50.007 Machine Learning

Course Project Brief

Roy Ka-Wei Lee
Assistant Professor, DAI/ISTD, SUTD

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Self-Introduction

- Call me Roy (Please avoid calling me Dr. Roy)
- Research: Social AI Studio
- Contact:
  - roy_lee@sutd.edu.sg
  - @sroylee (telegram)

Happenings in Second Half

Week	Topic	Instructor
8	Guest Lecture (I)	DSO
	Project Briefing	Roy
9	Dimension Reduction	
	Decision Tree & Random Forest	
10	Ensemble Methods & Tricks	
	Hidden Markov Models (I)	
11	Hidden Markov Models (II)	
	Project Consultation	
12	Guest Lecture (III)	Dyson
	Guest Lecture (IV)	SHIELD
13	Project Presentation	-

Tips to do well in course

1. **Be curious** - Ask questions in class!
2. **Be adventurous** - Venture beyond class material for your project! Google/YouTube for interesting materials!
3. **Be practical** - Think about the concepts and write codes to test it out!
4. **Be collaborative** - Discuss your ideas and doubts with your peers and me!
    a. I have a lot of respect for students who share and help peers.
    b. PS: I value non-anonymous engagement because I usually remember the student who asked me questions or illuminated my mind (easier to write recommendation letter).

Project Overview

- Objectives:
  1. Consolidate what we have learned in class
  2. Hands-on opportunity - Equip with practical ML skills for your future job and internship!
  3. Practical assessment!
- Check out: https://www.kaggle.com/competitions/50007-2022/overview

Feature Engineering

- Feature engineering is the process of selecting, manipulating, and transforming raw data into features that can be used in supervised learning.
  - Represent your model's x with *features*

label y

attributes/ features

Survival	Age	Sex	Fare	Cabin	…
Yes	67	F	Normal	003	…
No	32	M	Premium	021	…
…	…	…	…	…	…

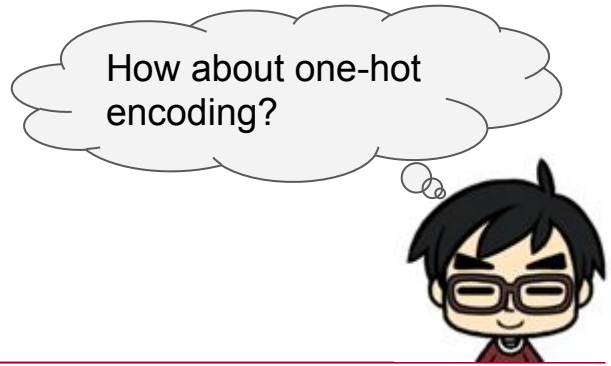
record/ instance x

How do we represent text as features?

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*

How do we represent text as features?

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*



One-Hot Encoding

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Step 1: Identify all unique words in corpus

baby	*the*	*do*
im	*dark*	
dancing	*oh*	
in	*shark*	

One-Hot Encoding

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Step 1: Identify all unique words in corpus
- Step 2: Represent each lyrics with a binary vector indicating if the word is present

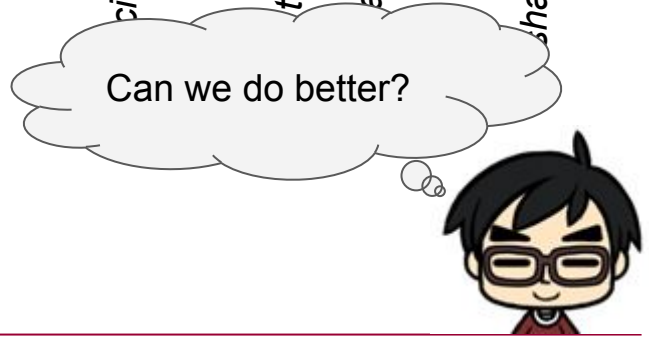
1	1	1	1	1	1	0	0	0
1	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	1	1
baby	*im*	*dancing*	*in*	*the*	*dark*	*oh*	*shark*	*do*

One-Hot Encoding

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Step 1: Identify all unique words in corpus
- Step 2: Represent each lyrics with a binary vector indicating if the word is present

