Workshop 11

COMP20008 Elements of Data Processing

Learning outcomes

By the end of this class, you should be able to:

- perform feature selection using the chi-squared method
- implement the item- and user-based recommender systems algorithms introduced in lectures
- discuss practical issues that may influence the design of real recommender systems

Q1: Feature selection

Feature A	Feature B	Class label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
F	F	−

Select the best feature using the χ^2 method.

Q1: Feature selection

Feature A	Feature B	Class label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
F	F	−

Select the best feature using the χ^2 method.

First write down the *observed* contingency table for feature A

		Feature A		
		T	F	Total
Class	+			
	−			
	Total			

Q1: Feature selection

Feature A	Feature B	Class label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
F	F	−

Select the best feature using the χ^2 method.

First write down the *observed* contingency table for feature A

		Feature A		
		T	F	Total
Class	+	4	0	4
	−	2	4	6
	Total	6	4	10

Q1: Feature selection

Feature A	Feature B	Class label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
F	F	−

Select the best feature using the χ^2 method.

Next write down the *expected* contingency table for feature A (assuming independence).

		Feature A		
		T	F	Total
Class	+			4
	−			6
	Total	6	4	10

Q1: Feature selection

Feature A	Feature B	Class label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
F	F	−

Select the best feature using the χ^2 method.

Next write down the *expected* contingency table for feature A (assuming independence).

		Feature A		
		T	F	Total
Class	+	2.4	1.6	4
	−	3.6	2.4	6
	Total	6	4	10

Q1: Feature selection

		Feature A		
		T	F	Total
Class	+	4	0	4
	−	2	4	6
	Total	6	4	10

		Feature A		
		T	F	Total
Class	+	2.4	1.6	4
	−	3.6	2.4	6
	Total	6	4	10

Select the best feature using the χ^2 method.

Now compute the χ^2 statistic:

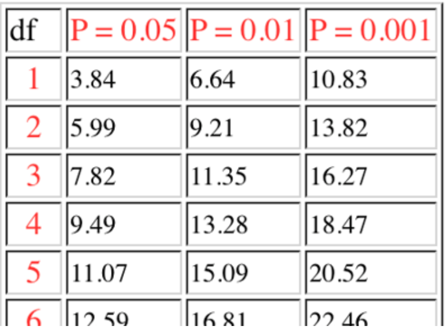
$$\chi^2 = \sum_{i\in\{+,-\}} \sum_{j\in\{\mathrm{T},\mathrm{F}\}} \frac{\left(o_{i,j} - e_{i,j}\right)^2}{e_{i,j}}$$

$$= \frac{(4-2.4)^2}{2.4} + \frac{(-1.6)^2}{1.6} + \frac{(2-3.6)^2}{3.6} + \frac{(4-2.4)^2}{2.4}$$

$$= 4.44$$

Degrees of freedom: $(2-1) \times (2-1) = 1$

Q1: Feature selection



df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46

Select the best feature using the χ^2 method.

Look up table to decide whether to reject the null hypothesis H_0 (that feature A and the class are independent).

The critical value is $\mathbf{3.84}$ at the $p = \mathbf{0.05}$ level for $\mathbf{df} = \mathbf{1}$. Since $\chi^2 = \mathbf{4.44}$ is greater than the critical value, we reject H_0.

Thus it’s highly likely feature A is not independent of the class label.

Q1: Feature selection

Feature A	Feature B	Class label
T	F	+
T	T	+
T	T	+
T	F	–
T	T	+
F	F	–
F	F	–
F	F	–
T	T	–
F	F	–

Select the best feature using the χ^2 method.

Repeat for feature B.

Q1: Feature selection

Feature A	Feature B	Class label
T	F	+
T	T	+
T	T	+
T	F	–
T	T	+
F	F	–
F	F	–
F	F	–
T	T	–
F	F	–

Select the best feature using the χ^2 method.

Repeat for feature B.

First write down the contingency tables.

