Homework 1 (CSC411)

Question 1, Part A

$$E(Z) = E[(X-Y)^2] = E[X^2 - 2XY + Y^2]$$
$$= E[X^2] - 2E[XY] + E[Y^2]$$
$$= \int x^2 dx - 2\int\int xy\,dx\,dy + \int y^2 dy$$
$$= \frac{x^3}{3} - 2\int y \frac{x^2}{2} dy + \frac{y^3}{3}$$
$$= \frac{x^3}{3} - x^2\frac{y^2}{2} + \frac{y^3}{3}$$
$$= \frac{1}{6}(2x^3 - 3x^2y^2 + 2y^3)$$

$$\therefore E(Z) = \mu = \frac{1}{6}(2x^3 - 3x^2y^2 + 2y^3)$$

Note: Since variables X & Y are sampled uniformly between 0 to 1, the probability density function for the interval is assumed to be 1, which is obtained via $\frac{1}{b-a}$ where $b=1$ & $a=0$.

$$V(Z) = E(Z^2) - \mu^2$$
$$= E[((X-Y)^2)^2] - \left[\frac{1}{6}(2x^3 - 3x^2y^2 + 2y^3)\right]^2$$
$$= E[(X^2 - 2XY + Y^2)^2] - \left[\frac{1}{36}(4x^6 + 4y^6 - 12x^5y^2 + 8x^3y^3 - 12x^2y^5 + 9x^4y^4)\right]$$

$$= E[x^4 + y^4 + 6x^2y^2 - 4x^3y - 4xy^3] - \left[\frac{1}{9}x^6 + \frac{1}{9}y^6 - \frac{1}{3}x^5y^2 + \frac{1}{4}x^3y^3 - \frac{1}{3}x^2y^5 + \frac{1}{4}x^4y^4\right] \quad (*)$$

$$= \int x^4 dx + \int y^4 dy + 6\iint x^2y^2\,dx\,dy - 4\iint x^3y\,dx\,dy - 4\iint xy^3\,dx\,dy - (*)$$

$$= \frac{x^5}{5} + \frac{y^5}{5} + 6\int \frac{x^3}{3}y^2 dy - 4\int \frac{x^4}{4} y\,dy - 4\int \frac{x^2}{2} y^3 dy - (*)$$

$$= \frac{x^5 + y^5}{5} + 2x^3\frac{y^3}{3} - \frac{x^4y^2}{2} - 4\cdot\frac{x^2}{2}\cdot\frac{y^4}{4} - (*)$$

$$= \frac{1}{5}x^5 + \frac{1}{5}y^5 + \frac{2}{3}y^3x^3 - \frac{1}{2}x^4y^2 - \frac{1}{2}x^2y^4 - \frac{1}{9}x^6 - \frac{1}{9}y^6 + \frac{1}{3}x^5y^2$$
$$- \frac{1}{4}x^3y^3 + \frac{1}{3}x^2y^5 - \frac{x^4y^4}{4}$$

$$= -\frac{x^6}{9} - \frac{y^6}{9} + \frac{x^5}{5} + \frac{y^5}{5} + \frac{5}{12}x^3y^3 + \frac{x^5y^2}{3} + \frac{x^2y^5}{3} - \frac{x^4y^4}{4}$$
$$- \frac{x^4y^2}{2} - \frac{x^2y^4}{2}$$

$$\therefore V(Z) = \sigma^2 = -\frac{x^6}{9} - \frac{y^6}{9} + \frac{x^5}{5} + \frac{y^5}{5} + \frac{5}{12}x^3y^3 + \frac{x^5y^2}{3} + \frac{x^2y^5}{3} - \frac{x^4y^4}{4} - \frac{x^4y^2}{2} - \frac{x^2y^4}{2}$$

Question 1, Part B

$$E[R] = E[Z_1 + \cdots + Z_d]$$
$$= E[Z_1] + \cdots + E[Z_d]$$
$$= \mu_1 + \cdots + \mu_d$$
$$= E[(X_1 - Y_1)^2] + \cdots + E[(X_d - Y_d)^2]$$
$$= \iint (X_1 - Y_1)^2 \, dx_1 dy_1 + \cdots + \iint (X_d - Y_d)^2 \, dx_d dy_d$$
$$= \sum_{d=1}^{d} \iint (X_i - Y_i)^2 \, dx_i dy_i$$

Note 1: $E[Z_1] = E[Z]$ (from Part A) $= \mu$ (denoted μ_1 here)

