CSC2515 Assignment 1

Parul Saini (parul.saini@mail.utoronto.ca)
Collaborators: Aditya Kharosekar, Subhayan Roy

October 6, 2021

1 Nearest Neighbours and the Curse of Dimensionality

1.1 Consider two independent uni-variate random variables X and Y sampled uniformly from the unit interval [0, 1]. Determine the expectation and variance of the random variable $Z = |X - Y|^2$, i.e., the squared distance between X and Y

Proof:

For any random variable X and Y properties of expectation are as follows:

$$\begin{aligned}
&E[X+Y] = E[X] + E[Y] \\
&E[2X+a] = 2E[X] \quad (where\ a\ is\ constant) \\
&E[XY] = E[X]E[Y]
\end{aligned} \tag{1}$$

The expectation of squared distance between X and Y:

$$\begin{aligned}
&Z = |X-Y|^2 \\
&Z = X^2 + Y^2 - 2XY \\
&E[Z] = E[X^2 + Y^2 - 2XY] \\
&E[Z] = E[X^2] + E[Y^2] - 2E[XY] \\
&E[Z] = E[X^2] + E[Y^2] - 2E[X]E[Y] \quad (using\ 1)
\end{aligned} \tag{2}$$

Since probability distribution function for uniform distribution is given as $\frac{1}{b-a}$ for $a \leq x \leq b$. Given b=1 and a=0

Expectation of X can be calculated as:

$$\begin{aligned}
E[X] &= \int_a^b x f(x) dx \\
&= \int_a^b \frac{x}{b-a} dx
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{b-a}[\frac{x^2}{2}]_a^b \\
&= \frac{1}{b-a}[\frac{b^2}{2} - \frac{a^2}{2}] \\
&= \frac{(b-a)(b+a)}{2(b-a)} \\
&= \frac{(b+a)}{2} = \frac{1}{2} \qquad (3)
\end{aligned}
$$

Similarly we can calculate $E[X^2]$ as follows:

$$
\begin{aligned}
E[X] &= \int_a^b x^2 f(x)dx \\
&= \int_a^b \frac{x^2}{b-a}dx \\
&= \frac{1}{b-a}[\frac{x^3}{3}]_a^b \\
&= \frac{1}{b-a}[\frac{b^3}{3} - \frac{a^3}{3}] \\
&= \frac{(b^3-a^3)}{3(b-a)} = \frac{1}{3} \qquad (4)
\end{aligned}
$$

Substituting (3) and (4) in equation (2), $E[|X-Y|^2]$ is:

$$
\begin{aligned}
E[Z] &= E[X^2] + E[Y^2] - 2E[X]E[Y] \\
&= 2(\frac{(b^3-a^3)}{3(b-a)}) - 2(\frac{(b+a)}{2}) \\
&= 2(\frac{1}{3}) - 2(\frac{1}{2}) \\
&= \frac{2}{3} - \frac{1}{2} \\
&= \frac{1}{6}
\end{aligned}
$$

Variance of a random variable in terms of Expectation can be written as:

$$
\begin{aligned}
Var(X) &= E[X^2] - (E[X])^2 \qquad (5) \\
&= \frac{(b^3-a^3)}{3(b-a)} - (\frac{(b+a)}{2})^2 \qquad (using\ 3\ and\ 4) \\
&= \frac{1}{3} - \frac{1}{4} \\
&= \frac{1}{12}
\end{aligned}
$$

Since we can calculate the Expectation for X and Y sampled from the unit interval [0, 1] as follows:

$$
\begin{aligned}
E[X] &= \int_a^b x f(x)dx = \frac{1}{2} \qquad (using\ 3)\\
E[X^2] &= \int_a^b x^2 f(x)dx = \frac{1}{3} \qquad (using\ 4)\\
E[X^3] &= \int_a^b x^3 f(x)dx = \frac{1}{4}\\
E[X^4] &= \int_a^b x^4 f(x)dx = \frac{1}{5}
\end{aligned}
$$

Therefore using properties from (2) , (5) we can write $Var[|X-Y|^2]$ in terms of Expectation as:

$$
\begin{aligned}
&Var[|X-Y|^2] = E[|X-Y|^4] - (E[|X-Y|^2])^2 \qquad (using\ 5)\\
&= E[X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4] - (\frac{1}{6})^2 \qquad (using\ E[|X-Y|^2] = \frac{1}{6}\ from\ above)\\
&= E[X^4] - 4E[X^3]E[Y] + 6E[X^2]E[Y^2] - 4E[X]E[Y^3] + E[Y^4] - (\frac{1}{36})\\
&= \frac{1}{5} - 4(\frac{1}{4})(\frac{1}{2}) + 6(\frac{1}{3})(\frac{1}{3}) - 4(\frac{1}{2})(\frac{1}{4}) + \frac{1}{5} - (\frac{1}{36})\\
&= \frac{2}{5} + \frac{2}{3} - 1 - (\frac{1}{36})\\
&= \frac{1}{15} - (\frac{1}{36})\\
&= \frac{21}{540}\\
&= \frac{7}{180}
\end{aligned}
$$

1.2 Using the properties of expectation and variance, determine $E||X-Y||_2^2 = E[R]$ and $Var||X-Y||_2^2 = Var[R]$ for two d-dimensional points X and Y sampled from a d-dimensional unit cube with a uniform distribution

