COMP 4332 / RMBI 4310
Big Data Mining (Spring 2022)

Project 1: Sentiment Analysis

TA: Jiayang CHENG(jchengaj@connect.ust.hk)

Sentiment Analysis

- Generally modeled as **classification** or regression task
  - predict a binary or ordinal label

Sentiment Analysis

- Simplest task:
  - Is the attitude of this text positive or negative?

- **More complex:**
  - **Rank the attitude of this text from 1 to 5**
  - (3/5) The room was clean and everything worked fine – even the water pressure
  - (1/5) …the worst hotel I had ever stayed at ...

- Advanced:
  - Detect the target, source, or complex attitude types

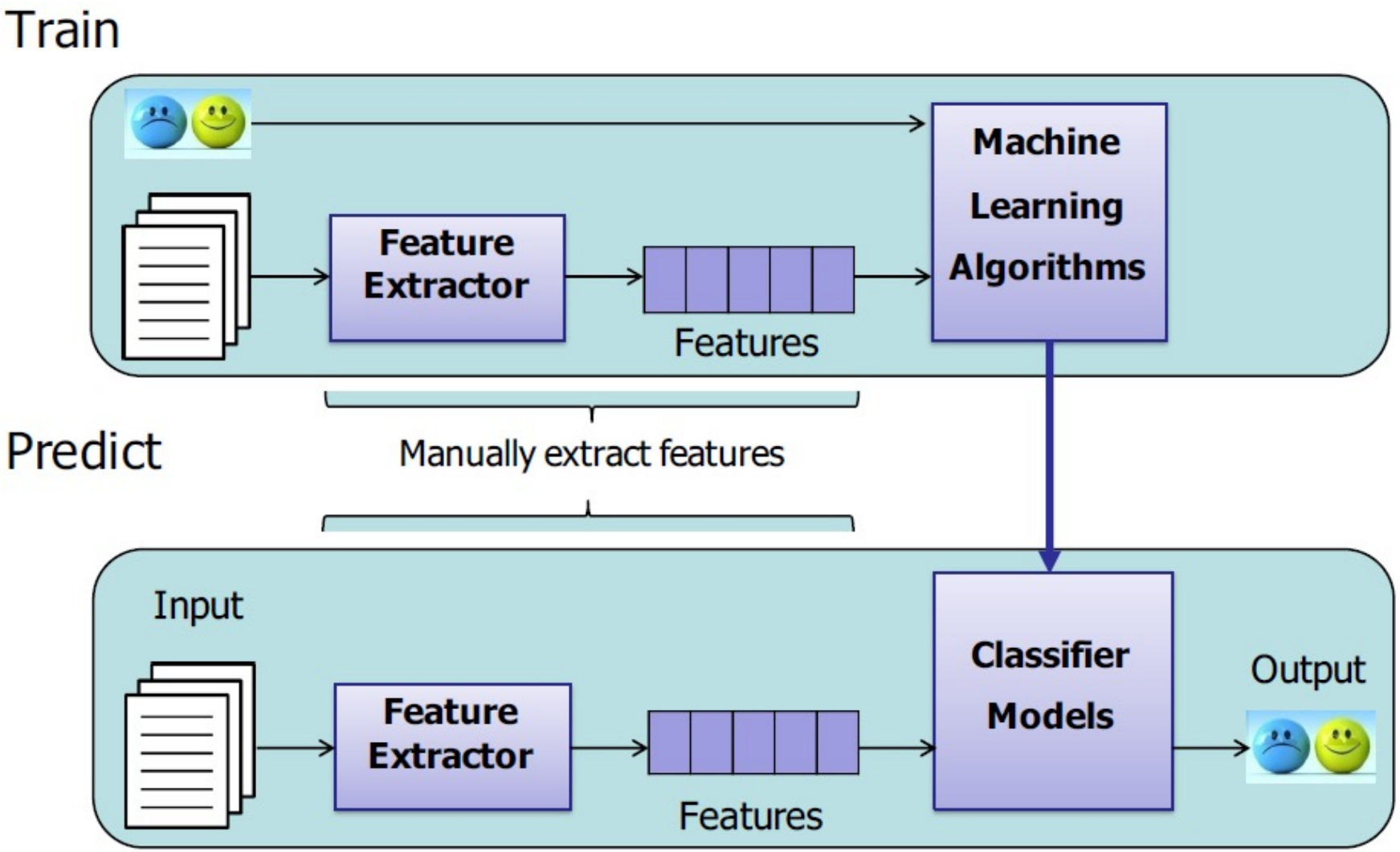
Pipeline

- Data Loader: Load data from disks
- Feature Extraction: Find useful features
- Learning: Classification via different classifiers

