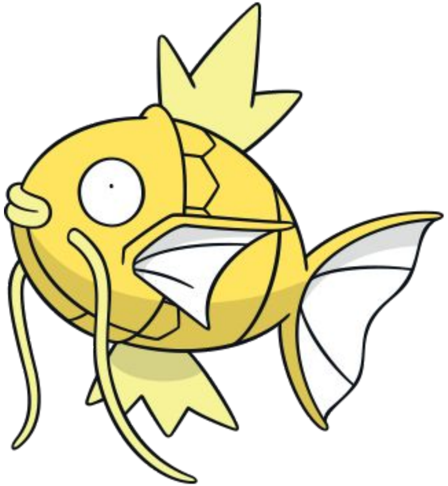
SDML Final Topic 3
鯉魚躍龍門



b05902004 陳心平
b06902020 唐浩
b06902025 黃柏瑋

Summary

- Data analysis
  - Rating (MF)
  - Item (Correlation & Supplementary matrix)
- Random
  - Label without Correlation
    - MF
    - One-hot rating
  - Label with Correlation
    - Pseudo / Soft label
    - KNN
- Rule

Data Analysis

Data Analysis

Rating $\Rightarrow$ Feature Input

Data Analysis - Rating

- 138493 users
- 26477 items
- 20,000,263 ratings
- We can regard user’s ratings as item’s feature inputs (138493-dim)
  - MF
  - only 0.5% true rating
  - Too many ambiguous ratings

Data Analysis - Rating

- Mean of rating count / variance

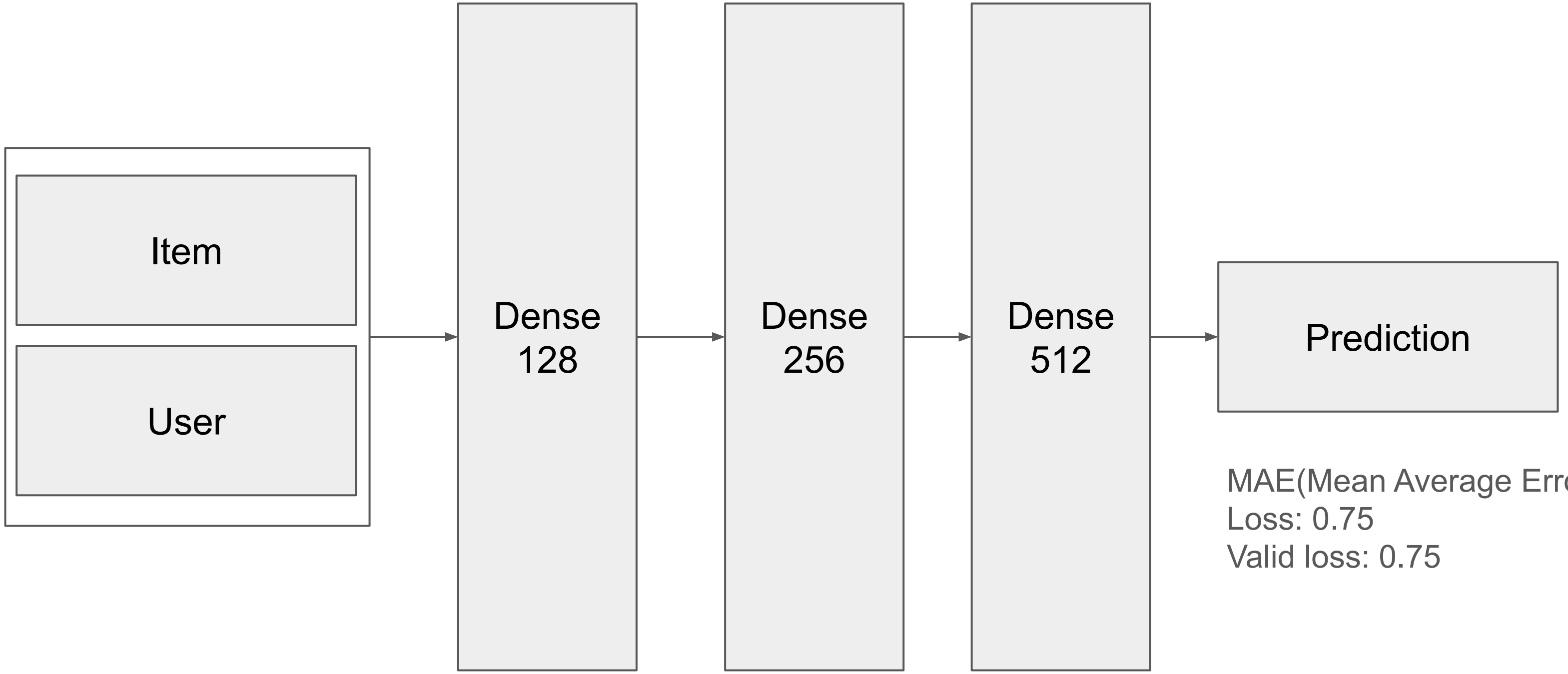
	count	variance
user	144.4135	0.9526
item	747.8411	0.9188

