Knowledge Discovery and Data Mining 1 (VO) (707.003)

Sample Examination Questions

Denis Helic

KTI, TU Graz

Jan 16, 2014

Example 1

Exercise

Suppose we have a utility matrix of a movie recommender system. This matrix keeps the user ratings for various movies. In our movies database we have only movies of two genres: science fiction and romance. The utility matrix:

User \ Movie	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	0	0	4	4
Jenny	0	0	0	5	5
Jane	0	0	0	2	2

Example 1

Exercise

For the purposes of recommending movies to new users we decompose the utility matrix using SVD decomposition. Thus, we map the users and movies into the concept space spawned by two movie genres: science fiction and romance. The SVD decomposition is given by:

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 2 & 2 \end{pmatrix} =$$

$$\begin{pmatrix} 0.14 & 0 \\ 0.42 & 0 \\ 0.56 & 0 \\ 0.70 & 0 \\ 0 & 0.60 \\ 0 & 0.75 \\ 0 & 0.30 \end{pmatrix} \times \begin{pmatrix} 12.4 & 0 \\ 0 & 9.5 \end{pmatrix} \times \begin{pmatrix} 0.58 & 0.58 & 0.58 & 0 & 0 \\ 0 & 0 & 0 & 0.71 & 0.71 \end{pmatrix}$$

Example 1

Exercise

1. What are these four matrices?
2. How do we interpret them?
3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?
4. Which other movies should we recommend to Quincy?
5. What about Leslie who rated *Alien* with 3 and *Titanic* with 4 stars.

Example 1

Solution

1. What are these four matrices?

The matrices are:

- the utility matrix $\mathbf{M}$
- $\mathbf{U}$ is a matrix of eigenvectors of $\mathbf{M}\mathbf{M}^T$
- $\mathbf{V}$ is a matrix of eigenvectors of $\mathbf{M}^T\mathbf{M}$
- $\mathbf{\Sigma}$ is the matrix of the square roots of eigenvalues (singular values) of $\mathbf{M}\mathbf{M}^T$ or $\mathbf{M}^T\mathbf{M}$.

Example 1

Solution

2. How do we interpret them?

Interpretation:

- **M** connects users to movies
- **U** connects users to concepts (genres)
- **V** connects movies to concepts
- **Σ** gives importance of concepts

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- First we need to represent Quincy in the utility matrix **M**. How can we do that?

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- First we need to represent Quincy in the utility matrix $\mathbf{M}$. How can we do that?
- Each row of $\mathbf{M}$ is a user. We represent Quincy with a row vector:

$$\mathbf{q}^T = \begin{pmatrix} 4 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- Now we need to assess Quincy's interests in different genres. How can we do that?

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- Now we need to assess Quincy's interests in different genres. How can we do that?
- We need to "map" Quincy into concept space. How?

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- Now we need to assess Quincy's interests in different genres. How can we do that?
- We need to "map" Quincy into concept space. How?
- What does $\mathbf{q}^T$ connect?

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- Now we need to assess Quincy's interests in different genres. How can we do that?
- We need to "map" Quincy into concept space. How?
- What does $\mathbf{q}^T$ connect?
- A user with movies
- What do we need?

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- Now we need to assess Quincy's interests in different genres. How can we do that?
- We need to "map" Quincy into concept space. How?
- What does $\mathbf{q}^T$ connect?
- A user with movies
- What do we need?
- The connection between the user and concepts

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- How to relate the user with concepts?

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- How to relate the user with concepts?
- $\mathbf{q}^T$ connects a user with movies, $\mathbf{V}$ connects movies to concepts
- $\mathbf{q}^T\mathbf{V}$ gives us connection between the user and concepts
- $\mathbf{q}^T\mathbf{V} =$?

Example 1

Solution

3. Suppose we have a new user Quincy. Quincy has only seen *Matrix* and rated it 4. How are Quincy's interests in different movie genres?

