11/21 HW2 Presentation

Logistics

- We will start 1hr earlier on 11/21: 13:20-17:20
  - Please let TA know if you cannot do 13:20~14:20
- Each group has 8 minutes (carefully timed)
- Ceiba submission (update, very important)
  - Slides submission deadline changes from 11/21 14:00 to 11/21 13:00
  - You **don’t need to submit any code**
  - You cannot modify the presentation file after 13:00
  - Only group leader needs to submit (don’t submit multiple copies)
-

Submission Format

- R08922xxx.zip/
  - R08922xxx_slides.pdf, or .ppt
    - Necessary for you presentation
  - R08922xxx_report.pdf (Optional)
    - Written report
    - If you think 8 minutes is not enough to describe all your findings (please clearly state which parts were not presented)
- Incorrect format will lead to penalty

Grading

- Your grade will be mainly based on the average of the grades from 2TAs and the instructor
- Please carefully manage your time
    - Penalty will be imposed if you are overtime, except that your materials are too good for us to miss.
- If you have ground-breaking findings, we will help you to publish the results

Final Project Description

Special Directions on Machine Learning (SDML)
Fall 2019, NTU CSIE
Prof. Shou-De Lin

Background

- Your team can choose one among the 4 projects we proposed
  - Please submit your preference here before 11/20 23:59 https://docs.google.com/forms/d/e/1FAIpQLSf08mWz6lG7CiXg_q7Cms9Z48SpZcfF8XvuwBiZvboCv7Cxbw/viewform
  - We will announce the project assignment and data on 11/21 5pm (after HW2 presentation)
- Each team will present their progress 3 times (including the final presentation)
  - Grading: 25%, 25%, 50%
  - Final report (optional): only if you have significant more stuffs to say

Follow-up Schedule

11月14日		Final Project Out (submit your preference by 11/20 23:59)
11月21日	HW2 presentation	Announce Final Project Assignments
11月28日	Lecture	
12月5日	Project Presentation	Projects 1 + 2
12月12日	Project Presentation	Projects 3 + 4
12月19日	Project Presentation	Projects 1 + 2
12月26日	Project Presentation	Projects 3 + 4
1月2日	Lecture	
1月9日	final project presentation	Starting from 1pm

Four Topics

- Team 1: URL Labelling
- Team 2: Classification with missing features in Testing Set
- Team 3: Predicting missing labels in data
- Team 4: Seq2Seq explanation

Four Topics

- Team 1: URL Labelling
- Team 2: Classification with missing features in Testing Set
- Team 3: Predicting missing labels in data
- Team 4: Seq2Seq explanation

Introduction

- It’s the data offered by Appier（沛星科技）

- Given a URL and a category, to predict whether the URL belongs to the category (binary classification)
  - P(label | url, category)
- Data:
  - Real world annotated URLs
  - Meta-data of some URLs (e.g. webpage description, title and keywords)
- Challenges:
  - the annotated URLS and the ones with Meta-data are not well-aligned
  - Incomplete training labels

URL Classifcation

- Goal: To predict whether the url belongs to a category
  - 510 categories/labels
  - **Note that a URL may have multiple labels e.g., belong to both category_1 and category_2.**
- For the task of this topic
  - Annotation is done by human.
  - Datasets
    - Dataset A: Partially annotated URL
      - For some (url, category) pairs, a label (true or false) is assigned
      - Note that a lot of (url, category) pairs are not labelled.
      - size: 9,000 (will be divided into training/testing)
    - Dataset B: Metadata
      - url, meta-description, meta-title, meta-keywords
      - size: 90,000
    - **Only partial Urls occur in both dataset A and dataset B**

Datasets

Dataset A

URL Category **Label**

	url	category	label
	http://www.99kubo.tv/vod-read-id-104339.html	c194	False
	https://www.cnyes.com	c79	True
	https://health.udn.com/health/story/5999/4070342	c216	False
	https://www.businesstoday.com.tw/article/categ...	c397	True
	https://roguelands.fandom.com/wiki/Creation_Ma...	c490	False
	https://m.juksy.com/archives/93047	c41	False
	https://www.techbang.com/posts/72015-the-new-h...	c206	False
	http://www.news18nepal.com/1332760/janeliu/nic...	c275	True
	https://travel.ettoday.net/article/1501401.htm...	c276	False
	http://blog.udn.com/frankbetty/6078389	c93	True