Observed

		Feature B		
		T	F	Total
Class	+	3	1	4
	–	1	5	6
	Total	4	6	10

Expected

		Feature B		
		T	F	Total
Class	+	1.6	2.4	4
	–	2.4	3.6	6
	Total	4	6	10

Q1: Feature selection

		Feature B		
		T	F	Total
Class	+	3	1	4
	−	1	5	6
	Total	4	6	10

		Feature B		
		T	F	Total
Class	+	1.6	2.4	4
	−	2.4	3.6	6
	Total	4	6	10

Select the best feature using the χ^2 method.

Now compute the χ^2 statistic:

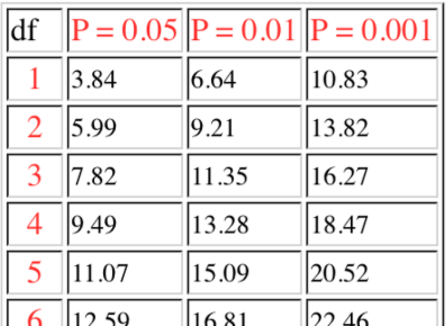
$$\chi^2 = \sum_{i \in \{+,-\}} \sum_{j \in \{\mathrm{T},\mathrm{F}\}} \frac{\left(o_{i,j} - e_{i,j}\right)^2}{e_{i,j}}$$

$$= \frac{(3-1.6)^2}{1.6} + \frac{(2.4-1)^2}{1} + \frac{(2.4-1)^2}{1} + \frac{(5-3.6)^2}{3.6}$$

$$= 3.40$$

Degrees of freedom: $(2-1) \times (2-1) = 1$

Q1: Feature selection



df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46

Select the best feature using the χ^2 method.

Look up table to decide whether to reject the null hypothesis H_0 (that feature B and the class are independent).

The critical value is $\mathbf{3.84}$ at the $p = \mathbf{0.05}$ level for $\mathbf{df} = \mathbf{1}$. Since $\chi^2 = \mathbf{3.40}$ is less than the critical value, we don't reject H_0.

Thus it's highly likely feature B is independent of the class label.

Q1: Feature selection

Feature A	Feature B	Class label
T	F	+
T	T	+
T	T	+
T	F	−
T	T	+
F	F	−
F	F	−
F	F	−
T	T	−
F	F	−

Select the best feature using the χ^2 method.

In summary:

- conclude feature A and class label are dependent at the 0.05 level
- conclude feature B and class label are independent at the 0.05 level

So we should select feature A

Q2: Feature selection in Python

Repeat Q1 in Jupyter notebook.

Q3a: Recommender systems

Use the item-based recommender systems approach discussed in lectures to predict Frank's rating for Batman.

User	Iron Man	Superman	Batman	Spiderman	Ant-Man	W Woman
Anne	3		3	3.5	2.5	3
Bob	4	3.5	2.5	4	3	3
Chris	3	3	3			4
Dave		3.5	2	4	2.5	
Eve		3		3	2	5
Frank	2.5	4		5	3.5	5
Gary		4.5	3	4	2	

Q3a: Recommender systems

Use the item-based recommender systems approach discussed in lectures to predict Frank's rating for Batman.

To predict a missing rating:

1. Find the three most similar items
2. Take Frank's average rating of these items, weighted by their similarity

Compute item similarity using:

$$\mathrm{sim}(i,j) = \frac{1}{1 + \mathrm{dist}(i,j)} \quad \text{where} \quad \mathrm{dist}(i,j) = \sqrt{\sum\nolimits_{u \in \mathrm{Users}} \left(r_{ui} - r_{uj}\right)^2}$$

(replace missing ratings with item-level mean).

Q3a: Recommender systems

Use the item-based recommender systems approach discussed in lectures to predict Frank's rating for Batman.