- How to relate the user with concepts?
- $\mathbf{q}^T$ connects a user with movies, $\mathbf{V}$ connects movies to concepts
- $\mathbf{q}^T\mathbf{V}$ gives us connection between the user and concepts
- $\mathbf{q}^T\mathbf{V} = ?$

$$\mathbf{q}^T\mathbf{V} = \begin{pmatrix} 2.32 & 0 \end{pmatrix}$$

- Quincy's interest in science fiction is 2.32 and he does not have interest in romance

Example 1

Solution

4. Which other movies should we recommend to Quincy?

   - Now we need to assess how Quincy would like other movies according to his interests. How we can do that?

Example 1

Solution

4. Which other movies should we recommend to Quincy?

- Now we need to assess how Quincy would like other movies according to his interests. How we can do that?
- We need again a relation between the user and movies, i.e. we need a row from the utility matrix $\mathbf{M}$
- $\mathbf{q}^T\mathbf{V}$ relates the user with concepts
- $\mathbf{V}$ relates movies with concepts

Example 1

Solution

4. Which other movies should we recommend to Quincy?

   - How do we obtain the relation between users and movies?

Example 1

Solution

4. Which other movies should we recommend to Quincy?

- How do we obtain the relation between users and movies?
- $\mathbf{q}^T\mathbf{V}\mathbf{V}^T$

$$\mathbf{q}^T\mathbf{V}\mathbf{V}^T = \begin{pmatrix} 1.33 & 1.33 & 1.33 & 0 & 0 \end{pmatrix}$$

- Quincy would like *Alien* and *Star wars*

Example 1

Solution

5. What about Leslie who rated *Alien* with 3 and *Titanic* with 4 stars.

   - We represent Leslie with a row vector:

$$\mathbf{q}^T = \begin{pmatrix} 0 & 3 & 0 & 0 & 4 \end{pmatrix}$$

Example 1

Solution

5. What about Leslie who rated *Alien* with 3 and *Titanic* with 4 stars.

   - Leslie's interests in genres:

$$\mathbf{q}^T\mathbf{V} = \begin{pmatrix} 1.74 & 2.84 \end{pmatrix}$$

   - Leslie's interest in science fiction is 1.74 and interest in romance is stronger: 2.84

Example 1

Solution

5. What about Leslie who rated *Alien* with 3 and *Titanic* with 4 stars.

   - User-movie matrix

$$\mathbf{q}^T\mathbf{V}\mathbf{V}^T = \begin{pmatrix} 1.00 & 1.00 & 1.00 & 2.00 & 2.00 \end{pmatrix}$$

   - Leslie would like *Casablanca* at most

Example 2

Exercise

For evaluation of the quality of a classifier we use the contingency table.

1. Sketch this table and write down the names for the table cells.
2. Using the terms from the contingency table explain how we measure accuracy of a classifier.
3. Explain what happens with accuracy in the presence of a skewed class distribution? Do we need alternative measures?
4. Define the precision and recall.
5. Explain precision-recall trade-off and F1 measure.

Example 2

Solution

1. Sketch this table and write down the names for the table cells.

Example 2

Solution

1. Sketch this table and write down the names for the table cells.

Prediction \ Real class	c	c^c
c	true positive (tp)	false positive (fp)
c^c	false negative (fn)	true negative (tn)

Table: Contingency table

Example 2

Solution

2. Using the terms from the contingency table explain how we measure accuracy of a classifier.

Example 2

Solution

2. Using the terms from the contingency table explain how we measure accuracy of a classifier.

$$A = \frac{tp + tn}{tp + fp + fn + tn}$$

Example 2

Solution

3. Explain what happens with accuracy in the presence of a skewed class distribution? Do we need alternative measures?

Example 2

Solution

3. Explain what happens with accuracy in the presence of a skewed class distribution? Do we need alternative measures?

   - We have one small and one huge class
   - $P(cancer) = 0.008$
   - $P(cancer^c) = 0.992$

