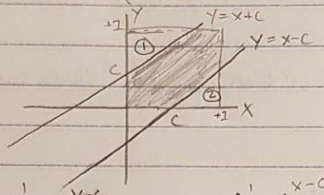
CSC411 Homework 1

Question 1

CSC411 HW1

1a) $P(Z<z) = P((X-Y)^2<z) = P(-\sqrt{z} < X-Y < \sqrt{z})$

$= P(-\sqrt{z} < X-Y \cap X-Y < \sqrt{z})$



Note: indep $\Rightarrow f_{X,Y}(x,y) = 1$

$① + ② = \int_c^1 \int_0^{y-c} 1 \, dx \, dy + \int_c^1 \int_0^{x-c} 1 \, dy \, dx$

$= \int_c^1 y-c \, dy + \int_c^1 x-c \, dx = \frac{1}{2}y^2 - cy \Big|_c^1 + \frac{1}{2}x^2 - cx \Big|_c^1$

$= \frac{1}{2} - c - \frac{1}{2}c^2 + c^2 + \frac{1}{2} - c - \frac{1}{2}c^2 + c^2$

$= 1 - 2c + c^2$

$\therefore$ shaded area: $1 - (1 - 2c + c^2) = 2c - c^2$

$\therefore P(Z<z) = 2\sqrt{z} - z$

Sanity check: $P(Z<1) = 2(1) - 1 = 1$ ✓

$f_Z(z) = 2 \cdot \frac{1}{2} z^{-1/2} - 1 = \frac{1}{\sqrt{z}} - 1$ → undefined at zero though?

More sanity: $\int_0^1 (z^{-1/2} - 1) dz = 2z^{1/2} - z \Big|_0^1 = 1$ ✓ OK...

$E[Z] = \int_0^1 z(z^{-1/2} - 1) dz = \int_0^1 (z^{1/2} - z) dz = \frac{2}{3}z^{3/2} - \frac{1}{2}z^2 \Big|_0^1 = \frac{2}{3} - \frac{1}{2} = \frac{4}{6} - \frac{3}{6} = \frac{1}{6}$

$E[Z^2] = \int_0^1 (z^{3/2} - z^2) dz = \frac{2}{5}z^{3/2} - \frac{1}{3}z^3 \Big|_0^1 = \frac{2}{5} - \frac{1}{3} = \frac{6}{15} - \frac{5}{15} = \frac{1}{15}$

$\therefore \text{Var } Z = \frac{1}{15} - \left(\frac{1}{6}\right)^2 = \frac{1}{15} - \frac{1}{36} = \frac{7}{180}$

1b) $E[R] = E[z_1] + \ldots + E[z_d] = d\,E[z] = \frac{d}{6}$

Fact: if $S = x_1 + \ldots + x_n$, $Var(x) = \sigma$, X independent

$Var(S) = n\sigma$

$\therefore Var(R) = d\,Var(z)$ z_i are indep. b/c each coord indep.

1c) Maximum euclidean (squared) distance: $d \cdot 1$

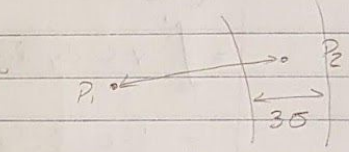
Define "most" points within 3σ:

$3\sqrt{Var(z)}$

$3\sqrt{d\frac{7}{180}}$

As proportion of max dist:

$\frac{3\sqrt{d\frac{7}{180}}}{d}$



$\lim_{d\to\infty} \frac{3\sqrt{d\cdot\frac{7}{180}}}{d} = 0$

Most points will be aprox same dist

What is "far away"? Define as $\frac{E[R]}{R_{max}} = \frac{1}{6}$

points will be $\frac{1}{6}^{th}$ max dist away.

As $d \to \infty$, $E[R] \to \infty$.

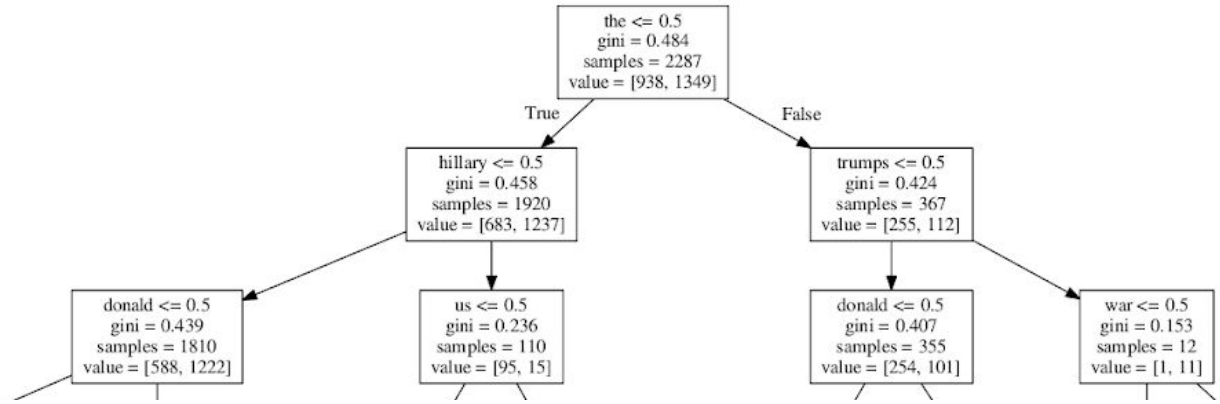
Validation Output

```
Tree with max depth=1, criterion=gini: validation_score=0.697959
Tree with max depth=1, criterion=entropy: validation_score=0.655102
Tree with max depth=2, criterion=gini: validation_score=0.730612
Tree with max depth=2, criterion=entropy: validation_score=0.610204
Tree with max depth=3, criterion=gini: validation_score=0.728571
Tree with max depth=3, criterion=entropy: validation_score=0.681633
Tree with max depth=4, criterion=gini: validation_score=0.732653
Tree with max depth=4, criterion=entropy: validation_score=0.724490
Top trees:
[(-0.7326530612244898, DecisionTreeClassifier(class_weight=None, criterion='gini',
max_depth=4,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best')), (-0.7306122448979592, DecisionTreeClassifier(class_weight=None,
criterion='gini', max_depth=2,
            max_features=None, max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, presort=False, random_state=None,
            splitter='best'))]
Top tree test set score=0.718941
Second tree test set score=0.718941
Information gain by splitting on trump:
0.0304843009658
Information gain by splitting on korea:
0.0182675975452
Information gain by splitting on hillary:
0.0410532561138
Information gain by splitting on the:
0.0457681727899
Information gain by splitting on economic:
1.35102190722e-06
Information gain by splitting on and:
0.00959761721536
Information gain by splitting on election:
0.000185478712508
Information gain by splitting on america:
0.00799189139389
Information gain by splitting on clean:
```

0.00225205884386
Information gain by splitting on black:
0.0112867708304

Tree Visualizations

Tree 1



Tree 2