**CS 484 Data Mining**

**Fall 2015**

**Professor Jessica Lin**

**HW 2 – Due 9/24**

**Total: 60 points**

**Question 1 (3 points each = 21 points total)**

For each of the following vectors, x and y, calculate the indicated similarity or the distance measures. Show your work.

1. x = (0 1 0 1 1), y = (1 0 1 0 0) Jaccard, Cosine, Euclidean, Correlation

(c) x= (0, -1, 0, 1), y = (1, 0, -1, 0) Cosine, Euclidean, Correlation

**Question 2 (9 points)**

You are given a set of points S in Euclidean space, as well as the distance of each point in S to a point x. (It does not matter if x ∈ S.) If the goal is to find all points within a specified distance ε of point y, y ≠ x, explain how you could use the triangle inequality and the already calculated distance to x to potentially reduce the number of distance calculations necessary? (Hint: The triangle inequality, d(x, z) ≤ d(x, y) + d(y, z), can be rewritten as d(x, y) ≥ d(x, z) – d(y, z)). Give a simple example.

**Question 3 (20 points)**

Given the following training examples for the target concept EnjoySport .

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Example | Sky | AirTemp | Humidity | Wind | Water | Forecast | EnjoySport |
| 1 | Sunny | Warm | Normal | Strong | Warm | Same | Yes |
| 2 | Sunny | Warm | High | Strong | Warm | Same | Yes |
| 3 | Rainy | Cold | High | Strong | Cool | Change | No |
| 4 | Sunny | Warm | High | Strong | Cool | Change | Yes |
| 5 | Sunny | Warm | Normal | Weak | Warm | Same | No |

Draw the decision tree that will be learnt based on the training examples, using entropy as the impurity measure. Show the value for information gain for each of the candidate attributes for each step in the growing tree. Show all your calculations.

**Question 4 (10 points)**

Consider the following data set for a binary class problem (below)

1. Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

b. Calculate the gain in the Gini index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

|  |  |  |
| --- | --- | --- |
| * A | * B | * Class Label |
| * T | * F | * + |
| * T | * T | * + |
| * T | * T | * + |
| * T | * F | * - |
| * T | * T | * + |
| * F | * F | * - |
| * F | * F | * - |
| * F | * F | * - |
| * T | * T | * - |
| * T | * F | * - |