Example 2

Solution

3. Explain what happens with accuracy in the presence of a skewed class distribution? Do we need alternative measures?

Example 2

Solution

3. Explain what happens with accuracy in the presence of a skewed class distribution? Do we need alternative measures?

- We always predict: $cancer^c$:

Prediction \ Real class	c	c^c
c	0	0
c^c	8	992

- $A = \frac{tp+tn}{tp+fp+fn+tn} = \frac{0+992}{0+0+8+992} = 0.992$
- We need alternatives

Example 2

Solution

4. Define the precision and recall.

Example 2

Solution

4. Define the precision and recall.

   - Recall $R = \frac{tp}{tp+fn}$
   - Precision $P = \frac{tp}{tp+fp}$

Example 2

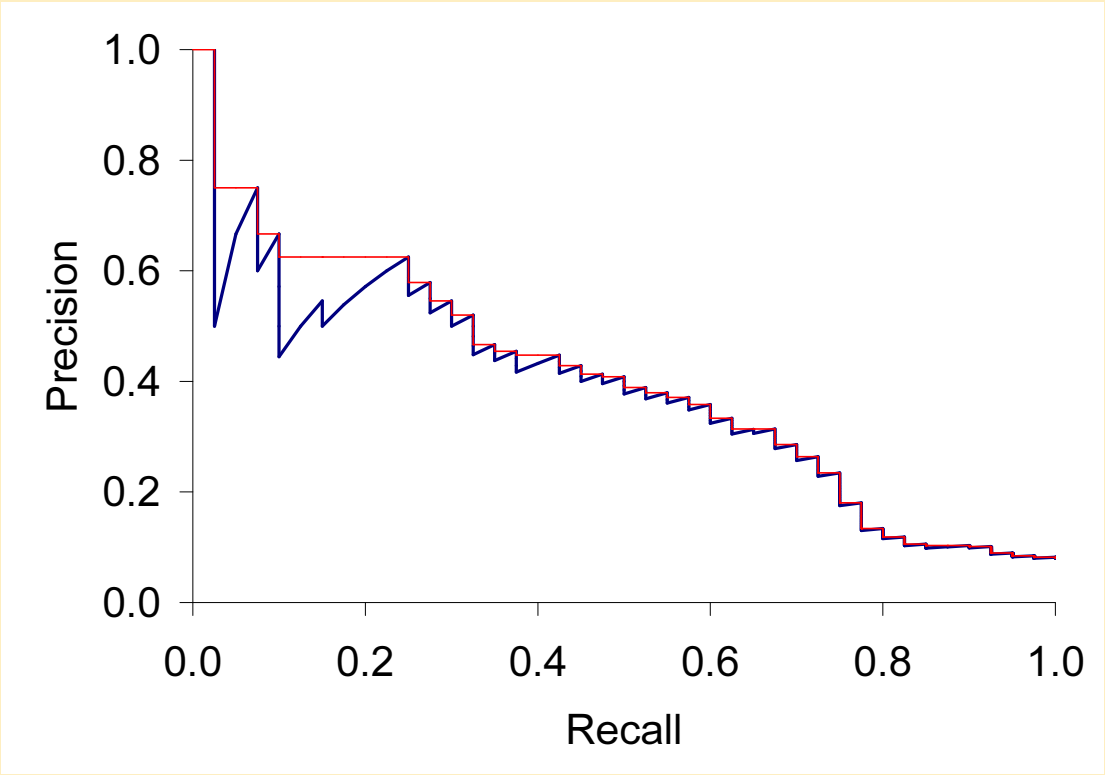
Solution

5. Explain precision-recall trade-off and F1 measure.

Example 2

Solution

5. Explain precision-recall trade-off and F1 measure.



Example 2

Solution

5. Explain precision-recall trade-off and F1 measure.

Example 2

Solution

5. Explain precision-recall trade-off and F1 measure.

   - $F1 = \frac{2PR}{P+R}$