Dataset B

URL Description Title keywords

	url	meta-description	meta-title	meta-keywords
	http://blog.udn.com/mobile/wangtao/3923781	「啟事」和「啟示」是兩個讀音相同，但卻完全不能互通的詞。我們常在各式媒體或招貼上看見諸如「尋...	張貼「啟事」可以獲得甚麼「啟示」？	[]
	https://www.mobile01.com/topicdetail.php?f=257...	關於 美食攝影的角度這問題 沒有一定!看起來好看就好如果硬是要規類出來大概 低角度/45度/...	[經驗分享] 自然感的美食攝影 關於拍攝角度	[]
	https://www.cool3c.com/article/112723	討喜小姐發佈一圖看完 三星Note 7爆炸案始末全紀錄，留言0篇於2018-08-11 07...	一圖看完 三星Note 7爆炸案始末全紀錄 (112723)	['產業消息', '一圖看懂', '三星', '爆炸', 'Note 7', '手機', '...
	https://bimeci.pixnet.net/blog/post/219048381	特約撰文：mangowalk　校稿：黑麵　編輯：ＢＭＣ　大家好我是芒果走路探險隊的隊...	台中望高寮13號碉堡探秘 @ 想和妳看棒球 :: 痞客邦 ::	['旅行', '秘境', '生活', '文字', '分享', '科普知識', '棒球', '...
	http://blog.udn.com/mobile/1688ku/32603939	破財漏財衰運的分析　破財，就是沒來由的突然花出不應花的一筆錢，比如：突然被人騙錢，或...	1688ku 當下光明	[]

Evaluation metrics

- Binary classification accuracy

Four Topics

- Team 1: URL Labelling
- Team 2: Classification with missing features in Testing Set
- Team 3: Predicting missing labels in data
- Team 4: Seq2Seq explanation

Motivation

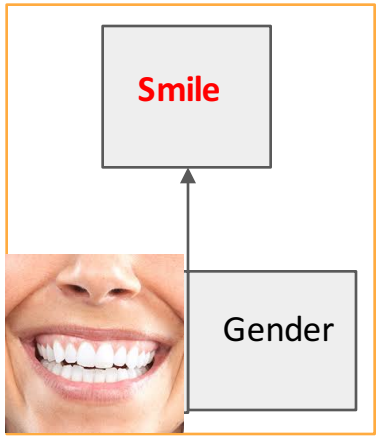
- In real-world scenario, some features are not available during the prediction time (only available in training data)
  - [Privacy] Unavailable features during test time due to privacy issue
    - For instance, we might not be able to get the demographic features (**gender** or **age)** for patients in disease classification, but those features are available during training.
  - [Fairness] Bias in training data
    - We might not want to explicitly use certain features for classification (e.g. using race as a feature for criminal prediction)

Privacy & Fairness

- https://www.zhihu.com/question/263336767
- http://dalimeeting.org/dali2017/fairness-and-privacy.html
- https://arxiv.org/pdf/1812.02696.pdf

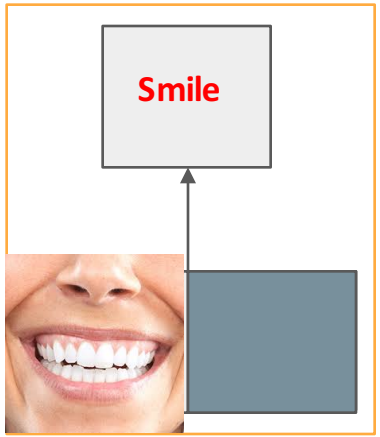
Problem definition

- Given: a set of training data for classification
- Prediction phase: some features are not available to be used



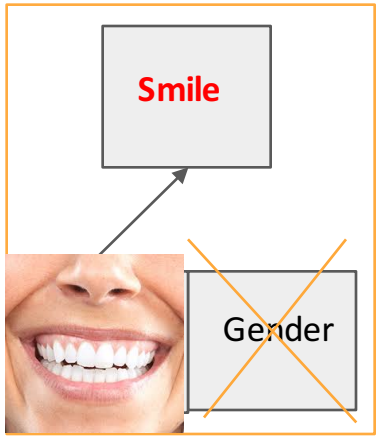
Problem definition

- Given: a set of training data for classification
- Prediction phase: some features are not available to be used



