Questions and Exercises

Course: Machine Learning

### Chapter 1: Introduction

* 1. Specify whether the following use statistical or syntactic pattern recognition

(a) The patterns are vectors formed by the features. The nearest neighbor based classifier is used.

(b) The patterns are complex and are composed of simple sub-patterns which are themselves built from simpler sub-patterns.

(c) Support vector machines are used to classify the patterns.

(d) The patterns can be viewed as sentences belonging to a language.

**1.2** Explain the following terms: pattern, class, classifier, feature, decision rule, and decision boundary.

**1.3** What is a training set? How it is chosen? Does the size of the training set influence to the performance of the classification process?

### Chương 2: Representation

**2.1** Find the centroid and medoid (most centrally located pattern) for the following set of patterns:

(1, 1), (1, 3), (1, 4), (2, 2), (2, 3), (3,1), (3,4), (4,2).

**2.2** Find the edit distance between the two strings “HOUSE” and “MOUND”.

**2.3**. Given the following distance measures. Let state which one is metric and which is non-metric.

Euclidean distance, edit-distance, Manhattan distance, Dynamic Time Warping distance.

**2.4** A dataset consists of the following patterns:

(1, 1, 1), (2, 2, 1), (1.5, 0.5, 1), (1, 3, 1), (4, 4, 2), (5, 5, 2), (4, 5, 2), (4, 6, 2)

Where each pattern consists of the *x* coordinate, the *y* coordinate and the class label. Find the direction of the ***w*** vector associated with Fisher’s Linear Discriminant.

**2.5** If there are 10 features and it is necessary to reduce the number of features to 6 so that the best set of six features is chosen, what is the number of feature sub-sets to be evaluated to find the optimal set of six features in the exhaustive search ?

**2.6** Give some differences between two feature abstraction methods LDA (Fisher Linear Discriminant) và PCA.

**2.7** Given a data set ***X*** consisting of 10 patterns (each pattern has 3 features) as follows: (7, 4, 3), (4, 1, 8), (6, 3, 5), (8, 6, 1), (8, 5, 7), (7, 2, 9), (8, 2, 2), (7, 4, 5), (9, 5, 8) và (5, 3, 3).

Compute covariance matrix of the dataset ***X***. (Check the result using the function *cov*(*A*) in MatLab).

**2.8** We need to classify a person as male or female, based on the three features: height, weight and foot length. The training set is given as follows (assume that the data set satisfies Gaussian distribution):

| Gender | Height (feet) | Weight  (lbs) | Foot length (inch) |
| --- | --- | --- | --- |
| M  M  M  M  F  F  F  F | 6  5.92  5.58  5.92  5  5.5  5.42  5.75 | 180  190  170  165  100  150  130  150 | 12  11  12  10  6  8  7  9 |

Given a test pattern with height = 6, weight = 130 and foot length = 8. Use covariance matrix and Mahalanobis distance to classify the test pattern.

**2.9** Find the eigen values and the eigen vector for the following matrix.



Hint: To check the result, use function *eig*(*A*) of MatLab.

**2.10** Given the confusion matrix of a classifier as follows:

|  | | Predicted | |
| --- | --- | --- | --- |
| C1 | C2 |
| Actual | C1 | 4 | 1 |
| C2 | 2 | 4 |

Compute sensitivity, specificity, precision and accuracy of the classifier.

**2.11**

a. Given a classifier for three classes C1, C2 and C3. Based on confusion matrix, given the formula to measure the accuracy of this classifier.

b. Given the confusion matrix of a classifier for three classes C1, C2 and C3 as follows:

|  | | Predicted | | |
| --- | --- | --- | --- | --- |
| C1 | C2 | C3 |
| Actual | C1 | 19 | 4 | 1 |
| C2 | 3 | 20 | 4 |
| C3 | 2 | 3 | 21 |

Compute the accuracy of this classifier.

