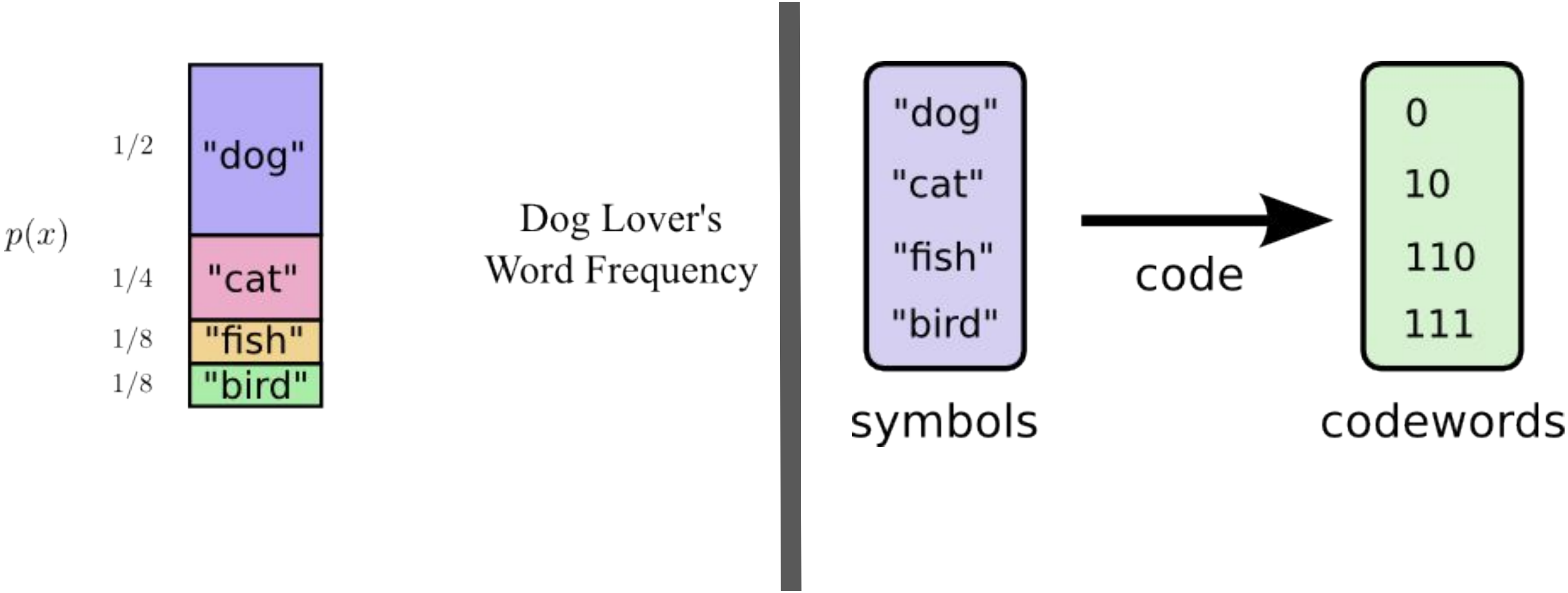
Information theory

You will see it in

- Entropy and Information gain in Decision Trees,
- KL divergence to measure distances between probability distributions,
- cross-entropy loss that is widespread in training classifiers.

Coding

What is the minimum length of code for communicating the messages of a dog lover?



https://colah.github.io/posts/2015-09-Visual-Information/

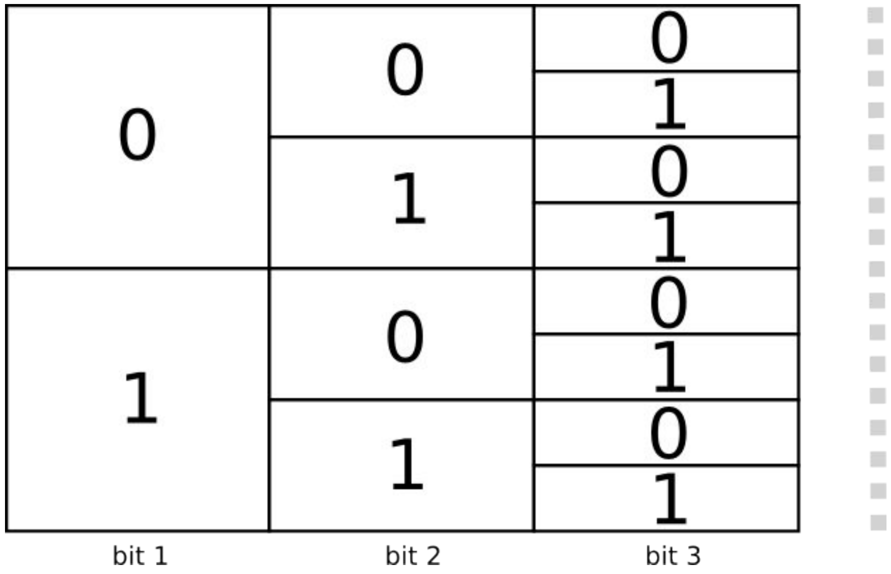
The space of codewords

2 codes of length 1 (0, 1)

4 codes of length 2 (00, 01, 10, 01)