For more information and examples, please refer to instuction.ipynb

If you want to quickly get familiar with the whole pipeline, please refer to general_pipeline.ipynb

Pipeline


Train
Machine Learning Algorithms
Feature Extractor
Features
Manually extract features
Predict
Input
Classifier Models
Output
Feature Extractor
Features

Feature Extraction

- word occurrence, word frequency, or TF-IDF
  - This room is clean.
  - `[0,0,1,1,0,1,0,0,1,0,1]`
- word embedding
  - cbow, skip-gram, GloVe, fasttext
- `contextualized word representation`
  - ELMo, BERT, GPT, GPT-2

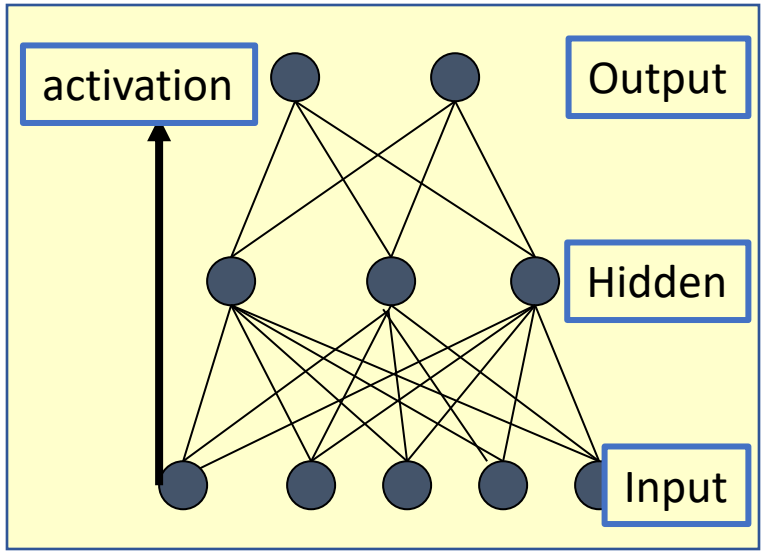
Feature Extraction

- user information
  - nationality
  - age
- date
  - weekday or weekend
  - holiday?
- hotel rating
  - Hilton Hotel
  - Youth Hostel
- data mining

Classification

- Naïve Bayes
- Logistic Regression
- Support Vector Machine
- **Deep Learning**

Multi Layer Perceptron



Demo: http://playground.tensorflow.org/

CNN

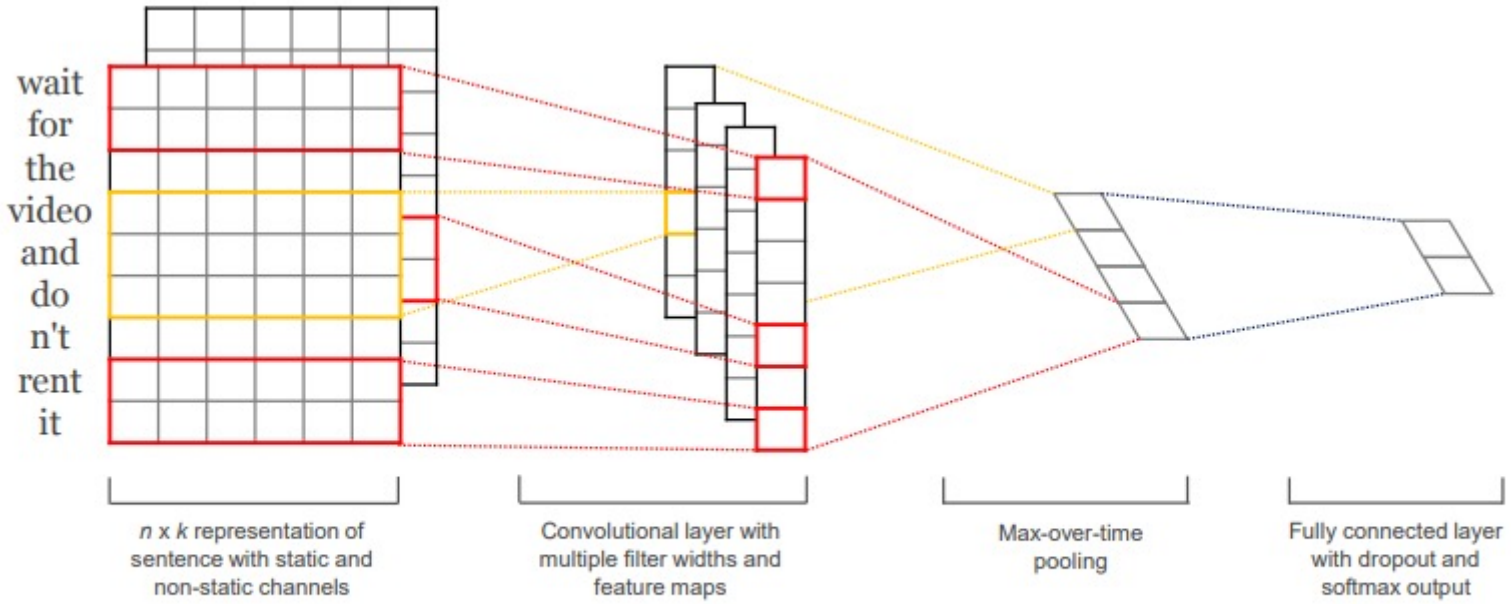
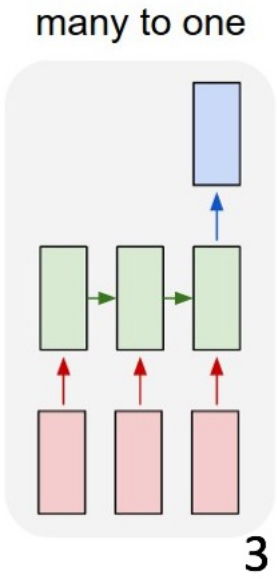


