CSC411 HW1
Ruijie Sun 1003326046

1.
(a)
Given that X, Y ~ U(0,1), so E(X) = E(Y) = $\frac{1}{2}$, V(X) = V(Y) = $\frac{1}{12}$

E(Z) = E((X-Y)^2) = E(X2 – 2XY + Y2)
= E(X^2) – 2E(XY) + E(Y^2)
= E(X^2) – 2E(X)E(Y) + E(Y^2) (Since X Y are independent)
= V(X) + (E(X))^2 – 2E(X)E(Y) + V(Y) + (E(Y))^2
= $\frac{1}{12} + \frac{1}{4} - \frac{1}{2} + \frac{1}{12} + \frac{1}{4}$
= $\frac{1}{6}$

Var(Z) = Var(((X-Y)^2)
= E((X - Y)^4) – (E(X-Y)^2) ^{2}
= E(X^4) – E(4X^{3}Y) + E(6X^{2}Y^2) – 4E(XY^3) + E(Y^4) - $\frac{1}{36}$
= E(X^4) – 4E(X^3) E(Y) + 6E(X^2)E(Y^2) – 4E(X)E(Y^3) + E(Y^4) - $\frac{1}{36}$

based on moment generating function for uniform distribution:
E(x^n) = $\frac{1}{n+1}\sum_{k=0}^{n} a^k b^{n-k}$
= $\frac{1}{5} - 4 \times \frac{1}{4} \times \frac{1}{2} + 6 \times \frac{1}{3} \times \frac{1}{3} - 4 \times \frac{1}{2} \times \frac{1}{4} + \frac{1}{5} - \frac{1}{36}$
= $\frac{7}{180}$

(b)
E(R) = E(Z_1+Z_2+…+Z_d)
= E(Z_1) + E(Z_2) + …+ E(Z_d)
= $\sum_1^d \frac{1}{6}$
= $\frac{d}{6}$

V(R) = Var(Z_1+Z_2+…+Z_d)
= Var(Z_1) + Var(Z_2) + …+ Var(Z_d) + 0 (since each Z_i are independent)
= $\frac{7d}{180}$

(c)
sd(R) = $\sqrt{\frac{7d}{180}}$
Max(squared Euclidean distance) = d

E(R) = $\frac{d}{6}$

So, based on sd(R), E(R) and Max(squared Euclidean distance), we can know that distance expectation increases linearly with max distance when dimension increases. But sd(R), the volatility, increase slower with rate $\sqrt{d}$. Thus, in high dimensions, most points are far away, and approximately the same distance.

2.

(b)

For the model with Gini coefficient and the tree max depth 2, the validation accuracy is 0.6551020408163265

For the model with Gini coefficient and the tree max depth 3, the validation accuracy is 0.6918367346938775

For the model with Gini coefficient and the tree max depth 4, the validation accuracy is 0.6836734693877551

For the model with Gini coefficient and the tree max depth 5, the validation accuracy is 0.6959183673469388

For the model with Gini coefficient and the tree max depth 6, the validation accuracy is 0.6918367346938775

For the model with Gini coefficient and the tree max depth 7, the validation accuracy is 0.6938775510204082

For the model with Gini coefficient and the tree max depth 8, the validation accuracy is 0.7

For the model with Gini coefficient and the tree max depth 9, the validation accuracy is 0.7040816326530612

For the model with Gini coefficient and the tree max depth 10, the validation accuracy is 0.7081632653061225

For the model with Entropy coefficient and the tree max depth 2, the validation accuracy is 0.6428571428571429

For the model with Entropy coefficient and the tree max depth 3, the validation accuracy is 0.6795918367346939

For the model with Entropy coefficient and the tree max depth 4, the validation accuracy is 0.6857142857142857

For the model with Entropy coefficient and the tree max depth 5, the validation accuracy is 0.6938775510204082

For the model with Entropy coefficient and the tree max depth 6, the validation accuracy is 0.6979591836734694

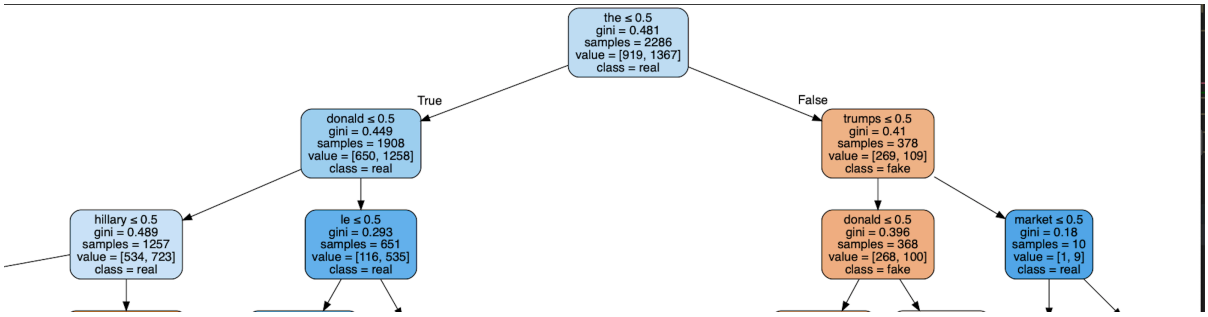
For the model with Entropy coefficient and the tree max depth 7, the validation accuracy is 0.7020408163265306

For the model with Entropy coefficient and the tree max depth 8, the validation accuracy is 0.6938775510204082

For the model with Entropy coefficient and the tree max depth 9, the validation accuracy is 0.6979591836734694

For the model with Entropy coefficient and the tree max depth 10, the validation accuracy is 0.7

(c)



(d)

When topmost splis is the, information gain is 0.047746612890889883

When topmost splis is donald, information gain is 0.04507907092238306

When topmost splis is trumps, information gain is 0.04057601943273359

When topmost splis is hillary, information gain is 0.04594498745640685

When topmost splis is le, information gain is 0.004040655814732008

When topmost splis is market, information gain is 0.0005128201655630882