$\Rightarrow$ User threshold: count > 100, variance > 0.9

Data Analysis - Rating

- 27814 users
- 23130 items
- 8455726 ratings
- We can regard user's ratings as item's feature inputs (27814-dim)
    - MF
    - 1.3% true rating

Data Analysis - Rating (MF model)



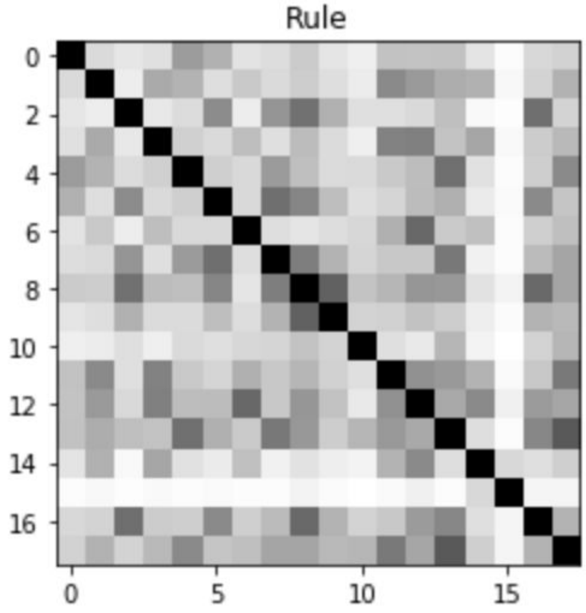
MAE(Mean Average Error)
Loss: 0.75
Valid loss: 0.75

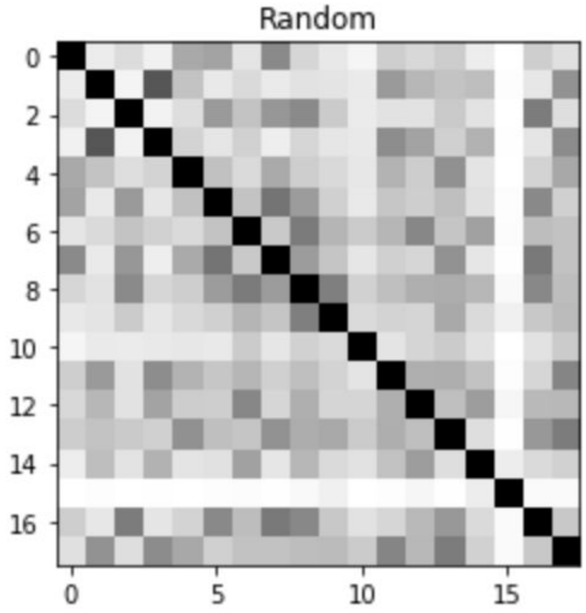
Data Analysis

Item → Feature Input / Label

Data Analysis - Item (Correlation matrix)

Cosine similarity





Data Analysis - Item (Supplementary matrix)

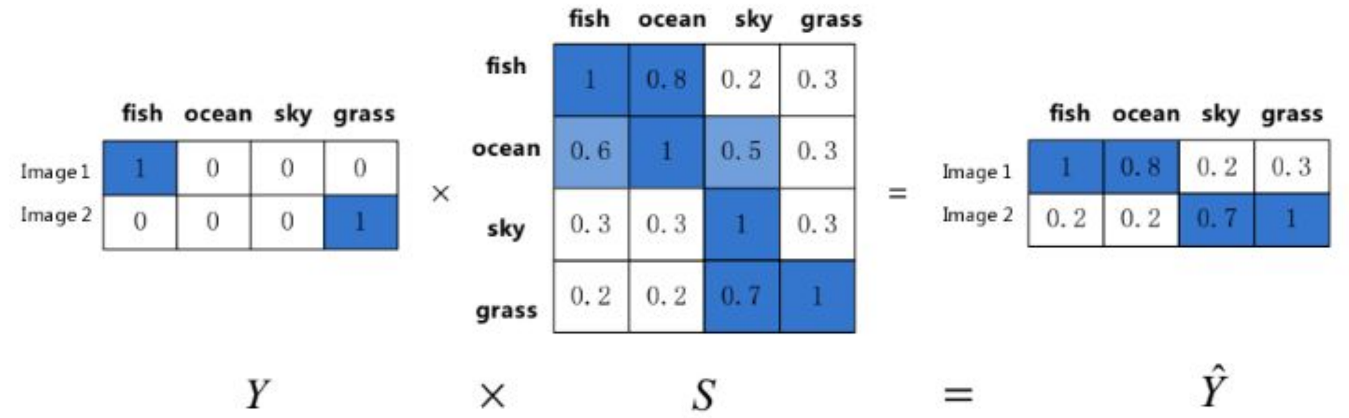


Figure 2. A supplementary label matrix $\hat{Y}$ obtained by multiplying the label correlation matrix S with the original label matrix Y