Figure 1: Model architecture with two channels for an example sentence.

RNN



Dataset

- training data: 18000 reviews
- validation data: 2000 reviews
- test data: 4000 reviews
- stars: 1.0-5.0
- given features: `business_id, cool, date, funny, review_id, text, useful, user_id`

business_id	cool	date	funny	review_id	stars	text	useful	user_id
cOJ1uIVIHCiefUyWG2wDfw	0	2016-04-19 01:53:32	0	knNl1wnZo3PdlHKr7Bd1JA	5.0	Great spot for Spark Plug shots (espresso, vod...	0	4mSdZyA7hut2s5t5WHR1mA
7m1Oa1VYV98UUuo_6i0EZg	0	2017-01-15 23:14:09	0	tX4vCH0zH79mqGONyhYziA	5.0	One of the most delicious burgers I've had in ...	0	j3t_Qv2SF1dRsYRVTnpZOQ
ZxUiFFSkxUPVQFx5iNnFrA	0	2011-04-08 06:11:45	0	k5Q5xyoIFPuIPrJlHzV4Kw	4.0	Great place for all your tobacco needs. Frien...	0	6wnuqs_HlS7rFAtxojH1wQ
f12Zv1B9crmSW58iyTR_mA	0	2015-06-15 21:04:20	0	HjqAN_SMiPPcHdaE2jcoeQ	5.0	We love the original Midwood location, so we w...	0	DronQMOA01-KIrX3UzqJFA
UIU7tug_Y-qVv_aLt7NN4g	0	2015-05-10 00:51:44	0	skjbbRmy4FiUUJSlOmsU-A	5.0	Absolutely delicious. They don't skimp on your...	0	V9H524ayC1oMfBT7b3BlhQ

Evaluation

- Macro F1 on **test data**
  - You would not get the test labels, but you can use the provided validation set to estimate your model's performance

Important dates

Three weeks in total

- [March 19, 2022] Project start
- [March 24, 2022] TA will release the validation performance of a weak baseline
- [March 31, 2022] TA will release the validation performance of a strong baseline
- [April 07, 2022, 23: 59] Submission deadline

Submission

- Predictions on **test data** (before submitting your test predictions, please make sure you can successfully evaluate your validation predictions on the validation data with the help of evaluate.py)
- Report (1~2 pages)
- Code (Frameworks and even programming languages are not restricted.)
- DDL: April 07, 2022
- Submission: Each **team leader** is required to submit the groupNo.zip file that contains pred.csv, the report, and your team's code on canvas.
- We will check your report with your code and the model performance (in terms of macro F1) on the test set.

Grading Rule

Grade	Classifier (80%)	Report (20%)
50%	example code in tutorials or in Project 1 without any modification	submission
60%	an easy baseline that most students can outperform	algorithm you used
80%	a competitive baseline that about half students can surpass	detailed explanation
90%	a very competitive baseline without any special mechanism	detailed explanation and analysis, such as explorative data analysis, hyperparameters and ablation studies
100%	a very competitive baseline with at least one mechanism	excellent ideas, detailed explanation and solid analysis

Thank You and Good Luck