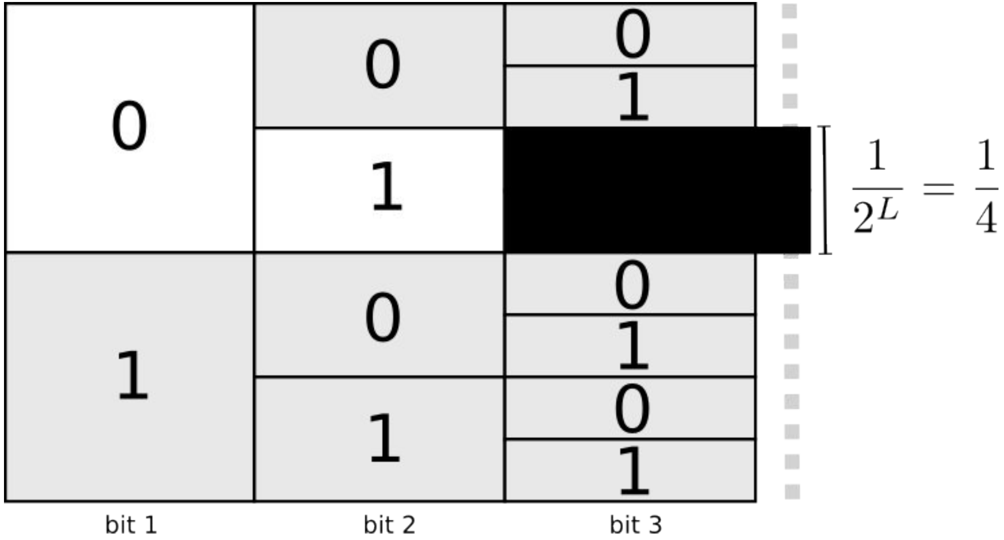
8 codes of length 3 (000, 001, ...)

….



Cost of 01 is the cost of all codes that cannot be used (black)

Cost of length 0 is 1



Greedy decoding:

0110111

Using both codes at the same time is ambiguous:

01 (dog) and 011 (cat)

Optimal Cost

The optimal cost for an event that happens with probability $p(x)$

is $p(x)$ of our total budget.

Pay more for frequently events and less for rare events.

Entropy

Cost of a message of length $L : \frac{1}{2^L}$

Invert to get the length of a message that costs $C : \log_2(\frac{1}{C})$

Since we spend $p(x)$ on the codeword for x, it has length $\log_2(\frac{1}{p(x)})$

Entropy of a distribution: the average length of the best possible code

$$H(p) = \sum_x p(x) \log_2(\frac{1}{p(x)})$$

Entropy is measured in bits (log base-2) or nats (log base-e)

Entropy

$$H(p) = \sum_x p(x) \log_2(\frac{1}{p(x)})$$

What is the entropy of each distribution?

- Bernoulli(0)
- Bernoulli(1)
- Bernoulli(0.5)

Information Gain (used in Decision Trees)

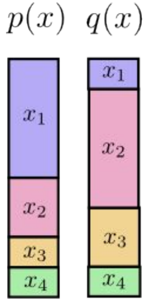
Conditional Entropy: $$H(X|Y) = \sum_{x,y} p(x, y) \log \left(\frac{1}{p(x|y)} \right)$$

How much information I gain by observing X after I had observed Y:

Information Gain: $$IG(X, Y) = H(X) - H(X|Y)$$

Cross-entropy (used in Classification)

The average length of communicating an event from one distribution with the optimal code for another distribution



Cross-Entropy: $H_p(q)$

Average Length
of message from q(x)
using code for p(x).

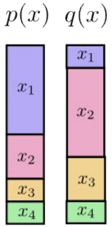
Entropy: $$H(p) = \sum_x p(x) \log\left(\frac{1}{p(x)}\right)$$

Cross-Entropy: $$H_p(q) = \sum_x q(x) \log\left(\frac{1}{p(x)}\right)$$

Classification

$$H_p(q) = \sum_x q(x) \log\left(\frac{1}{p(x)}\right)$$

Input data, ground-truth target, $(\boldsymbol{x}_i, t_i)$ (a single data point i)



Cross-Entropy: $H_p(q)$

Average Length
of message from q(x)
using code for p(x).

Prediction, $y_i \in \{1, \cdots, C\}$ (categorical random variable)

Probabilistic classifier, $p(y_i | \boldsymbol{x}_i; \boldsymbol{\theta})$

Ground-truth distribution, $q(y_i|\boldsymbol{x}_i) = \begin{cases} 1 & y_i = t_i \\ 0 & y_i \neq t_i \end{cases}$

Cross-entropy: average length of the ground truth ground-truth using the optimal

code for p: $$H_p(q) = -\log\left(p(t_i|\boldsymbol{x}_i; \boldsymbol{\theta})\right)$$

Kullback–Leibler (KL) divergence

Distance between two probability distributions

$$D_q(p) = H_q(p) - H(p)$$

$$D_q(p) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right)$$

How much longer our messages are (from p) because we used a code optimized for a different distribution (q). If the distributions are the same, this difference will be zero.

KL Divergence in Classification

Probabilistic classifier, $p(y_i|\boldsymbol{x}_i;\boldsymbol{\theta})$

Ground-truth distribution, $q(y_i|\boldsymbol{x}_i) = \begin{cases} 1 & y_i = t_i \\ 0 & y_i \neq t_i \end{cases}$

Cross-entropy, $H_p(q) = \sum_x q(x) \log \left(\frac{1}{p(x)} \right)$

KL Divergence, $D_q(p) = H_q(p) - H(p)$

What is the Entropy of q?