Random

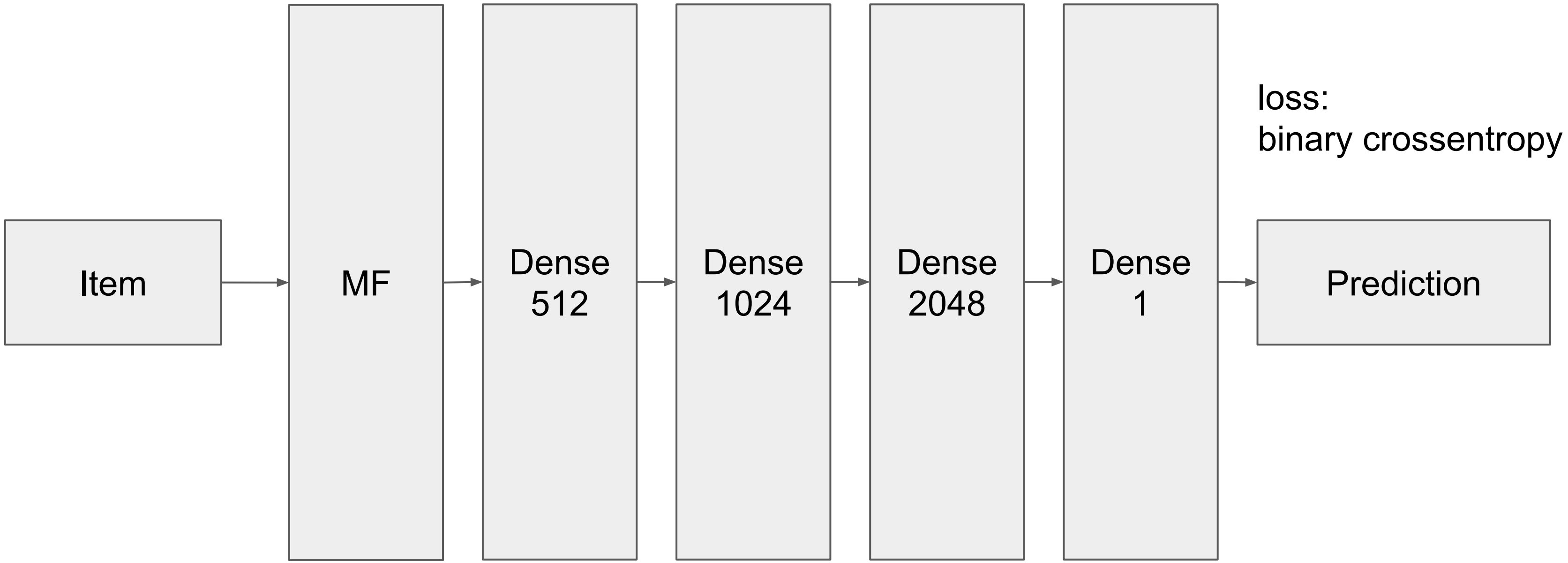
Random

Label without correlation

Assume label Independency

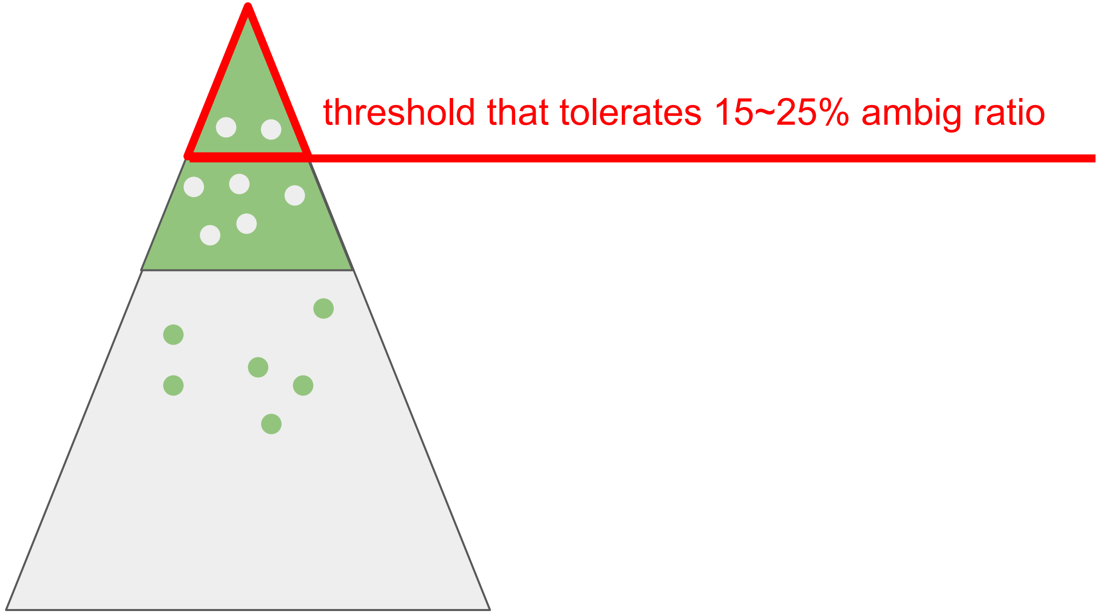
Random - MF

For each label,



Random - MF

Threshold:



Random - result

Predicted result:

	MF			
ambig ratio	0.25			
avg. label	1.6450			
F1 score	0.88153			

Random

- Question
  - Does rating value really matter?
  - No rating
    - Unseen
    - No interest

Not even an action film
Rating: None(Unseen)

Bad action film
Rating: 1.0

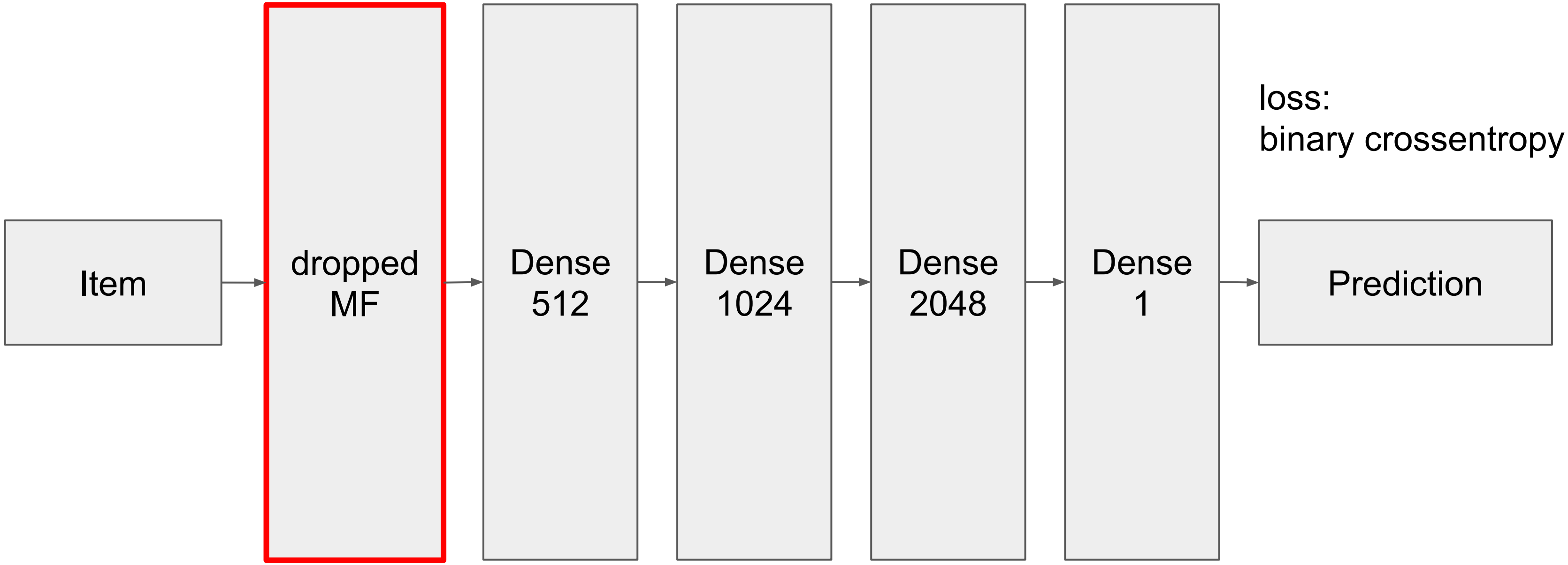
Good action film
Rating: 5.0

Random

- Solution
  - Dropped MF
    - Give a probability of 0.2 to drop non-rated ones in MF
  - Original Rating
    - Rated = rating value, Non-rated = 0
  - One-hot Rating
    - Rated = 1, Non-rated = 0

