Sample Exam (solutions)

Instructions:

You have **four (4)** hours to complete this exam. No textbooks, notes, or calculators are allowed during the exam. The total score that can be obtained in this exam is **100 points**. In order to pass the exam you should obtain at **least 60 points**. The Exam consists of two parts. Part A is worth **20 points** and Part B is worth **80 points**. You are allowed to use a dictionary (English to any language).

Part A

IMPORTANT: Your answers should be written on the provided answer sheet.

Part B

You have three questions to answer**.** For each question, you are expected to answer all parts. You should provide concise answers. Note that short and precise answers are preferred to lengthy answers.

The mapping to letter grades will tentatively be as follows:

- A: 100 – 93
- B: 92 – 86
- C: 85 – 77
- D: 76 – 68
- E: 67 – 60
- Fx: 59 – 50
- F: 49 – 0

Part A (20 points)

All questions should be answered by indicating whether statements are correct or not.

(a) Given the following confusion matrix that describes the results of classifying a test data set (the columns represent the predicted values and the rows represent the actual values):

		Predicted		
	Classes	apple	orange	pear
Actual	apple	4	5	1
	orange	2	10	5
	pear	0	3	4

For each of the following statements, indicate whether it is correct or not:

1: The accuracy is (4+10+4) / (4+5+1+2+10+5+0+3+4). **T**

2: The precision for class 'apple' is 4 / (4+5+1). **F**

3: The recall for class 'pear' is 4 / (1+5+4). **F**

4: The test data used in our experiment consists of 10 apples. **T**

(b) For each of the following statements about classification and clustering, indicate whether it is correct or not:

5: The training error of the 1-Nearest Neighbor classifier is always 0. **T**

6: In the presence of outliers in our data, using K-means is preferable to K-medoids. **F**

7: One difference between K-means and K-medoids is that in K-medoids the cluster representative (cluster center) is never an observed data example whereas in K-means this is not always the case. **F**

(c) For each of the following statements about frequent itemset mining, indicate whether it is correct or not:

8: All closed itemsets are also maximal. **F**

9: If an itemset is frequent then all its subsets have a relative support that is higher than the relative support of that itemset. **F**

10: Association rule {apples, oranges}→{bananas} with 75% confidence means that if a customer buys bananas, then with 75% probability that customer may also buy apples and oranges. **F**

Part B (80 points)

Question 1 (Frequent itemsets and association rules) **[30 points]**

(a) Explain the *Apriori principle* and stress its importance for itemset mining? [10 points]

The apriori principle states that if an itemset is frequent then all its subsets are frequent. It is useful for itemset mining since it can help pruning the search space of candidate itemsets, hence resulting in a reduced computational time.

(b) Suppose you discover the following two association rules for a given set of transactions:

{A}→{B} and **{B}→{A}**

If the confidence of both rules is 1, what can you tell about the (relative) supports of items A and B? Motivate your answer carefully. Recall from our lecture that

confidence(X→Y) = support(X, Y) / support(X)

[10 points]

if	conf (A→B)	=	conf (B→A)
then	support (A,B) / support (A)	=	support (A,B) / support (B)
hence	support (A)	=	support (B)

(c) Assume the Apriori algorithm identified the following *seven* 4-itemsets that satisfy a user given support threshold:

abcd, abce, abcf, acde, adef, bcde, bcef

What are the initial candidate 5-itemsets generated by the Apriori algorithm? Which of those candidates survive subset pruning? [10 points]

We should look for all pairs of itemsets that share the same prefix of size 3. There are three such pairs:

1. abcd and abce
2. abcd and abcf
3. abce and abcf

From pair 1 we obtain abcde

From pair 2 we obtain abcdf

From pair 3 we obtain abcef

Finally, for each of the three candidates we check whether there exists any subset of size 4 that is not already mined, i.e., that is not in the set of frequent itemsets of size 4. **All three candidates are eliminated**!! For candidate 1, abde is not frequent; for candidate 2, abdf is not frequent; and for candidate 3, abef is not frequent.

