Name: Tianhong, Zhang

Email: zhan4868@umn.edu

Student ID: 5293616

Question 1.

a. True. Redundant attributes receive similar weights and do not degrade the quality of the classifier. However, if the number of irrelevant or reduntant attributes is large, the learning of the ANN model may suffer from overfitting, leading to poor generalization performance.
b. False. SVM and ANN does not handle very large training data sets equally well. SVN is more robust to the presence of a large number of irelevant and redundant attributes than other classifiers.
c. False. The SVM learning problem can be formulated as a convex optimization problem, in which efficient algorithms are available to find the global minimum of the objective function. Neural Network tend to find only locally optimum solutions. They do not always produce the same decision boundaries.

,

Question 2.

a. 1. TP = 90, FP = 5, TN = 95, FN = 10
   2. TP = 90, FP = 50, TN = 950, FN = 10
   3. TP = 90, FP = 500, TN = 9500, FN = 10

b. 1. p = 90/95 = 0.947, r = 0.9, F = 180/195 = 0.923, TPR/FPR = 18
   2. p = 90/140 = 0.643, r = 0.9, F = 180/240 = 0.75, TPR/FPR = 18
   3. p = 90/590 = 0.153, r = 0.9, F = 180/690 = 0.261, TPR/FPR = 18

c. T3 is strictly better than T4 (irrespective of skew). T3 has a higher TPR and a lower FPR than T4, and both TPR and FPR do not take into account the skew among the classes.

d. T1 has a lower TPR and lower FPR than T4, so it is worse but not strictly worse than T4; T2 has a higher TPR and a higher FPR than T4, so it is worse but not strictly worse than T4; Regarding with TPR/FPR, T1 = 50, T2 = 9.9, T4 = 18; T2 is worse than T4 (irrespective of skew) in this measure.

e. Which classifiers may be better or worse than T4 (depending on skew)?
   1. For balanced skew case: T3 is better than T4 because of its high precision and high F-measure. T1 and T2 may be better or worse than T4, depending on different factors, e.g., if we want most of a classfier's positive predictions correct, we want a classifier with a high precision, than T1 is better suited. However, in cases that correctly identifying all the positive instances is what matters most, a high recall is preferred, so that T2 is better than T4.
   2. For medium skew case, T2 and T3 may be better than T4 due to their higher precision and recall. T1 may be better or worse than T4, take different factors into consideration.
   3. For high skew case, T1, T2 and T3 may all be better than T4. T1 may be worse than T4, since it has a lower recall.

Question 3.

a. $R = 3^d - 2^{d+1} + 1 = 602$

b. $i = 2^6 -1 = 63$

c. $\binom{6}{3} = 20$

d. $\sigma(Bread) = 6$, $\sigma(Milk) = 5$,$\sigma(Bread\cup Milk) = 3$

e. $c(Bread \to Milk) = 3/6 = 0.5$

$c(Milk \to Bread) = 3/5 = 0.6$

f. $\sigma(Butter) = \sigma(Milk) = 5$, Milk and Butter is a pair of items that have the same confidence.

Question 4.

- a) True
- b) False. ACD is a frequent itemset, thus CD must be frequent.
- c) True.
- d) False, AB is not a frequent itemset, ABDM cannot be a frequent itemset.
- e) False. ABCDE cannot be a frequent itemset.

Question 5.

- a. {C,D,E}, {H,I}
- b. {C,D}, {D,E}, {C,E}, {C},{D}, {E},{H}, {I}, {C,D,E}, {H,I}
- c. {C}, {D}, {I}, {C,D,E},{H,I}

Question 6.

- a. Candidate Generation: ${abcp},{abcw}, {abpw}, {acpw}$, candidate pruning: ${abcp}$
- b. Candidate Generation: ${abcp}$, candidate pruning: ${abcp}$
- c. No, it is impossible to generate a frequent 5-itemset. For Fk-1 x Fk-1 method, we need at least two 4-itemsets that are frequent, however, we have only one candidate 4-itemsets.

Question 7.

- a. The number of frequent itemsets for Dataset 1 = $2^4-1 = 15$
  The number of frequent itemsets for Dataset 2 = $2^3-1 = 7$
  The number of frequent itemsets for Dataset 3 = $2^2-1 = 3$
- b. Dataset 1 will produce the longest frequent itemset.
- c. Dataset 2 will produce frequent itemsets with highest maximum support.
- d. Dataset 2 will produce frequent itemsets containing items with widely varying support levels.
- e. The number of maximal frequent itemsets for Dataset 1 = 1
  The number of maximal frequent itemsets for Dataset 2 = 1
  The number of maximal frequent itemsets for Dataset 3 = 1
  They all produce the same number of maximal frequent itemsets.
- f. The number of closed frequent itemsets for Dataset 1 = 1
  The number of closed frequent itemsets for Dataset 2 = 3
  The number of closed frequent itemsets for Dataset 3 = 1
  Dataset 2 will produce the greatest number of closed frequent itemsets.

Question 8.

1. (b) false. $s(A) = s(A,B)$
2. (c) cannot decide.
    - if ${B}$ is a closed itemset,e.g., $s(B)=100$ and no other items have 100 support.
    - if ${B}$ is not a closed itemset, e.g.,$s(B)=80$
3. (c) cannot decide.
    - if ${D,E}$ is a closed itemset,e.g., none of its immediate supersets has exactly the same support count as 40.
    - if ${D,E}$ is not a closed itemset,e.g., $s(A,D,E)=40$
4. (a) True. Because all immediate supersets of ${D,E,F}$ has support counts less than or equal to 27, and $s(D,E,F)=30$
5. (b) false. $s(D,E,F,G) = s(D,E,F,G,H)$
6. (a) True. ${A, B, C,D,E,F,G,H,I,J}$ has no supersets.