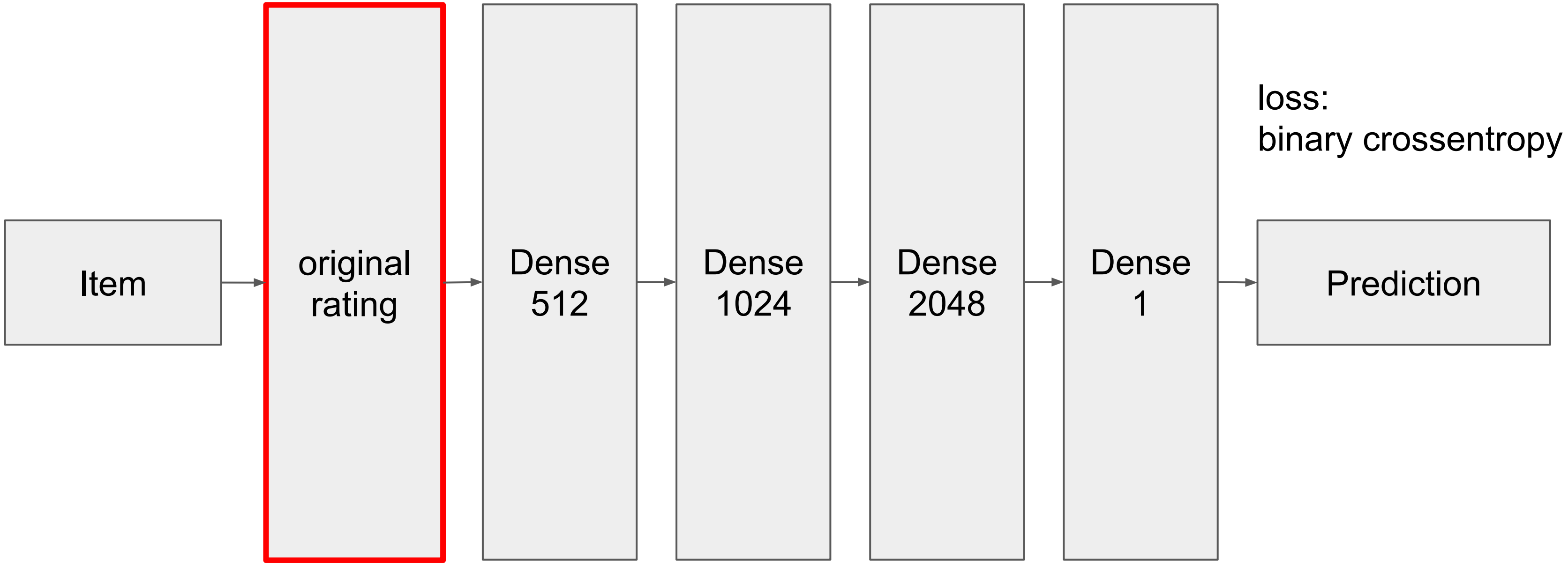
Random - Dropped MF

For each label,



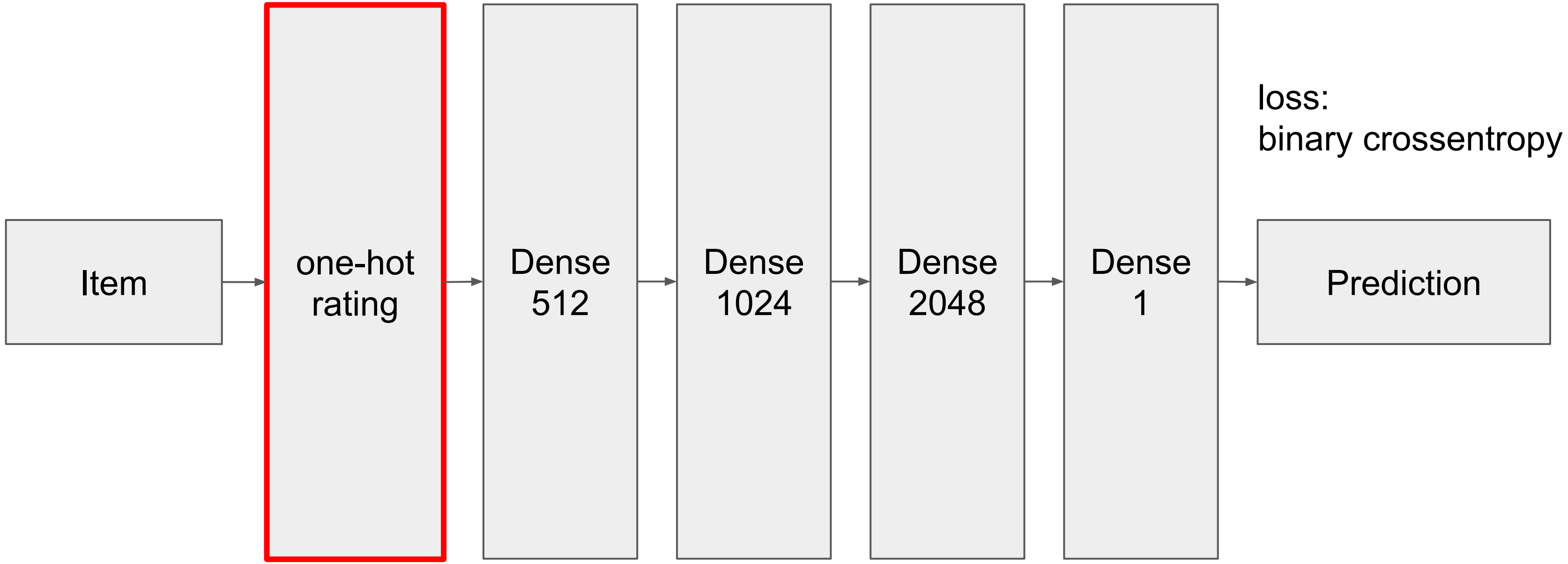
Random - Original Rating

For each label,



Random - One-hot rating

For each label,



Random - result

Predicted result:

	MF	Dropped MF	Original Rating	One-hot rating
ambig ratio	0.2	0.2	0.2	0.3
avg. label	1.6450	1.63	1.6374	1.7312
F1 score	0.88121	0.8703	0.8866	0.8903 0.8907

Random

Label with correlation

Consider supplementary matrix

Data Analysis - Item (Supplementary matrix)

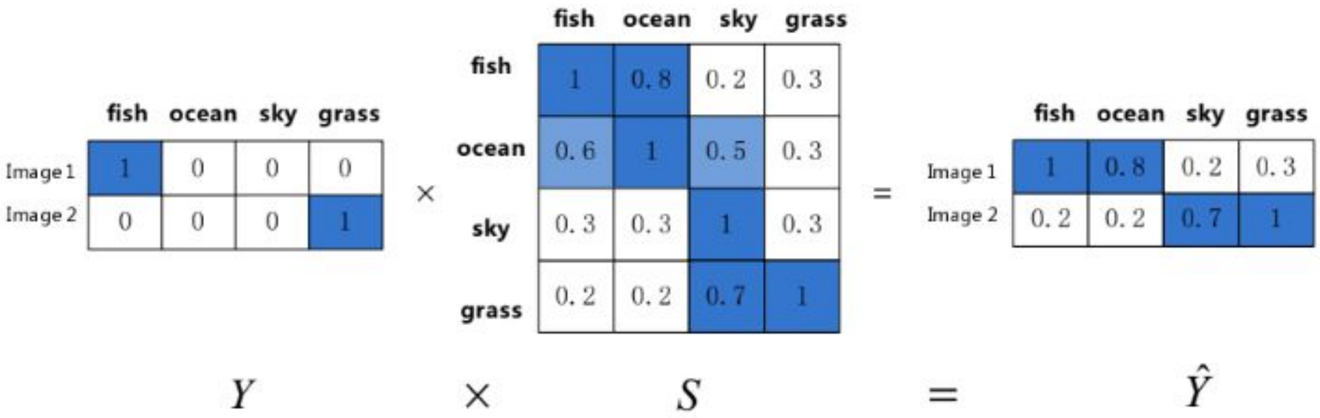


Figure 2. A supplementary label matrix $\hat{Y}$ obtained by multiplying the label correlation matrix S with the original label matrix Y

Random - pseudo label

- random pseudo label
  - For item i’s label j,
    P(unknown = positive) = 0.2 * normalized supplementary matrix[i][j]

- confident pseudo label
  - For each label,
    those >= min(known positive’s normalized supplementary value) is positive