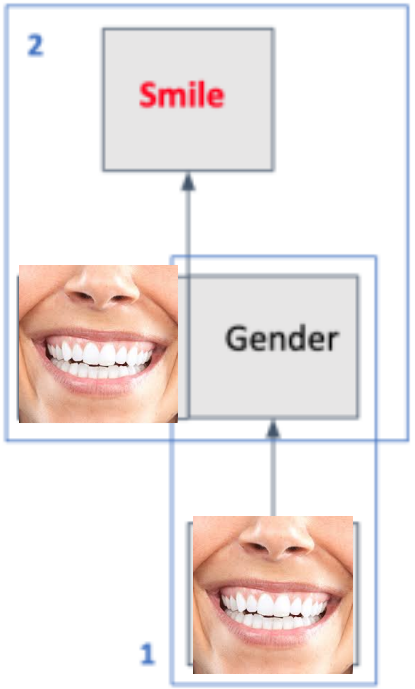
Baseline 1

- Ignore the unavailable features during training



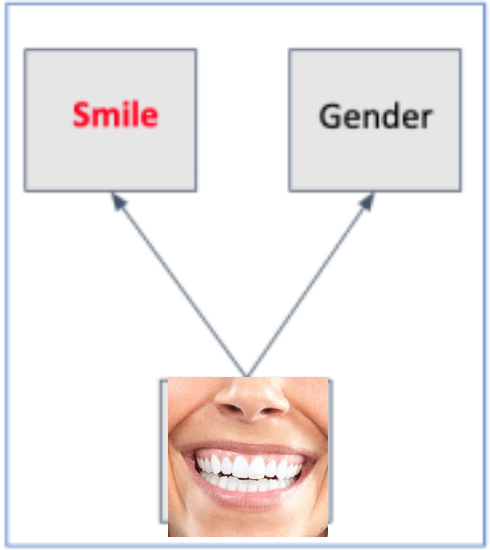
Baseline 2

- 2-stage training
  1. **Feature predictor**
  2. **Target task predictor**



Baseline 3: Multi-task Training

- Treat given input features as **auxiliary** labels and learn the target task by **multi-task training**



Kaggle Competition Site will be Set

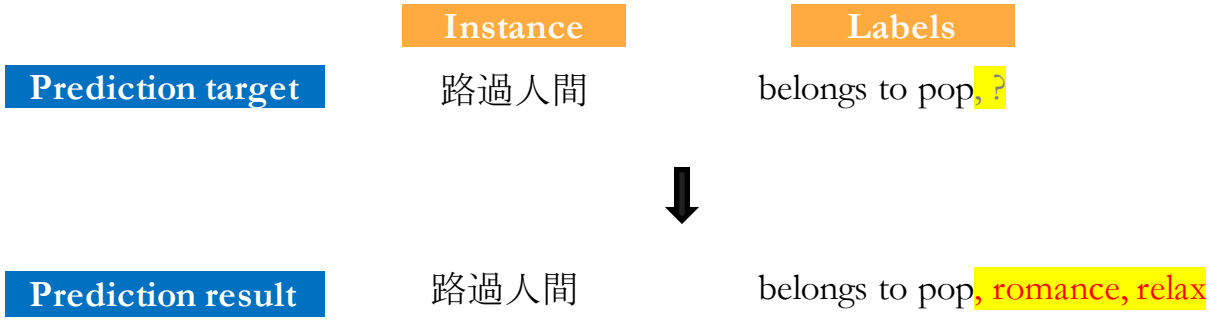
- Two scenarios:
  - Only one feature are not available
  - More features are not available
- **Your first goal is to beat those baselines !!**
- For performance, we care about 2 things:
  - The absolution performance (show in Kaggle)
  - Relative performance after considering the missing features (show in your presentation)

Four Topics

- Team 1: URL Labelling
- Team 2: Classification with missing features in Testing Set
- Team 3: Predicting missing labels in data
- Team 4: Seq2Seq explanation

Multi-label with Missing Labels problem

- In multi-label learning, each instance can be assigned to multiple class labels simultaneously
- Instead of assuming a complete label assignment is provided for each instance, only partial labels are assigned with instances, while the rest are missing
- The goal is to recover the full label assignment for each instance



Why is Multi-label with Missing Labels important?