User	Iron Man	Superman	Batman	Spiderman	Ant-Man	W Woman
Anne	3	3.58	3	3.5	2.5	3
Bob	4	3.5	2.5	4	3	3
Chris	3	3	3	3.92	2.58	4
Dave	3.13	3.5	2	4	2.5	4
Eve	3.13	3	2.7	3	2	5
Frank	2.5	4	2.7	5	3.5	5
Gary	3.13	4.5	3	4	2	4
dist	1.94	2.76	0	3.70	1.75	4.10
sim	0.34	0.27	1	0.21	0.36	0.20

Temporarily replace missing values with item mean.

Then compute the similarities between Batman and the other movies.

Q3a: Recommender systems

Use the item-based recommender systems approach discussed in lectures to predict Frank’s rating for Batman.

Item	Frank’s rating	Similarity to Batman
Iron Man	2.5	0.34
Superman	4	0.27
Spiderman	5	0.21
Ant-Man	3.5	0.36
W Woman	5	0.20

Now compute the average rating of the three most similar items, weighted by their similarity to Batman:

$$\frac{0.36 \times 3.5 + 0.34 \times 2.5 + 0.27 \times 4}{0.36 + 0.34 + 0.27} = 3.29$$

Q3b: Recommender systems

Use the user-based recommender systems approach discussed in lectures to predict Frank’s rating for Batman.

User	Iron Man	Superman	Batman	Spiderman	Ant-Man	W Woman
Anne	3		3	3.5	2.5	3
Bob	4	3.5	2.5	4	3	3
Chris	3	3	3			4
Dave		3.5	2	4	2.5	
Eve		3		3	2	5
Frank	2.5	4		5	3.5	5
Gary		4.5	3	4	2	

Q3b: Recommender systems

Use the user-based recommender systems approach discussed in lectures to predict Frank's rating for Batman.

To predict a missing rating:

1. Find the three most similar users
2. Take the user's average rating of these items, weighted by similarity

Compute user similarity using:

$$\mathrm{sim}(u,v) = \frac{1}{1+\mathrm{dist}(u,v)} \quad \text{where} \quad \mathrm{dist}(u,v) = \sqrt{\sum\nolimits_{i \in \mathrm{Items}} (r_{ui} - r_{vi})^2}$$

(replace missing ratings with user-level mean).

Q3b: Recommender systems

Use the user-based recommender systems approach discussed in lectures to predict Frank's rating for Batman.

User	Iron Man	Superman	Batman	Spiderman	Ant-Man	W Woman	dist	sim
Anne	3	3	3	3.5	2.5	3	3.08	0.245
Bob	4	3.5	2.5	4	3	3	3.16	0.240
Chris	3	3	3	3.25	3.25	4	2.52	0.284
Dave	3	3.5	2	4	2.5	3	3.24	0.236
Eve	3.25	3	3.25	3	2	5	2.89	0.257
Frank	2.5	4	4	5	3.5	5	0	1
Gary	3.38	4.5	3	4	2	3.38	2.81	0.262

Temporarily replace missing values with user-level mean.
Then compute the similarities between Frank and the other users.

Q3b: Recommender systems

Use the user-based recommender systems approach discussed in lectures to predict Frank's rating for Batman.

User	Rating for Batman	Similarity to Frank
Anne	3	0.245
Bob	2.5	0.240
Chris	3	0.284
Dave	2	0.236
Eve		0.257
Garry	3	0.262

Now compute the average rating for Batman from the three most similar users, weighted by their similarity to Frank:

$$\frac{0.284 \times 3 + 0.262 \times 3 + 0.245 \times 3}{0.284 + 0.262 + 0.245} = 3$$

Q3c: Recommender systems

Identify advantages and disadvantages of the item- and user-based approaches.

Q3c: Recommender systems

Identify advantages and disadvantages of the item- and user-based approaches.

Item-based systems tend to perform better in practice:

- An index of similar items can be computed offline. Thus an item-based system can make fast recommendations, even for new users.
- In contrast, an index of similar users would quickly be out-of-date.
- Similarity scores between items are more likely to converge over time than similarity scores between users
- Item-based systems can produce good recommendations for new users with limited data

Q4: Cold-start problem

Recommender systems are challenged by the “cold start” problem: how to make recommendations to new users and new items. Suggest three strategies that might be used to address this.