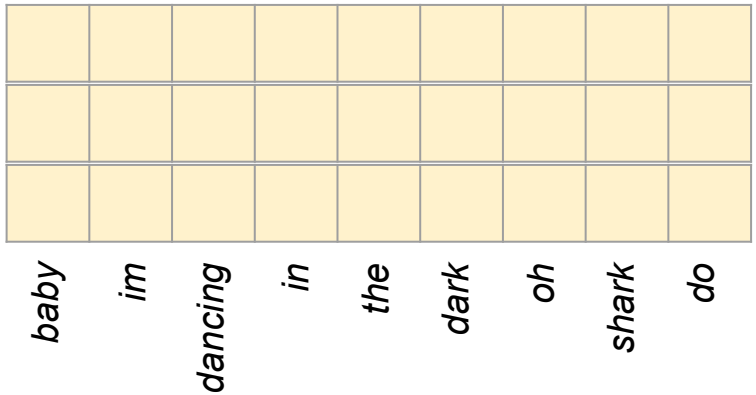
1	1	1	1	1	1	0	0	0
1	0	0	0	0	0	1	0	0
1	0	0	0	0	0	0	1	1

baby *im* *cing* *in* *the* *ark* *oh* *shark* *do*



Term Frequency

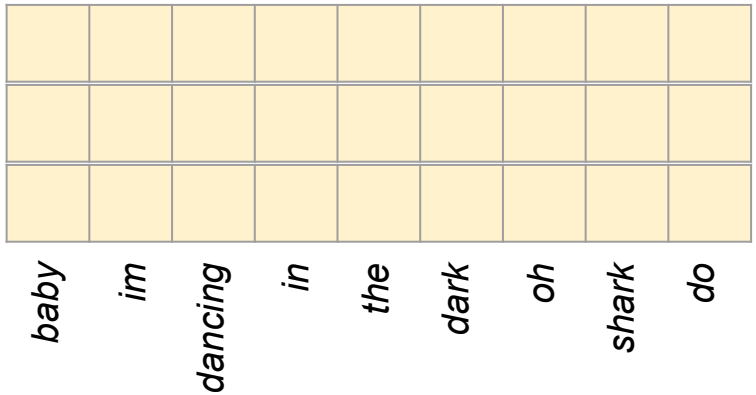
- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Term Frequency (TF) tells you how “important” is the term in the document



$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Term Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Term Frequency (TF) tells you how “important” is the term in the document



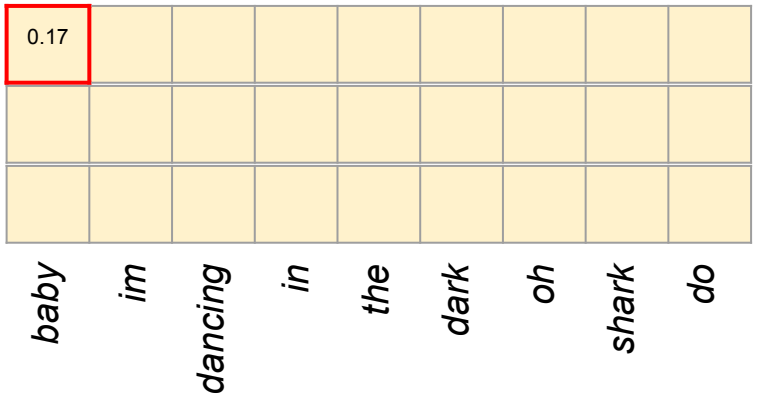
$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Number of times term t occur in document d

Total number of terms t' in document d

Term Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Term Frequency (TF) tells you how “important” is the term in the document



