Machine Learning
Homework Assignment 2

Due Date: September 16th (BEFORE CLASS)

(Each problem is 10 points)

CONCEPT LEARNING

Problems 1 to 2 use the following:

Consider a dataset with 5 *discrete* or *symbolic* features: **A, B, C, D, E** where:

$$\mathbf{A} \in \{a_1, a_2\}, \mathbf{B} \in \{b_1, b_2, b_3\}, \mathbf{C} \in \{c_1, c_2, c_3, c_4\},$$
$$\mathbf{D} \in \{d_1, d_2, d_3, d_4, d_5\}, \mathbf{E} \in \{e_1, e_2, e_3, e_4, e_5, e_6\}$$

Problem 1

Consider the **traditional** concept learning where a concept is given by a tuple (NOTE: that we **don't allow** $\varnothing$)

$$\Lambda = \langle \lambda_A, \lambda_D, \lambda_C, \lambda_D, \lambda_E \rangle, \text{ where } \lambda_X \in \{?\} \cup \mathbf{X}$$

(a) **[2 points]** How many "**most specific hypotheses**" are there? (NOTE: most specific hypotheses have no "?" and each element of the tuple is a valid variable value)

(b) **[2 points]** What is the **total number** of the hypotheses in this version space?

(c) **[6 points]** In the following 6 hypothesis, identify ALL pairs $\Lambda_S < \Lambda_G$ of hypotheses where Λ_S is more specific than Λ_G:

$$\Lambda_1 = \langle a_1, b_3, c_2, d_4, e_5 \rangle,\ \Lambda_2 = \langle a_2, ?, c_3, d_5, ? \rangle,\ \Lambda_3 = \langle a_1, b_3, ?, d_4, ? \rangle$$
$$\Lambda_4 = \langle a_2, b_1, c_3, d_5, e_2 \rangle,\ \Lambda_5 = \langle ?, b_3, c_2, d_4, e_5 \rangle,\ \Lambda_6 = \langle a_2, ?, c_3, ?, ? \rangle$$

Problem 2

In the traditional version space each Λ_X can either be "?" or only **ONE** of the possible values that variable **X** can take. Lets expand this definition such that each Λ_X can take any **non-empty** subset of the values of variable **X** i.e.:

$$\Lambda = \langle \lambda_A, \lambda_D, \lambda_C, \lambda_D, \lambda_E \rangle, \text{ where } \lambda_X \in 2^{\mathbf{X}} \setminus \varnothing$$

where $2^{\mathbf{X}} \setminus \varnothing$ is the power set of **X** minus the null set. Consider the following hypothesis:

$$\Lambda_1 = \langle \{a_1\}, ?, \{c_2, c_4\}, \{d_1, d_3, d_4\}, \{e_5\} \rangle$$

This hypothesis is **true** for all instances where:

$$(A \in \{a_1\}) \wedge (C \in \{c_2, c_4\}) \wedge (D \in \{d_1, d_3, d_4\}) \wedge (E \in \{e_5\})$$

(a) **[2 points]** What is the size of this new version space?

(NOTE: Remember we don't allow empty set for any variable)

(b) **[2 points]** List all the **minimal specialization** of hypothesis = <?, ?, ?, ?, ?>
(NOTE:
$$\Lambda = \langle ?,?,?,? \rangle = \langle \{a_1,a_2\},\{b_1,b_2,b_3\},\{c_1,c_2,c_3,c_4\},\{d_1,d_2,d_3,d_4,d_5\},\{e_1,e_2,e_3,e_4,e_5,e_6\} \rangle$$

(c) **[2 points]** List all possible instances for which the following hypothesis is true.
$$\Lambda_1 = \langle \{a_1\},?,\{c_2,c_4\},\{d_1,d_3,d_4\},\{e_5\} \rangle$$

(d) **[2 points]** List all the **minimal specializations** of the above hypothesis.
(e) **[2 points]** List all the **minimal generalizations** of the above hypothesis.

Problem 3 [10 points] Instead of using the **candidate elimination algorithm** for learning version spaces, consider the new **greedy algorithm** that goes from maximally general to specific hypothesis by picking the one that maximally increases the accuracy of a hypothesis. (NOTE: We are going to use the version space defined in problem 2).

$$\text{Iteration: } t \leftarrow 0$$
$$h_t \leftarrow \langle ?,?,...? \rangle,$$
$$\mathbf{S}(h_t) \leftarrow \text{Minimally Specific hypotheses of } h_t$$
$$h_{t+1} \leftarrow \arg\max_{h \in \mathbf{S}(h_t)} \{Accuracy(h)\}$$
$$\textbf{while } \left(Accuracy(h_{t+1}) > Accuracy(h_t)\right) \{$$
$$\quad t \leftarrow t+1$$
$$\quad \mathbf{S}(h_t) \leftarrow \text{Minimally Specific hypotheses of } h_t$$
$$\quad h_{t+1} \leftarrow \arg\max_{h \in \mathbf{S}(h_t)} \{Accuracy(h)\}$$
$$\}$$

Code the above algorithm. Generate a sequence of hypothesis on the **Mushroom data** (ignore the data points that have even one missing value). Submit the code as well as the accuracies of the hypothesis in the sequence.

