**CPT\_S 534 HW2**

**Yang Zhang**

**11529139**

1. **(a)**

We define g(y)=y2 and . Since g(y)’’ = 2 which is bigger than 0, therefore g(y) is a convex function. By Jensen’s inequality, g[f(y)] < f[g(y)] holds true for random y and convex function g(y). Thus , then we substitute y with (x-z):

[

And since d is large, so 1/d < 1. Therefore,

Which is equivalent to

**(b)**

We could use the lower bound of the Euclidean distance (the right side of Jensen’s inequality) to estimates the actual Euclidean distance of two random points in R-space. Then the estimation distance between x and z is

Which can be done in linear time instead of matter of polynomial

1. The key idea of LSH is to precompute a hash table that puts points with same probability in the same bucket. For input point p, the algorithm retrieves all the points in the same bucket with p (i.e. points that have the same probability as p). Then looping over the points we retrieved and calculating the distance from p. Finally, records the point if it is a correct answer.
2. It is not always possible to construct a decision tree from set of rules. If the rules in the rule set don’t have intersection (common if-condition), then there is no such common node to connect the tree branch construct from the rule. For example, suppose the the rule-sets has two rules:
3. If today is Sunday and sunny Jayce would play basketball
4. If it is windless and Jayce is happy then he would play basketball

There is no way to adapt a decision based on the two rules.

1. For each of the features (Outlook, Temp, Humidity, Wind) has 9 yes and 5 no

So, the entropy H(Outlook) = H(Temp) = H(Humidity) = H(Wind) =

Note: (x / y) denotes (# of yes / # of no)

The Outlook feature has 3 attributes Sunny (2 / 3), Overcast (4 / 0), Rain (3 / 2)

H(Sunny) =

H(Overcast) =

H(Rain) =

G(Outlook) = H(Outlook) – 5/14H(Sunny) – 4/14H(Overcast) – 5/14H(Rain)

= 0.65– 0.24 – 0 -0.24 =0.17

The Temp feature has 3 attributes Hot (2 / 2), Mild (4 / 2), Cool (3 / 1)

H(Hot) =

H(Mild) =

H(Cool) =

G(Temp) = H(Temp) – 4/14H(Hot) – 6/14H(Mild) – 4/14H(Cool)

= 0.019

The Humidity feature has 2 attributes High (3 / 3), Normal (6 / 2)

H(High) =

H(Normal) =

G(Humidity) = H(Humidity) – 6/14H(High) – 8/14H(Normal)

= 0.034

The Wind feature has 2 attributes Strong (3 / 3), Weak (6 / 2)

H(Strong) =

H(Weak) =

G(Wind) = H(Wind) – 6/14H(Strong) – 8/14H(weak)

= 0.034

Therefore, we pick Outlook feature to split first since it has the max information gain.

**5.**

**(b)** The validation accuracy for Tree 1: 0.6007031642390759

The testing accuracy for Tree 1: 0.6033182503770739

Confusion matrix for test 1:

~0 ~1

0 690 309

1 480 510

The validation accuracy for Tree 2: 0.57

The testing accuracy for Tree 2: 0.5742904841402338

Confusion matrix for test 2:

~0 ~1

0 206 94

1 161 138