In the Name of God

Machine Learning

Homework I

Concept Learning

Assignment Date: **16 Esfand** Submission Deadline: **25 Esfand**

Contents

Objectives and Precautions

In this homework you will learn:

- Data manipulation and preprocessing
- Basic learning methods : Rote learner, Find-S, Candidate Elimination

Also note that:

- You are welcome to use any programming language you want, but we suggest using python. And for this homework you will only need the pandas library.
- You are required to prepare a report containing your results and your answers to the homework questions. Your submission then will be a zip file containing your code files and your report. Naming format is “your names.zip” for example ”Achraf_Hakimi_Mohamed_Salah.zip”.
- Late submission strategy is: 70 percent score for one day delay and 50 percent for 2 days delay. Submissions after 2 days will not be graded.
- Feel free to ask your questions in the telegram channel.
- Do not copy other works, write your own code.

Thank you. Good Luck.

1 Concept Learning

This homework is a practice on the concepts you learned form Tom M. Mitchell, Machine learning, Chapter 2. That is the concepts of a Rote Learner, Hypothesis Space, Version Space, Find-S, and the Candidate Elimination algorithm.

1.1 Data Preprocessing: adults.csv

Here, you will be using a subset of the adults income dataset originally released by the UCI Machine Learning repository in 1996. Follow the instructions to prepare your data for the next section:

- Load your data and check the number of its features and samples. What are the target values (also referred to as labels)?
- You are going to use only the following 5 features and the target column, therefore drop the rest of the feature columns except for: 'workclass', 'education', 'marital.status', 'race', 'sex', 'income'.
- When you drop a number of features from a dataset, you take away what may distinguish between two samples. Meaning that some samples will become identical. Therefore, drop the duplicates from your dataset. How may duplicates where there?
- Now check your dataset for nan values and drop any sample that has a nan value in one of its features. These nan values sometimes are represented by a question mark.
- At this point you will have only 1995 samples in your dataset. Plot a histogram of every feature and see the distribution of your samples over each feature.
- Now we are going to simplify your dataset by reducing the number of unique values in its features. Therefore, use the mappings shared with you in the text file 'mappings.txt' and map the original feature values to more general ones.
- You may find that some feature values are mapped to '?'. This is because We want to drop any sample with these feature values. So, we mark them with '?' to drop them later. After you have done the mappings on the whole dataset, drop samples with '?' in any of their features; don't forget to drop duplicate samples as well.
- Now you have 138 samples! Plot the histogram of every feature again. Your dataset is ready.

1.2 Practice the Learning Algorithms

Pick the first 8 samples in your dataset as training samples and the ones located at indices 9 and 10 as test samples. Then use the following learners to predict class label for the test samples: (you can either do this part on paper or write a program)

1. Rote learner
2. Find-s (find the maximally specific hypothesis)
3. Candidate Elimination (find the version space)

And answer the following questions:

- What's a hypothesis?
- What does it mean when we say a hypothesis is consistent with training samples?
- Are your training samples consistent? Randomly select another 8 samples from your dataset, are those consistent as well?
- What's a hypothesis space? What is the size of hypothesis space in this example?
- What's a version space? Find the version space in this example.
- What is a query sample in candidate elimination algorithm? What is a good query sample for the version space you found?
- Can all the above methods predict class labels for the test samples? Which ones can't? Why?
- How accurate are the class predictions using each of the above methods?