- Multi-label classification is an essential problem in many application domains, including image annotation, video classification, document categorization, and gene function prediction
- However, complete training labels are hard to collect in real world problems
    - Human can miss some relevant labels (e.g. Thriller and horror)
- Label incompleteness can severely degrade the performance of the learned multi-label classification models, since it will build negative prediction patterns between the input instances and the missing labels and further propagate the mistakes into the prediction phase on the test data

Task description: the scenario of recommendation

- Given: 1. Ratings for each user-item pair 2. labels (genra) of items
- Predict: Despite the given labels of an item, which other labels should this item belong to?
- Evaluation: F1 score for all items

	Instance	Labels
Prediction target	路過人間	belongs to 流行樂, ?

Prediction result	路過人間	belongs to 流行樂, romance, relax, ~~comedy~~

Dataset description – ratings.csv

- **Ratings Data File Structure (ratings.csv)**
  - userId,itemId,rating,timestamp
  - Eg. 1,2,3.5,1112486027
- Rating range from 0.5 to 5.0
- Total 27,278 items, itemID might not be sequential
- Total 138,493 users, userID might not be sequential
- Total 20,000,263 ratings

Dataset description – items.csv

- **Item Data File Structure (items.csv)**
  - itemId, labels
  - 249, 2 3

- There are total 18 distinct labels
- You need to predict the labels for ALL items (include the existing labels)
- Two scenarios:
  - Labels missing by random
  - Labels are not missing by random (’similar labels’ have higher chance to drop)
- Before drop: 60% items have more than one genres – average: 1.99

Evaluation – Average F score for 27,278 items

- Precision
  - $P = \frac{N_{rs}}{N_s}$
- Recall
  - $R = \frac{N_{rs}}{N_r}$
- F-measure (F_1 score)
  - $F = \frac{2PR}{P+R}$

		Recommended Items		
		Selected	Not Selected	Total
Relevancy	Relevant	N_{rs}	N_{rn}	N_r
	Irrelevant	N_{is}	N_{in}	N_i
	Total	N_s	N_n	N

Prediction result output format and Kaggle

- Your prediction result file should follow the following format and upload it in Kaggle

- **Result file**
  - itemId, labels you predict(separated by ”|”)
  - 249,2 3

Kaggle Competition

- Testing data will be divided into public and private testing.
  - You will be evaluated based on the performance of **public+private** testing.
- Maximum 2 submissions a day are permitted.
  - Competition ends on 1/5 23:59.
  - Please declare one submission before the deadline as your final submission (if not, then we will use your last submission)
- Using *extra data* from the Internet is prohibited.
- Please use the SDML_<Student-ID> or SDML_<Group Name> as the Kaggle nickname to show on the leaderboard.
  - Please form a team (not more than 3 persons).
  - For example, SDML_r07922000 or SDML_Team01 as the nickname
  - Do not create multiple kaggle accounts for more submissions.
- Competition pages:
  - will be announced later

Grading Policy

- Performance (50%)
  - Except task 1.4 we will focus on the solidness of the analysis
- Others (50%)
  - *Efforts:* #methods you tried; please describe and analyze the approaches with experiment results.
  - *Clearness on the presentation*
  - *Novelty:* how novel is your model designed.
    (Ensemble techniques are valid, however we encourage novel single models.)

Final Reports (Optional)

- You can submit when there is not sufficient amount of time to present your results in the final presentation.
- Should be formatted in PDF files through CEIBA.
- The report should include:
  - What you have said in the presentation
  - What you have not yet got a chance to say in the final presentation
- Submission Deadline: Jan 6th 23:59
- No page limit. ☺
  - Feel free to include all the experiment results, reference theorems or other appendices.

Why working hard on your project?

- We have carefully choose some interesting and important topics for you
    - Any breakthrough will lead to a nice publication
    - Since anyway you will need to work on this project, why not work harder to make it meaningful
- It’s a chance for you to experience what it looks like to work on a practical ML research topic

Shou-de Lin and Craig Knoblock, "Exploiting a Search Engine to Develop More Flexible Web Agents", in Proceedings of the 2003 IEEE/WIC International Conference on Web Intelligence (WI 2003), Halifax, Canada. (Best Paper Award)
Shou-de Lin and Kevin Knight. "Discovering the linear writing order of a two-dimensional ancient hieroglyphic script", in Artificial Intelligence v.170/4-5, Elsevier, 2006. One of the Top 25 hottest papers in Artificial Intelligence in 2006.