Qualify Exam – Data Mining, 2018

1. (20 points) Please briefly describe the following terminologies. (1) Sequential Pattern (2) Entropy (3) Information Gain (4) Ensemble method (5) OLTP

2. (20 points) For **each** following evaluation criteria, please **describe** and **explain ONE** retrieval system in which this criterion is important. (1) R-Precision (2) NDCG (3) P@3 (4) AUC (5) Specificity.

3. (20 points) What is "overfitting" problem in classification modeling? How to do "overfitting"? How to deal with it?

4. (20 points) Please apply FP-growth algorithm to find association rules in the following transaction data.

TID	Items bought
100	{a, c, d, f, g, i, m, p}
200	{a, b, c, f, i, m, o}
300	{b, f, h, j, o}
400	{b, c, k, s, p}
500	{a, c, e, f, l, m, n, p}

5. (20 points) You have designed a retrieval system on a data set with 1000 web pages. The test results of 2 queries are listed below. Please answer the following questions.
   (1). (8 points) If the numbers of positive pages for these 2 queries are 5 and 6, please answer the R-precision and MAP of M2 (just list your result, don't need to calculate out).
   (2). (5 points) Use the result of Q1 to draw the ROC curve of M1 and M2.
   (3). (7 points) Use ROC curve to judge which method is better? How to improve it?

Top-10 results	Result (M1)	Result (M2)
Q1	OOXXO XXXOO	XXXOX OOOOX
Q2	OXOXO XOXOX	OOOOX XXOOX