Question 2 (Clustering and classification reduction) [30 points]

(a) In class we discussed that K-means is sensitive to *initialization*. Provide an illustrative example of this claim. Briefly describe a solution to this initialization problem. [10 points]

The answer can be found in Lecture 4, slides 24-28. In this example, we assume that the two initial centers are the means of the two corresponding points. If this initialization is assigned, then K-means converges immediately; nonetheless the cluster assignments are far from optimal.

(b) Assume we are trying to learn a decision tree. Our input data consists of N samples, each with k attributes ($N \gg k$). We define the depth of a tree as the maximum number of nodes between the root and any of the leaf nodes (including the leaf, not the root).

a. If all attributes are binary, what is the maximal number of leaf (decision) nodes that we can have in a decision tree for this data? What is the maximal possible depth of a decision tree for this data? [10 points]

If the attributes are binary, in the worst case, we will perform a binary split for each and every attribute in each path of the tree. Moreover, in the worst case each attribute may appear in each path once. Hence, this leads to 2^k leaf nodes.

The maximum depth of the tree is k (equal to the number of attributes per path in the tree) plus the final leaf node of the path.

b. If all attributes are continuous, what is the maximum number of leaf nodes that we can have in a decision tree for this data? What is the maximal possible depth for a decision tree for this data? [10 points]

In the worst case, we will end up splitting using the same attribute several times on the same path until we run out of examples. This means that the depth is N-1 (as many as the examples minus one, since the last node in the path will have two leaves), and the number of leaf nodes is N (as many as the examples).

Question 3 (Deep Learning and Model Evaluation) [20 points]

(a) Mary and Bob are given a classification problem on a dataset containing 1000 real-valued features (among which, the last feature is the class label) and 1 million examples (records), with no missing values. Their task is to compare the performance of 1-NN and random forests on the given dataset. Bob decides to build both classifiers using all data examples and the first 999 features. To his surprise, when he evaluates the performance of the built classifiers on the same data examples and same features, he observes that both classifiers achieve close to 99.9% precision and recall. Mary decides to take an alternative approach and instead performs 10-fold cross-validation using all 1000 features. Again, both classifiers achieve recall and precision close to 99.9%. When they both present their approaches and results to their boss, they instantly get fired. Explain what was wrong with their approaches. [10 points]

Bob made the terrible mistake of training and testing on the same data examples. He did not test on an independent (to the training set) set.

Even though Mary avoided making the terrible mistake of Bob, she did sth even more tragic: she used the class label as a feature for training.

(b) Prof. Birdfreak wants to use an artificial neural network (ANN) to automatically determine the species of Galapagos finches (birds of the subfamily Geospizinae) in images using the following measurements: (i) beak length, (ii) beak height, (iii) eye diameter, (iv) head length, and (v) body length. Given the location where the pictures were taken, the possible species are: (i) Large Ground Finch Geospiza magnirostris, (ii) Medium Ground Finch Geospiza fortis, (iii) Small Tree Finch Camarhynchus (formerly Geospiza) parvulus, and (iv) Green Warbler-Finch Certhidea olivacea. He has a database of a few hundred labelled images of individuals of these species on which to train his ANN. The are several design aspects Prof. Birdfreak needs to take into account:

1) How many input and how many output units should the ANN have? [2 points]

Input units: 5 (as many as the image features to be used for learning)
Output units: 4 (as many as the species)

2) Should the ANN have hidden neurons? [2 points]

It should have at least one (by definition!).

3) What activation functions should the ANN use? [2 points]

Any activation function is a good answer as long as you mention the name of the function: e.g., the sigmoid, or Relu.

4) How should the ANN weights be trained? [4 points]

They should be initialized randomly and then trained using the forward backpropagation algorithm.