Random - soft label

- label
    - Normalized supplementary matrix

- loss
    - Mean Square Error

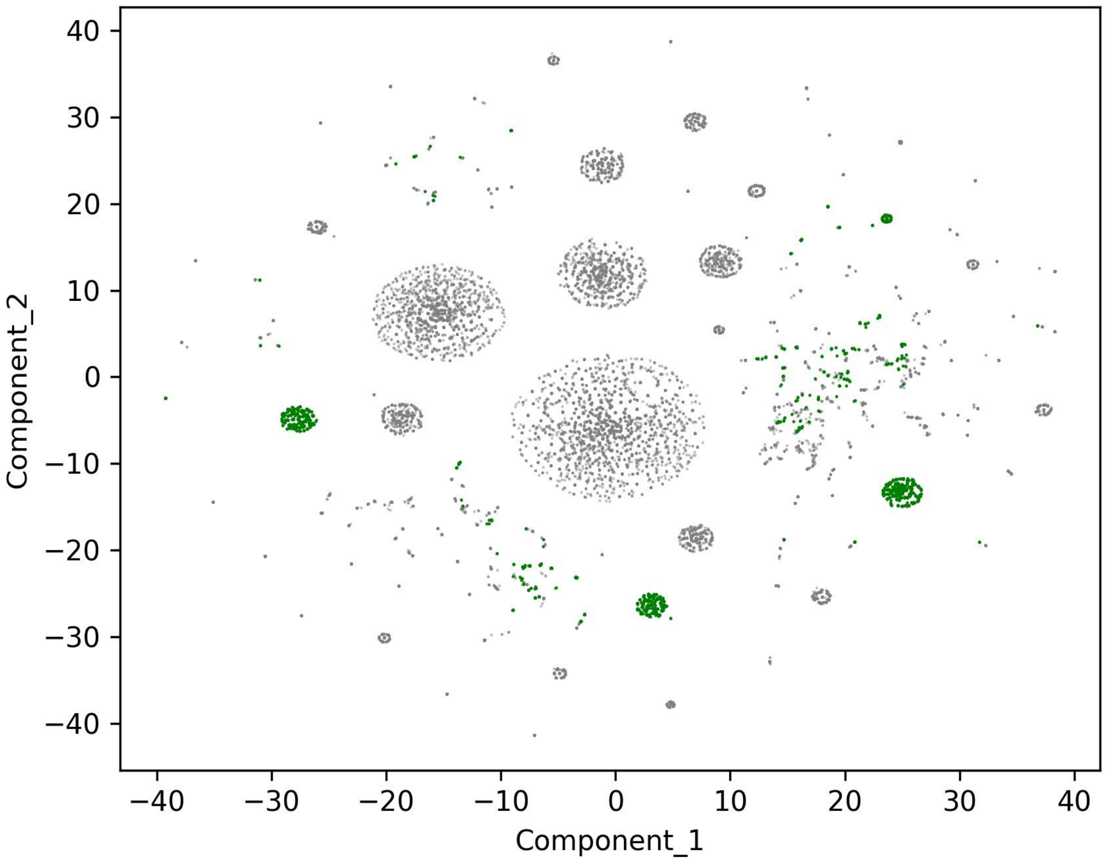
Random - result

Predicted result:

	MF	random_pseudo	conf_pseudo	soft label
ambig ratio	0.2	0.2	0.2	0.2
avg. label	1.6450	1.646	1.6418	1.6034
F1 score	0.88121	0.8801	0.8806	0.8779