Q4: Cold-start problem

Recommender systems are challenged by the "cold start" problem: how to make recommendations to new users and new items. Suggest three strategies that might be used to address this.

New users

- Use personal info (age, gender, etc) to recommend items liked by similar users
- Preference elicitation—i.e. ask questions
- Recommend popular items
- Use cross-domain info—e.g. book preferences to recommend films
- User's social network

New items

- Ask users to rate new items
- Initialise using info of neighbouring (similar) items. Requires a hybrid recommender system.

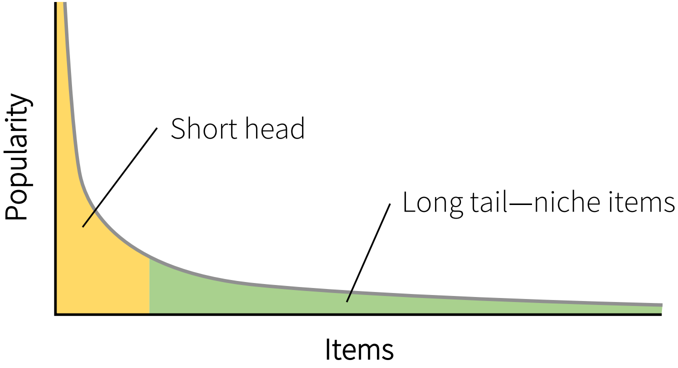
Q5: Long-tail problem

Recommender systems are sometimes criticised for over-recommending popular items and under-recommending rarer items. Why do you think this happens? How might it be addressed?

Q5: Long-tail problem

Recommender systems are sometimes criticised for over-recommending popular items and under-recommending rarer items. Why do you think this happens? How might it be addressed?

- Partly explained by a feedback loop:
  - popular items are recommended
  - they receive more ratings
  - popularity is amplified…
- Recommending popular items can be a decent strategy
- But the long-tail problem is why recommender systems exist



Q5: Long-tail problem

Recommender systems are sometimes criticised for over-recommending popular items and under-recommending rarer items. Why do you think this happens? How might it be addressed?

Managing the long-tail problem:

- use a hybrid system (content filtering + collaborative filtering)
- learn to rank items from the tail to improve relevance
- incorporate diversity and timeliness of an item

Revision

2016 Exam

Euclidean distance: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$

3. Consider the following dataset D which describes 3 people:

Age	Weight
20	40
30	50
25	25

a) (2 marks) Using a Euclidean distance function, construct the dissimilarity matrix for D. Show all working. (You may leave any square root terms unsimplified).

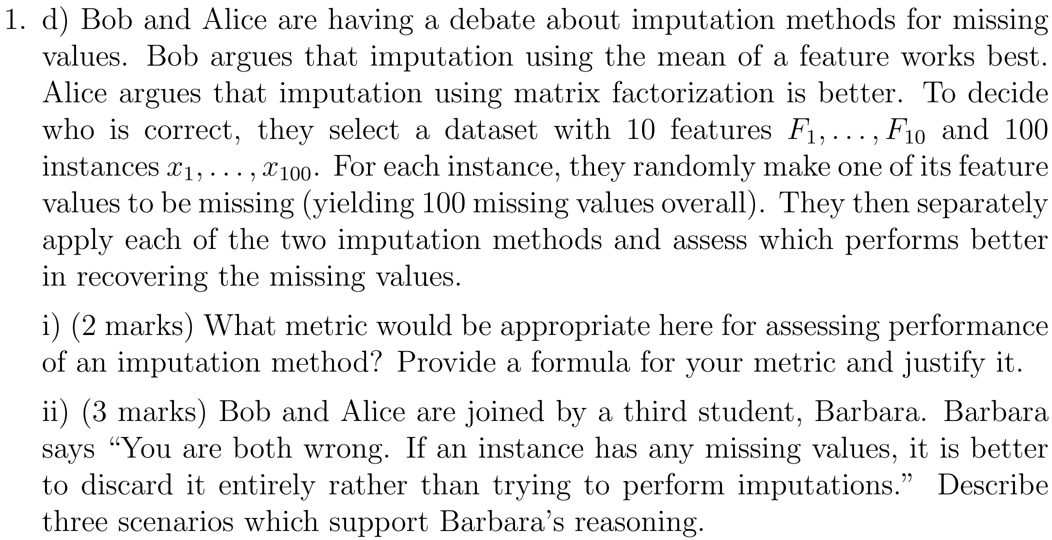
b) (2 marks) A dissimilarity matrix can be visualised as a heat map. Explain why it can be useful to reorder a dissimilarity matrix, using a technique such as the VAT algorithm. What information can this provide and why?