$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

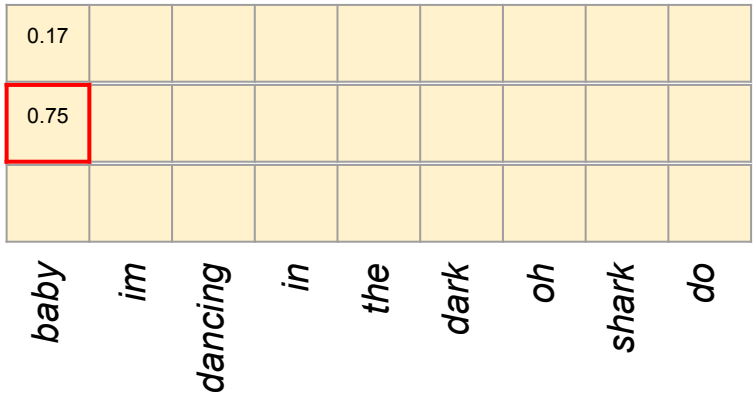
Number of times term t occur in document d

Total number of terms t' in document d

$$tf(baby, lyric1) = \frac{1}{6} = 0.17$$

Term Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Term Frequency (TF) tells you how "important" is the term in the document



$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Number of times term t occur in document d

Total number of terms t' in document d

$$tf(baby, lyric2) = \frac{3}{4} = 0.75$$

Term Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Term Frequency (TF) tells you how “important” is the term in the document

0.17	0.17	0.17	0.17	0.17	0.17	0	0	0
0.75	0	0	0	0	0	0.25	0	0
0.13	0	0	0	0	0	0	0.13	0.75
baby	*im*	*dancing*	*in*	*the*	*dark*	*oh*	*shark*	*do*

$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Number of times term t occur in document d

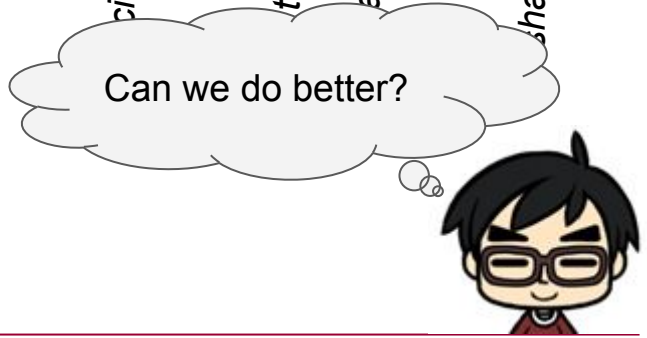
Total number of terms t' in document d

Term Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Term Frequency (TF) tells you how "important" is the term in the document

0.17	0.17	0.17	0.17	0.17	0.17	0	0	0
0.75	0	0	0	0	0	0.25	0	0
0.13	0	0	0	0	0	0	0.13	0.75

baby *im* *cing* *in* *the* *ark* *oh* *shark* *do*



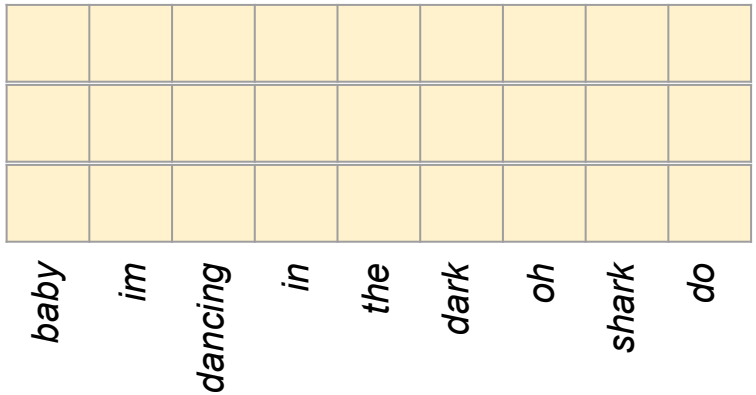
$$\mathrm{tf}(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}}$$