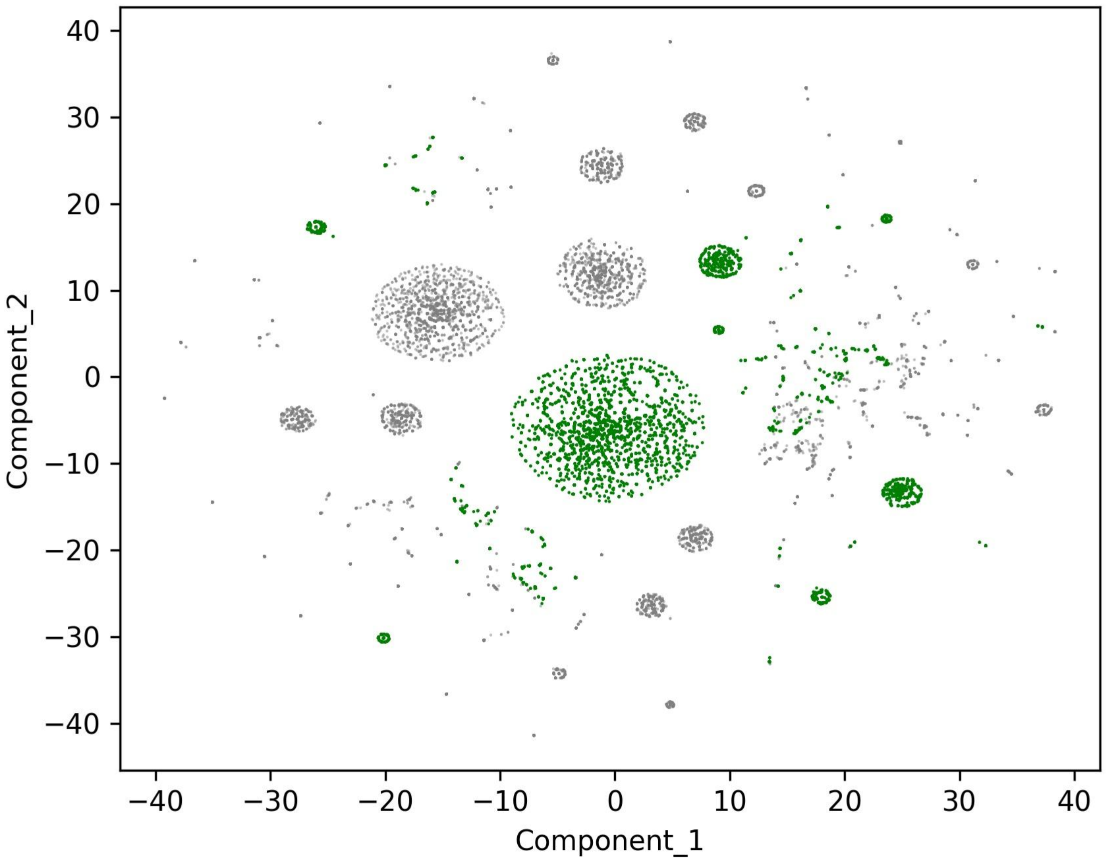
KNN

label = 6
supplementary matrix



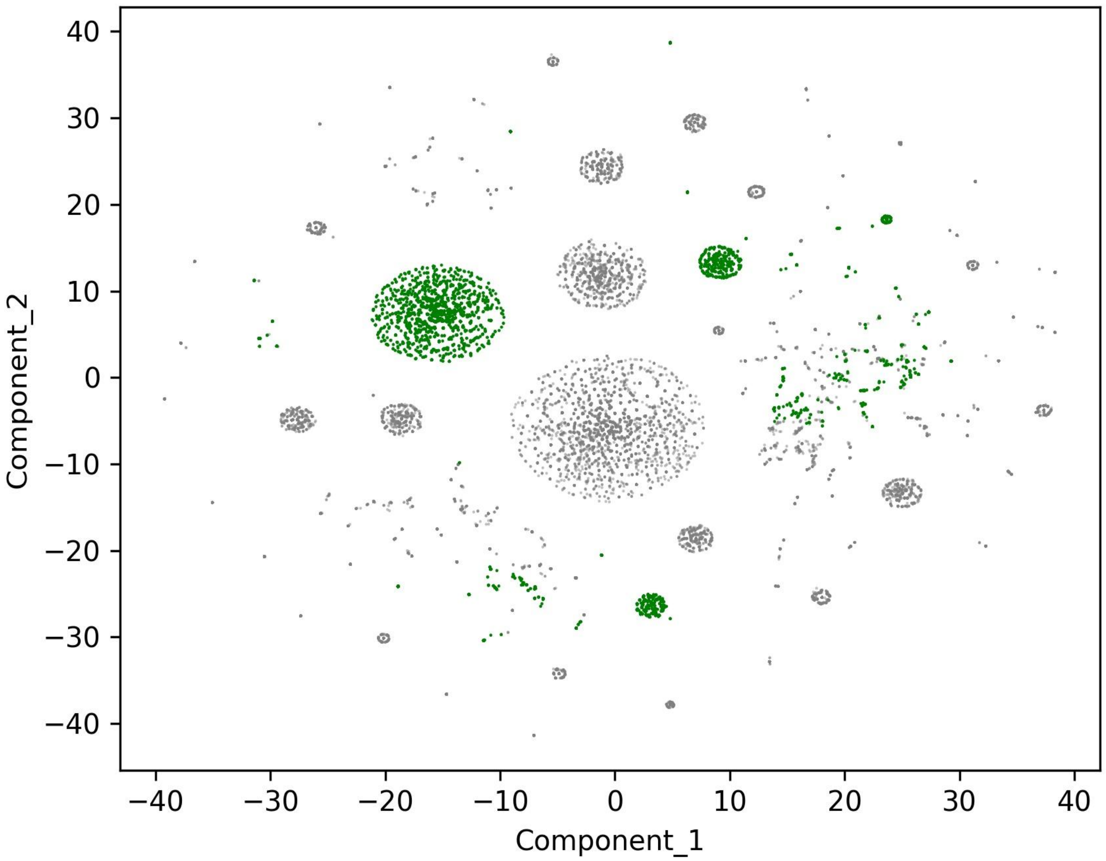
KNN

label = 8
supplementary matrix



KNN

label = 12
supplementary matrix



KNN (Supplementary matrix)

Public & Private Score: 0.87489
Label%: 1.555

Failed:

… the example above is based on the assumption that the correlation matrix can accurately capture the real relations shared among different labels, which will lead to the supplementary matrix with richer label information.

Rule

Rule

Thought:

1. Try to find other rules by ratings. => decreasing the parts of random.

2. Try those labels which should appear together for many times.
   ex. (1) labels that appear together in “random” case for many times, but not in “rule” case.
   (2) labels that with higher correlation. (by the result of previous part)

Rule

Result:

1. No matter we try ratings with MF, ratings with “important users”, or ratings with all ‘1’ and ‘0’, the F1 score does not get better.

   (Some public score: 0.94004(MF), 0.94086(‘1’, ‘0’))

=> We didn’t find other possible rules by applying ratings.

Rule

Result:
2-1. Because the result of the last page, we reduced the ratio of ratings while training model.

=> More speed, without loss.

Rule

Result:
2-2. Since we did not find other rules, we still need to train model by randomize some situation.

=> Performance not getting better.(Best: 0.94187) Why?

Rule

Guess:
Those labels we focus on are too “complicated”. (because they appear many times.)

=>   When randomizing situations, we might lead to bad results.

Rule

Possible solution:

1. Find other rules.
2. Try those “simple” labels. (which improve our score last time)
3. Change predicting strategy.

Rule

Predicting strategy:

Before:

Predict fewer labels, but we have high confidence about the prediction.

=> Not good when we have too few rules.

Rule

Predicting strategy:

After: (future work)

Try to predict more labels. (Hypothesize boldly, while prove it carefully.)

Future work

- More extension on rating data analysis (based on one-hot rating)
- Better methods about correlation
- Those possible solutions we have mentioned.

Reference

L. Xu, Z. Wang, Z. Shen, Y. Wang, E. Chen, "Learning low-rank label correlations for multi-label classification with missing labels", *IEEE International Conference on Data Mining*, pp. 1067-1072, 2014.

Responsibility

- Data analysis
  - Rating (MF) 黃柏瑋
  - Item (Correlation & Supplementary matrix) 陳心平
- Random
  - Label without Correlation
    - MF 黃柏瑋
    - One-hot rating 黃柏瑋、陳心平
  - Label with Correlation
    - Pseudo / Soft label 黃柏瑋
    - KNN 陳心平
- Rule 唐浩