

Homework 3

Due: Apr 3

1. Calculate the entropy gain for splitting on each attribute of the following data set. Which attribute would a decision tree for this data select as its root split?

Attributes			Class
f_1	f_2	f_3	c
feathers	omnivore	green	bird
scales	herbivore	green	reptile
feathers	herbivore	brown	bird
fur	omnivore	black	mammal
fur	herbivore	brown	mammal
feathers	carnivore	brown	bird
scales	omnivore	black	reptile
fur	carnivore	black	bird
scales	carnivore	green	reptile

2. The perceptron classifier uses a linear hyperplane as a decision surface to separate binary class data. Give a graphical example of a 2-dimensional data set that is *linearly inseparable*, i.e., a data set for which no possible linear surface can correctly differentiate the positive from negative instances. Now, describe a function in n dimensions that is guaranteed to be linearly inseparable.
3. The Euclidean distance function in \mathbb{R}^n can be written (longhand)

$$d(X, Y) = |X - Y| = \sqrt{\sum_{i=1}^n (X(i) - Y(i))^2}$$

or (linear algebraically):

$$d(X, Y) = |X - Y| = ((X - Y)^T (X - Y))^{1/2}$$

- . Show that the k -NN classifier will still make the same classifications if the square root is omitted.
4. In the programming language of your choice, write a program to implement the k -nearest neighbor learning algorithm for continuous spaces. Your program must be able to accept two data sets, a **training** data set, from which it builds its model, and a **test** data set, which it uses to evaluate the data (i.e., measure classification accuracy). You will find data sets in <http://www.cs.unm.edu/~terran/classes/cs427-s03/data/>

Each file contains one line per data point (a.k.a., instance). Each line is N comma delimited fields. The first $N - 1$ fields are numerical feature values; the N^{th} field is the class label. The class label is a discrete, categorical value and may be a string. For each data file, perform the following operations:

- (a) Randomly shuffle the instances in the data. (Hint: this can be accomplished efficiently with an index array, a random number generator, and an off-the-shelf `sort()` function.)
- (b) Split the shuffled data in half. Use one half for training and the other for testing.
- (c) Train your k -NN classifier on the training data and test its classification accuracy on the test data.
- (d) Repeat steps (a)-(c) 5 times, and calculate the average and standard deviation of the testing accuracies.

Perform the above tests for $k \in \{1, 3, 5, 7\}$.

Deliverables Hand in your complete code and a table giving the mean and std dev accuracy values for your classifier for each data file and each value of k . In addition, answer the following questions:

- (a) Why do you randomly shuffle the instances in step (a) above?
- (b) Why do you repeat the accuracy test 5 times in step (c) above?
- (c) What effect does changing k have?

5. (**Extra Credit**) The Mahalanobis distance function is defined (linear algebraically) by:

$$d_W(X, Y) = ((X - Y)^T W^{-1} (X - Y))^{1/2}$$

where W is a symmetric, positive definite matrix defining a linear transform between two \mathbb{R}^n spaces.

- Show that by choosing W correctly, d can be rendered invariant to nonsingular linear transformations of the data space. (I.e., if you replace X and Y by AX and AY for some symmetric, positive definite matrix A , then d_W remains unchaned.)
- Show that the isodistance surfaces (i.e., constant d_W) under this function are hyper-elliptical in \mathbb{R}^n , degenerating to hyperspheres when $W = I$.
- Extend your program from Question 4 to employ the Mahalanobis distance function with W an arbitrary (square) matrix loaded from a file. Use your result from the first sub-part of this problem to select appropriate W for the data files from Question 4 and derive new accuracy statistics for them under the resulting distance functions. Describe how and why doing this changes your accuracy values and which data sets it has most impact on.