DECISION TREES

Problem 4: **Decision Tree Classifier on Mushroom Data**.
(a) Randomly partition the data into 5 buckets.
(b) In each EXPERIMENT take two of the 5 buckets as TEST data and remaining 3 buckets as TRAINING data. So we can do (5 choose 2) = 10 EXPERIMENTS.
(c) In each EXPERIMENT build a DECISION TREE classifier using the TRAINING data and evaluate it on the TEST data.
(d) Report the MEAN and STANDARD DEVIATION of the TEST set and TRAINING set accuracies for the 10 experiments.
(e) Limit the decision tree with different DEPTHS (depth 4, 8, 12, 16, 20) and see the effect of this on accuracy. Plot depth vs. accuracy.

K-MEANS CLUSTERING

There are FOUR aspects of a K-Means clustering algorithm that we will explore.

(i) **Number of clusters** – K = 5, 10, 15, 20, 25.
(ii) **Cluster Initialization** – ***random*** initialization vs. ***farthest_first_point*** initialization
(iii) **Projection Method** – (raw data, PCA(9) projection, and Fisher(9) projection)
(PCA(9) = first 9 dimensions obtained by doing PCA projection of MNIST data)
(Fisher(9) = all 9 dimensions obtained by doing Fisher projection on MNIST data)
(iv) **Clustering Metric** – method used to EVALUATE the cluster.
(a) **MEAN SQUARED ERROR** and **PURITY**

Definition of Cluster PURITY Metric

Normally it is **not** obvious how to EVALUATE an unsupervised method such as Clustering. However since for MNIST data has class labels, we can use these labels as "ground truth" to define a measure of *Purity*:

$$\text{Let: } \mathbf{L} = \left\{ \mathbf{L}_c = \text{Set of points in class } c \right\}_{c=1}^{C}$$

$$\text{Let: } \mathbf{M} = \left\{ \mathbf{M}_k = \text{Set of points in cluster } k \right\}_{k=1}^{K}$$

$$Purity(\mathbf{L},\mathbf{M}) = \frac{1}{N} \sum_{k=1}^{K} \max_{c=1 \ldots C} \left| \mathbf{L}_c \cap \mathbf{M}_k \right|$$

<u>Problem 5</u>: [10 points] K-Means Clustering with Random Initialization

- Do Clustering on MNIST data with K = 5, 10, 15, 20, 25 clusters.
- Do random initialization for each clustering run.
- Repeat random initialization 30 times for the same K
- Draw the images of cluster means for one of the clustering for each K.
- Compute MEAN and STANDARD DEVIATION of BOTH cluster metrics:
    (a) Cluster Mean Squared Error and
    (b) PURITY for various K's.

<u>Problem 6</u>: [5 points] K-Means Clustering with Farthest First Point Initialization

- Do Clustering on MNIST data with K = 5, 10, 15, 20, 25 clusters.
- Do farthest first point initialization for each clustering.
- Measure MEAN SQUARED ERROR and PURITY for various K's.
- Draw the images of cluster means for one of the clustering for each K.
- Compare with corresponding metrics with random initialization.

<u>Problem 7</u>: [5 points] K-Means Clustering of PCA projected MNIST data

- Project MNIST data into top **9 PCA** dimensions.
- Do Clustering on PCA projected MNIST data with K = 5, 10, 15, 20, 25 clusters.
- Do farthest first point initialization for each clustering.
- Measure MEAN SQUARED ERROR and PURITY for various K's.
- Draw the images of cluster means for one of the clustering for each K.
- Compare with corresponding metrics with problems 6 and 7.

<u>Problem 8</u>: [5 points] K-Means Clustering of FISHER projected MNIST data

- Project MNIST data into top **9 FISHER** dimensions.
- Do Clustering on Fisher projected MNIST data with K = 5, 10, 15, 20, 25 clusters.
- Do farthest first point initialization for each clustering.
- Draw the images of cluster means for one of the clustering for each K.

- Measure MEAN SQUARED ERROR and PURITY for various K's.
- Compare with corresponding metrics with problems 6, 7 and 8.

Problem 9: **Spherical K-Means Clustering on Newsgroup text data**.

(a) Compute the number of documents in which each word occurred (DocFreq(word)).
(b) Compute the INVERSE DOCUMENT FREQUENCY of each word (See class notes).
(c) Compute the TFIDF representation of each document – sparse, normalized.
(d) Modify the Farthest First Point algorithm for Cosine similarity and sample initial cluster centers using this modification.
(e) Perform Clustering for K = 5, 10, 15, 20, 25 clusters of Newsgroup data.
(f) Print the TOP 10 words for each cluster center for each K.
(g) Compute the PURITY measure for various K.