Number of times term t occur in document d

Total number of terms t' in document d

Inverse Document Frequency

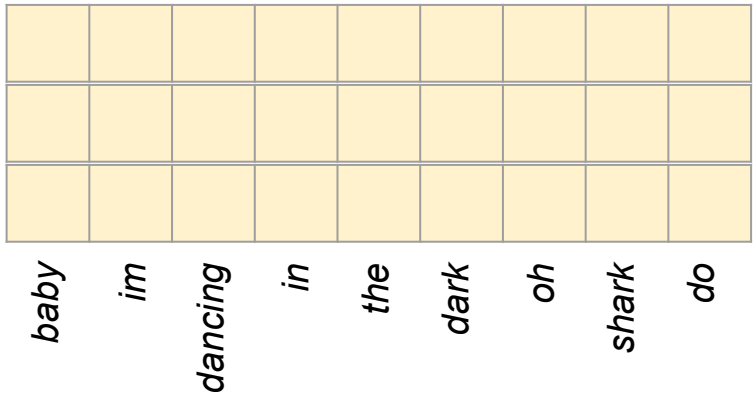
- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Inverse Document Frequency (IDF) measures how much information the word provides



$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Inverse Document Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Inverse Document Frequency (IDF) measures how much information the word provides



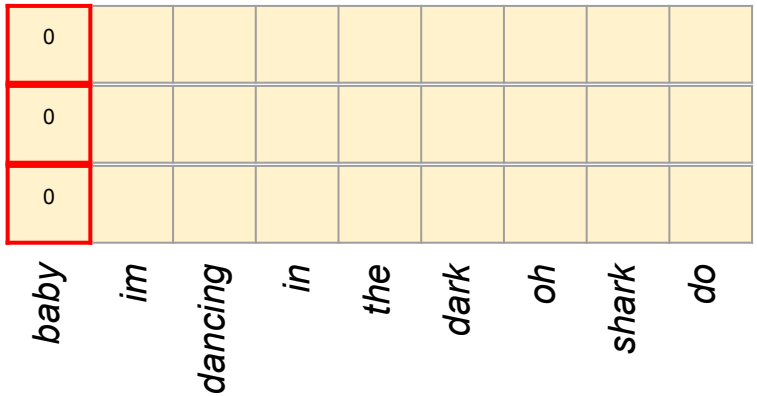
$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

N: Total number documents in dataset

$|\{d \in D : t \in d\}|$: Number of document d with term t

Inverse Document Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Inverse Document Frequency (IDF) measures how much information the word provides



$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Total number documents in dataset

Number of document d with term t

$$idf(baby, D) = log\frac{3}{3} = 0$$

Inverse Document Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Inverse Document Frequency (IDF) measures how much information the word provides

0							0.48	
0							0.48	
0							0.48	
baby	*im*	*dancing*	*in*	*the*	*dark*	*oh*	*shark*	*do*

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Total number documents in dataset

Number of document d with term t

$$idf(baby, D) = log\frac{3}{1} = 0.48$$

Inverse Document Frequency

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- Inverse Document Frequency (IDF) measures how much information the word provides

0	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
0	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
0	0.48	0.48	0.48	0.48	0.48	0.48	0.48	0.48
baby	*im*	*dancing*	*in*	*the*	*dark*	*oh*	*shark*	*do*

$$\text{idf}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Total number documents in dataset

Number of document d with term t

TF-TDF

- How do you represent the following texts as features?
  - *Lyrics 1: baby, Im dancing in the dark*
  - *Lyrics 2: baby, baby, baby, oh*
  - *Lyrics 3: baby shark, do do, do do do do*
- TF-IDF calculated as:

$$\mathrm{tfidf}(t, d, D) = \mathrm{tf}(t, d) \cdot \mathrm{idf}(t, D)$$

baby	*im*	*dancing*	*in*	*the*	*dark*	*oh*	*shark*	*do*
0	0.08	0.08	0.08	0.08	0.08	0	0	0
0	0	0	0	0	0	0.12	0	0
0	0	0	0	0	0	0	0.06	0.36

Tips to do well in course project

1. Read the instructions clearly! Clarify any doubts!
2. Read the grading metrics to know how to score well!
3. DO NOT spend too much time on the project (especially Task 3!) - There is no perfection 100% Macro-F1 score (that's overfitting).
    a. DO NOT manually label the test set - WE WILL KNOW!
4. Source out solutions beyond what was taught in class! Understand and be able to explain the solutions clearly!
5. DO NOT harass the TAs on the Blue and Red baselines! You are encouraged to ask them questions, but not the baselines.
6. Share insights with the class in "Discussion" (I will browse this often)!