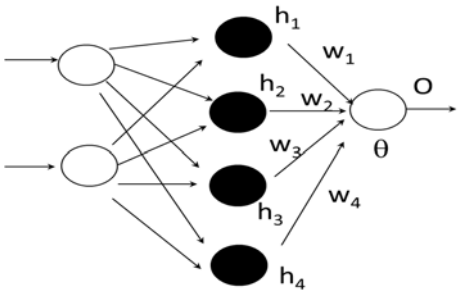
Data Mining Research & Practices –Final Exam

1. (a) (4 %) Design a multilayer feed-forward neural network for the given data in Table 1. Label the nodes in the input and output layers.

Table 1

	Outlook	**Wind**	**PM2.5**	**Play baseball**
Day 1	Rainy	Strong	1.7	No
Day 2	Sunny	Weak	0.8.	Yes
Day 3	Rainy	Weak	0.2	Yes
Day 4	Overcast	Strong	0.7	Yes
Day 5	Rainy	Weak	1.5	No

Given the neural network below. Assume that T and O are the actual and predicted values of the output node, respectively; h_1, h_2, h_3 and h_4 are the output values of the hidden layer. w_1, w_2, w_3 and w_4 are the link weights; θ is the bias of the output node. Sigmoid function ($h(x) = 1/(1+e^{-x})$) is used as the activation function of the output node. λ is the learning rate. The derivative of h(x) is h(x)(1-h(x)).



(b) (3%) Derive the input and output (O) of the output node, respectively.

(c) (6%) Briefly explain how to update the link weight w_2 by using the gradient descent approach. You need to derive the equation for updating w_2.

(d) (3%) Derive the equation for updating θ.

2. (a)(3%) Given the following large (frequent) **3-itemsets**:

< 1 2 3 >
< 1 2 4 >
< 1 3 4 >
< 1 3 6 >
< 2 3 4 >
< 2 3 5 >
< 2 4 5 >
< 3 4 5 >

Find the candidate 4-itemsets according to the **Apriori-generate** algorithm. Find the candidate 4-itemsets after pruning.

(b)(3%) Given the following large **3-sequences**:

< 1 2 4 >
< 1 2 5 >
< 1 4 5 >
< 1 5 2 >
< 1 5 4 >
< 2 3 5 >
< 2 5 4 >
< 3 1 4 >
< 3 1 5>

Find the candidate 4-itemsets according to the **Apriori-generate** algorithm. Find the candidate 4-itemsets after pruning.

3.

(a) (4%) Explain the concept of support vectors, maximum marginal hyperplane and linear separation between classes in SVM. You should draw a diagram to aid your explanation.

(b) (4%) Use the following example to explan the usage of Kernel function.

$$\Phi(u) = (u_1^2, u_2^2, \sqrt{2}u_1, \sqrt{2}u_2, 1); \Phi(v) = (v_1^2, v_2^2, \sqrt{2}v_1, \sqrt{2}v_2, 1); \Phi(u) \bullet \Phi(v) = ?$$

(c) (4%) Explain the following equation. You need to explain z, x_i, y_i , λ_i , z, x_i • **z,** and the usage of f (z). Are all the training instances used in computing f (z)?

$$f(\mathbf{z}) = sign(\mathbf{w} \cdot \mathbf{z} + b) = sign\left(\sum_{i=1}^{N} \lambda_i y_i \mathbf{x_i} \cdot \mathbf{z} + b\right).$$

4. A database has five transactions. Let min_sup=50% and min_conf=75%.

Table

TID	DATE	ITEMS_BOUGHT
T1	1/01/20	B, C, D, E F
T2	1/02/20	B, C, D, E, G
T3	1/03/20	A B, C, E, F
T4	1/04/20	C, E F, G
T5	1/05/20	A, B, E

(a) (5%) List all of the *strong* association rules (with support *s* and confidence *c*) matching the following metarule, where X is a variable representing customers and $item_i$ denotes variables representing items (e.g., "A", "B", etc.):

$\forall x \in transaction, buys(X, item_1) \wedge buys(X, item_2) \Rightarrow buys(X, item_3)[s, c]$

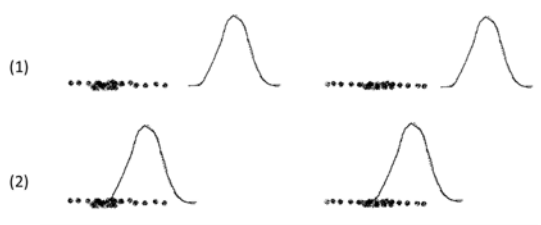
(b) (4%) Establish the ***FP-tree*** and find out "*Conditional Pattern Base", "Conditional FP-tree"* and "*Frequent Patterns Generated"* for **item F.**

5.

(a) (7%) Explain the basic concept of EM (Expectation-Maximization) clustering. What are the differences between K-means and EM clustering in terms of **the assignment of data points to clusters** and **the computation of centroids / model parameters**?

(b) (2%) Given the following two mixture models (1) and (2). Which one has higher expected likelihood? Why?

$$P(\mathbf{O}|\Theta) = \prod_{i=1}^{n} \sum_{j=1}^{k} \omega_j P_j(o_i|\Theta_j)$$



6. Assume that there are two latent topics A and B, and the LDA (Latent Dirichlet Allocation) outputs the following assignments of the words to the topics.

