Introduction to Machine Learning-Assignment 1

Due: September 30, 2019

1. Maximum Posteriori Hypothesis (10 points)

When a test for steroids is given to soccer players, 98% of the players taking steroids test positive and 12% of the players not taking steroids test positive. Suppose that 5% of soccer players take steroids. What is the maximum posteriori hypothesis for a soccer player who tests positive? What are the exact posterior probabilities?

2. Naive Bayes Classifier (20 Points + 10 Points)

2.1 Consider the hypothesis space defined over these instances (Table 1), in which each hypothesis is represented by a pair of 4-tuples. Using the native Bayes classifier to predict the target value PlayTennis = Yes/No to the following instance.

a) <Sunny, Mild, Normal, Weak>
b) <Rain, Cool, High, Strong>

2.2 Consider the following example and calculate the accuracy of the classifier with precision, recall, F1-score, sensitivity, specificity and ROC curve using Python.

Table 1: Training examples for the target concept PlayTennis

Day	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	Normal	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	Normal	Strong	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Mild	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Rain	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Strong	Yes
D10	Rain	Hot	Normal	Strong	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Sunny	Mild	Normal	Weak	No
D13	Overcast	Hot	Normal	Weak	Yes
D14	Sunny	Cool	High	Weak	No

3. KNN (20 Points + 10 Points)

3.1 Table 2 is a gene expression microarray data, where each row represents a sample for a person, each column represents a gene, and each entry represents gene expression value of a gene over a sample. The output of each sample indicates whether a specific disease exists in this sample. Please use Euclidean Distance-Weighted KNN (K=3) to predict the output of the sample 11 and sample 12.

Sample	G1	G2	G3	G4	G5	G6	G7	Output
1	1.0	2.3	5.2	1.2	5.3	2.6	2.3	Yes
2	2.0	3.6	1.8	2.3	1.6	2.1	1.5	No
3	1.5	1.5	4.1	1.3	1.2	3.1	1.6	Yes
4	2.2	1.9	9.5	1.5	1.5	4.2	1.4	No
5	3.9	2.4	5.3	1.7	1.6	2.5	2.9	Yes
6	5.1	3.6	2.7	2.6	1.7	2.8	3.4	Yes
7	1.8	4.2	3.6	3.5	1.6	3.4	1.3	No
8	2.3	1.5	7.2	4.1	7.1	3.1	1.8	No
9	4.2	2.4	6.2	2.9	2.5	3.3	2.5	Yes
10	3.6	5.6	1.9	3.2	2.6	5.2	2.7	No

Table 2: Gene expression data

Sample	G1	G2	G3	G4	G5	G6	G7	Output
11	2.1	2.2	3.2	1.4	5.1	2.4	1.4	?
12	2.4	2.3	3.4	3.8	2.3	5.7	5.2	?

3.2 Consider the following example and calculate the accuracy of the classifier with precision, recall, F1-score, sensitivity, specificity and ROC curve using Python.

4. Decision Tree (20 Points + 10 Points)

4.1 Consider the hypothesis space defined over these instances (Table 3), in which each hypothesis is represented by a pair of 5-tuples. Please provide a hand trace of the ID3 algorithm to build a Decision Tree Classifier.

Table 3: Training examples for the target concept PlayTennis

D	Outlook	Temperature	Humidity	Wind	Play
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	Yes
D3	Rain	Hot	High	Weak	Yes
D4	Rain	Mild	Normal	Strong	No
D5	Rain	Cool	High	Weak	No
D6	Rain	Mild	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Cool	High	Weak	No
D9	Sunny	Cool	High	Strong	Yes
D10	Rain	Mild	High	Strong	Yes

4.2 Consider the following example and calculate the accuracy of the classifier with precision, recall, F1-score, sensitivity, specificity and ROC curve using Python.