**2.12** How to represent a chromosome in the population of the genetic algorithm that is used for feature selection.

### Chapter 3 (k-Nearest-Neighbor Classifier)

**3.1** Consider the set of two-dimensional patterns:

(1, 1, 1), (1, 2, 1), (1, 3, 1), (2, 1, 1), (2, 2, 1), (2, 3, 1), (2, 3.5, 1),

(2.5, 2, 1), (3.5, 1, 1), (3.5, 2, 1), (3.5, 3, 2), (3.5, 4, 2), (4.5, 1, 2)

(4.5, 2, 2), (4.5, 3, 2), (5, 4, 2), (5, 5, 2), (6, 3, 2), (6, 4, 2), (6, 5, 2)

where each pattern is represented by feature 1, feature 2 and the class.

1. If the pattern *P* is (3.8, 3.1), find the class of *P* using the 1-nearest-neighbor algorithm.
2. Find the class of *P* using the *k*-nearest-neighbor algorithm (*k* = 3).

**3.2** Consider the set of two-dimensional patterns:

(1, 1, 1), (1, 2, 1), (1, 3, 1), (2, 1, 1), (2, 2, 1), (2, 3, 1), (2, 3.5, 1),

(2.5, 2, 1), (3.5, 1, 1), (3.5, 2, 1), (3.5, 3, 2), (3.5, 4, 2), (4.5, 1, 2)

(4.5, 2, 2), (4.5, 3, 2), (5, 4, 2), (5, 5, 2), (6, 3, 2), (6, 4, 2), (6, 5, 2)

where each pattern is represented by feature 1, feature 2 and the class.

Find the centroids of the two classes. Use the minimum-distance classifier to find the class of the pattern *P* = (3.8, 3.1).

**3.3** Consider the set of two-dimensional patterns:

(1, 1, 1), (1, 2, 1), (1, 3, 1), (2, 1, 1), (2, 2, 1), (2, 3, 1), (2, 3.5, 1),

(2.5, 2, 1), (3.5, 1, 1), (3.5, 2, 1), (3.5, 3, 2), (3.5, 4, 2), (4.5, 1, 2)

(4.5, 2, 2), (4.5, 3, 2), (5, 4, 2), (5, 5, 2), (6, 3, 2), (6, 4, 2), (6, 5, 2)

where each pattern is represented by feature 1, feature 2 and the class.

a) Find the condensed set using the condensed nearest neighbor algorithm.

b) If the patterns of Class 2 appears first, given the set

(3.5, 3, 2), (3.5, 4, 2), (4.5, 1, 2) (4.5, 2, 2), (4.5, 3, 2), (5, 4, 2), (5, 5, 2),

(6, 3, 2), (6, 4, 2), (6, 5, 2), (1, 1, 1), (1, 2, 1), (1, 3, 1), (2, 1, 1), (2, 2, 1),

(2, 3, 1), (2, 3.5, 1), (2.5, 2, 1), (3.5, 1, 1), (3.5, 2, 1),

find the condensed set using the condensed nearest neighbor algorithm.

**3.4** Consider the set of two-dimensional patterns:

(1, 1, 1), (1, 2, 1), (1, 3, 1), (2, 1, 1), (2, 2, 1), (2, 3, 1), (2, 3.5, 1),

(2.5, 2, 1), (3.5, 1, 1), (3.5, 2, 1), (3.5, 3, 2), (3.5, 4, 2), (4.5, 1, 2)

(4.5, 2, 2), (4.5, 3, 2), (5, 4, 2), (5, 5, 2), (6, 3, 2), (6, 4, 2), (6, 5, 2)

where each pattern is represented by feature 1, feature 2 and the class.

Rank the data set using the Naïve Rank Numerosity Reduction algorithm.

**3.5** State the difference between *prototype selection* and *prototype abstraction.* Is Naïve

Rank Numerosity Reduction algorithm belongs to the category of prototype abstraction?

**3.6** We use a weighted 5-NN classifier to determine the class of object P. Assume that the distances between P and the five nearest neighbors (X1, X2, X3, X4 and X5) be d1 = 1, d2 = 3, d3 = 4, d4 = 5 and d5 =8. If X1, X2 belong to class + and the three others belong to class -. Which class is P assigned to?