Doc1: Apple: B, Banana: B, Potato: B, Hamster: A

Doc2: Banana: B, Potato: B, Kitten: A, Cute: A

Doc3: Hamster: A, Cute: A, Kitten: B, Apple B

(a) (4%) Derive the topic distributions (proportion) for each document.

(b) (2%) Derive the probability distributions of words for topic A.

(c) (4%) Explain the idea of updating the topic assignment of current word *w* in document *d* based on the topic distributions of all the documents (p(topic k | document d)) and word distributions of all the topics (p(word w | topic k)).

7. (a) (4%) Explain how k-means clustering is executed in Map-Reduce by using four Map tasks and three Reduce tasks to cluster data into three clusters. You should use examples (partial data) and draw a diagram to aid your explanations. Clearly indicate the key values that are shuffled to the Reduce tasks.

(b)(4%) Explain the map method and reduce method for k-means clustering. You also need to clearly indicate the input and output of the two methods.

8. (a) (5%) Briefly explain the Hadoop Distributed File System (HDFS). You should draw a diagram to aid your explanations. You need to explain the functions of namenode and datanode.

(b)(2%) Briefly explain the advantage for the job tracker to decide on where to run each map task based on the concept of locality.

(c)(3%) Briefly explain one key idea of Spark **R**esilient **D**istributed **D**atasets (RDD).

9. (a) (4%) There are three types of layers to build CNN architectures: Convolutional Layer, Pooling Layer, and Fully-Connected Layer. In the Convolution Layer, an image is convolved with a filter. Given the following matrix below, please calculate the output of ? after the Convolutional Layer with ReLU as the activation function. (ReLU function is defined as y = max(0, x).)

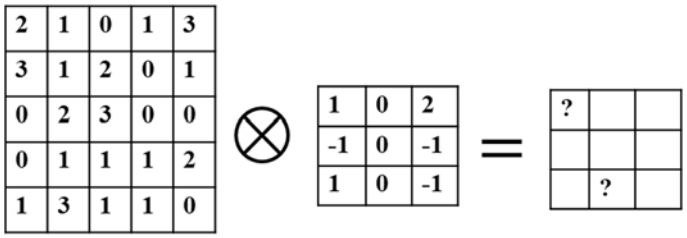
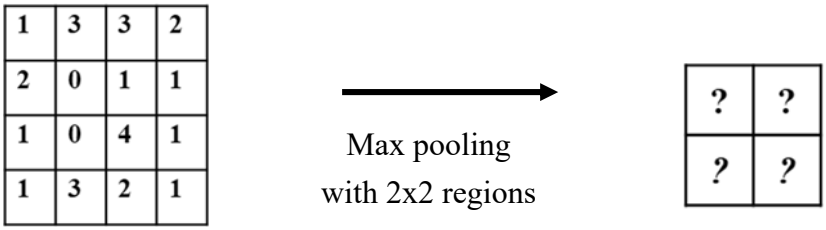


Image Filter

(b) (2%) Please calculate the output after Max Pooling Layer.



10. Give the following CNN neural network model. The size of the input image is 32x32 with RGB colors.

```
model.add(Conv2D(filters=16, kernel_size=(3,3), padding='same', input_shape=(32,32,3), activation='relu'))
model.add(MaxPooling2D(pool_size=(2, 2)))
model.add(Conv2D(filters=24, kernel_size=(3,3), padding='same', activation='relu'))
```

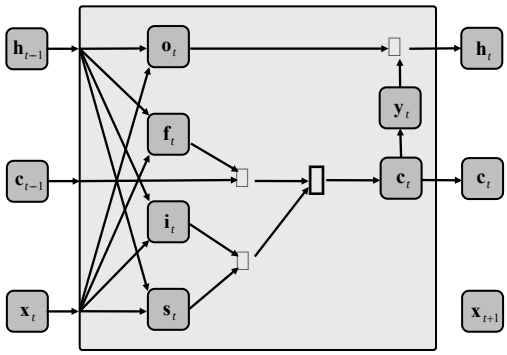
(a) (4%) Derive the number of link weights and biases respectively for conv2D layer 1.

(b) (2%) Derive the number of link weights and biases respectively for max pooling layer.

(c) (4%) Derive the number of link weights and biases respectively for conv2D layer 2.

You need to draw the size of the neural nodes and the number of filters for the two connecting layers to answer each of the above questions.

11. (a) (6%) Given the LSTM architecture below. Briefly explain the usages of i_t, f_t and c_t, respectively.



Given the following LSTM neural network model in Keras.

```
modelLSTM.add(Embedding(output_dim=16, input_dim=5000, input_length=100))
modelLSTM.add(LSTM(64))
modelLSTM.add(Dense(units=2,activation='sigmoid'))
```

(b) (2%) Derive the number of parameters for the embedding layer.

(c) (4%) Derive the number of link weights and biases respectively for the LSTM layer.

(d) (2%) Derive the number of link weights and biases respectively for the dense layer.

You need to draw the size of the neural nodes and the links connecting to the layer to answer questions (c) and (d), respectively.