c) (2 marks) Would there be a benefit of applying principal components analysis to D to assist in visualisation? Explain.

2016 Exam

Euclidean distance: $d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

2016 Exam



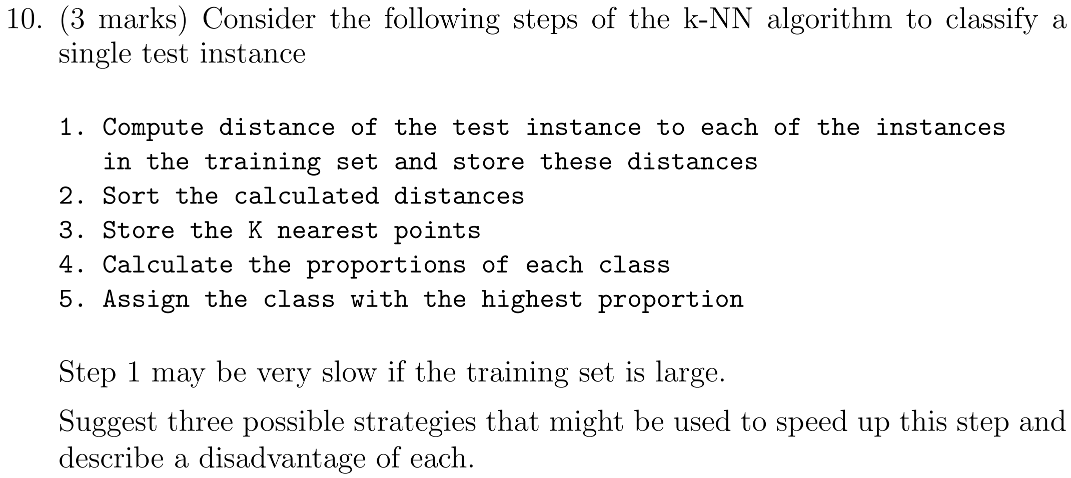
1. d) Bob and Alice are having a debate about imputation methods for missing values. Bob argues that imputation using the mean of a feature works best. Alice argues that imputation using matrix factorization is better. To decide who is correct, they select a dataset with 10 features $F_1, \ldots, F_{10}$ and 100 instances $x_1, \ldots, x_{100}$. For each instance, they randomly make one of its feature values to be missing (yielding 100 missing values overall). They then separately apply each of the two imputation methods and assess which performs better in recovering the missing values.

   i) (2 marks) What metric would be appropriate here for assessing performance of an imputation method? Provide a formula for your metric and justify it.

   ii) (3 marks) Bob and Alice are joined by a third student, Barbara. Barbara says "You are both wrong. If an instance has any missing values, it is better to discard it entirely rather than trying to perform imputations." Describe three scenarios which support Barbara's reasoning.

2016 Exam

2018 Exam



10. (3 marks) Consider the following steps of the k-NN algorithm to classify a single test instance

```
1. Compute distance of the test instance to each of the instances
   in the training set and store these distances
2. Sort the calculated distances
3. Store the K nearest points
4. Calculate the proportions of each class
5. Assign the class with the highest proportion
```

Step 1 may be very slow if the training set is large.

Suggest three possible strategies that might be used to speed up this step and describe a disadvantage of each.

2018 Exam