Note 2: $E[Z_i]$, where $1 \le i \le d$ can be calculated in a similar manner as in Part A using integrals.

$$\therefore E(R) = \sum_{i=1}^{d} \iint (X_i - Y_i)^2 \, dx_i dy_i = \mu_R = d\,E(Z_i) \text{ where } 1 \le i \le d$$

$$V[R] = E[R^2] - (E[R])^2$$
$$= E[(Z_1)^2 + \cdots + (Z_d)^2] - (E[R]^2)$$
$$= E[(Z_1)^2] + \cdots + E[(Z_d)^2] - (E[R])^2$$
$$= E[((X_1 - Y_1)^2)^2] + \cdots + E[((X_d - Y_d)^2)^2] - (E[R])^2$$
$$= \iint ((X_1 - Y_1)^2)^2 \, dx_1 dy_1 + \cdots + \iint ((X_d - Y_d)^2)^2 \, dx_d dy_d - (E[R])^2$$
$$= \sum_{i=1}^{d} \iint ((X_i - Y_i)^2)^2 \, dx_i dy_i - (E[R])^2$$

$$\therefore V(R) = \sum_{i=1}^{d} \iint ((X_i - Y_i)^2)^2 \, dx_i dy_i - \mu_R^2$$
$$= d\,E(Z_i^2) - (d\,E(Z_i))^2$$
$$= \sigma_R^2$$

Question 1, Part C

These calculations support the Curse of Dimensionality in nearest neighbours calculations because as dimensionality increases, the polynomial growth in Expectation and variance values increases exponentially with integral evaluations.

Question 2, Part B:

Output from validation code yields the following:

Information Gain accuracy for max_depth 2: 0.7402862985685071
Gini accuracy for max_depth 2: 0.7402862985685071

Information Gain accuracy for max_depth 3: 0.7402862985685071
Gini accuracy for max_depth 3: 0.7402862985685071

Information Gain accuracy for max_depth 5: 0.7402862985685071
Gini accuracy for max_depth 5: 0.7402862985685071

Information Gain accuracy for max_depth 11: 0.7402862985685071
Gini accuracy for max_depth 11: 0.7402862985685071

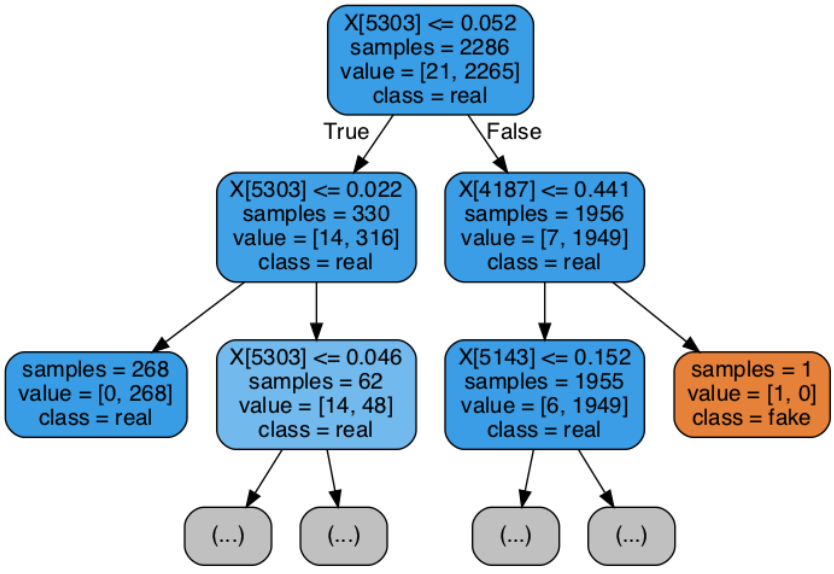
Information Gain accuracy for max_depth 17: 0.7402862985685071
Gini accuracy for max_depth 17: 0.7402862985685071

Question 2, Part C:

Since all accuracies were approximately the same, the following shows the tree with *criterion* Information Gain and *max_depth* of 5:

Note 1: the following values extracted from the vocabulary in the vectorizer for interpretation of the tree:

- X[5303] = “trump”
- X[4187] = “reject”
- X[5143] = “the”



Question 2, Part D:

The following is the output for the word “trump” chosen to be the root:

- Information Gain for word "trump" is: 0.04370012447843627

Here are a few other sample keywords’ computations for Information Gain:

- Information Gain for word "hillary" is: 0.047341783016359024
- Information Gain for word "debate" is: 0.08555126897517264
- Information Gain for word "the" is: 0.06649755612035199