Proof:

L_2-Norm (Euclidean Distance) for variables sampled in a d dimensional space can be written as:

$$
\begin{aligned}
L_2(X,Y) &= \sqrt{|X_i - Y_i|^2 + ... + |X_d - Y_d|^2}\\
L_2(X,Y)^2 &= |X_i - Y_i|^2 + ... + |X_d - Y_d|^2
\end{aligned}
$$

Since $Z_i = |X_i - Y_i|^2$

$$\begin{aligned} & ||X-Y||_2^2 = R \ (given\ ) \\ & = Z_i + ... + Z_d \\ & = |X_i - Y_i|^2 + ... + |X_d - Y_d|^2 \\ & = Z_i + ... + Z_d \end{aligned} \tag{6}$$

$$\begin{aligned} & E||X-Y||_2^2 = E[R] \\ & = E[Z_i] + ... + E[Z_d] \\ & = dE[Z] \\ & = \frac{d}{6} \ (using\ E[Z] = \frac{1}{6}\ calculated\ in\ section\ 1.1) \end{aligned}$$

Similarly for Variance, using (6)

$$\begin{aligned} & Var||X-Y||_2^2 = Var[R] = Var[Z_i + ... + Z_d] \\ & = Var[Z_i] + ...Var[Z_d] \\ & = dVar[Z] \\ & = \frac{7d}{180} \ (using\ Var[Z] = \frac{7}{180}\ calculated\ in\ section\ 1.1) \end{aligned}$$

1.3 Compare the mean and standard deviation of $||X - Y||^2$ to the maximum possible squared Euclidean distance between two points within the d-dimensional unit cube

Proof:

Maximum possible squared Euclidean distance between two points within the d-dimensional unit cube is given as:

$$\begin{aligned} & L_2(X,Y) = \sqrt{|X_i - Y_i|^2 + ... + |X_d - Y_d|^2} \\ & = \sqrt{|1-0|^2 + ... + |1-0|^2} \\ & = \sqrt{d} \end{aligned}$$

Therefore, $L_2(X,Y)^2 = d$

Mean $(\mu) = E[R] = E||X-Y||_2^2 = \frac{d}{6}$ (using $E[R] = \frac{d}{6}$ calculated in section 1.2)

Standard Deviation $(\sigma) = \sqrt{V} = \sqrt{\frac{7d}{180}} = 0.2\sqrt{d}$ (using $Var[R] = \frac{7d}{180}$ calculated in section 1.2)

The substantial difference of 0.83d $(d - \frac{d}{6})$ between the mean distance of $||X - Y||^2$ and the maximum possible squared Euclidean distance between two points within the d-dimensional unit cube suggests that most of the points are far away. In addition to this a low value of standard deviation $(0.2\sqrt{d})$ suggests that data points are clustered around the mean thereby resulting in data points being positioned at approximately the same distance.

2 Information Theory

Definition of the entropy of a discrete random variable X with probability mass function p:

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} \tag{7}$$

The summation is over all possible values of x ϵ X , which (for simplicity) we assume is finite. For example, X might be {1,2,...,N}

2.1 Prove that the entropy H(X) is non-negative.

Proof :

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)}$$

since $\log_a \frac{1}{b} = -\log_a b$

$$\Rightarrow H(X) = -\sum_x p(x) \log_2 p(x)$$

since $0 \leq p(x) \leq 1$

$$\begin{aligned}
&\Rightarrow \log_2 p(x) \leq 0 \\
&\Rightarrow -\log_2 p(x) \geq 0 \\
&\Rightarrow -\sum_x p(x) \log_2 p(x) \geq 0 \\
&\Rightarrow H(x) \geq 0
\end{aligned}$$

2.2 If X and Y are independent random variables, show that H(X, Y) = H(X) + H(Y)

Proof :

Joint entropy is given as

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$

Using conditional probability: $P(A|B) = \frac{P(A,B)}{P(B)}$ if $P(B) \neq 0$

$$= -\sum_x \sum_y p(x,y) \log p(x)p(y|x))$$

Since x , y are independent random variables $P(X|Y) = P(X), P(Y|X) = P(Y), P(X,Y) = P(X)P(Y)$, therefore

$$= -\sum_x \sum_y p(x,y) \log p(x) - \sum_x \sum_y p(x,y) \log p(y|x))$$
$$= -\sum_x p(x) \log p(x) - \sum_y \sum_x p(x,y) \log p(y))$$
$$= -\sum_x p(x) \log p(x) - \sum_y p(y) \log p(y))$$
$$= H(X) + H(Y)$$

Hence Proved that $H(X,Y) = H(X) + H(Y)$ using (7)

2.3 Prove the chain rule for entropy: $H(X,Y) = H(X) + H(Y|X)$

Proof :

Conditional probability:

$$P(A|B) = \frac{P(A,B)}{P(B)} \tag{8}$$

Conditional entropy of Y given another random variable X is given as

$$H(Y|X) = \sum_x p(x)\ H(Y|X=x)$$
$$= -\sum_x p(x) \sum_y p(y|x)\ \log p(y|x)$$

$$= -\sum_x \sum_y p(x,y) \log p(y|x) \quad (using\ 8) \tag{9}$$

Joint entropy is given as

$$H(X,Y) = -\sum_x \sum_y p(x,y) \log p(x,y)$$
$$= -\sum_x \sum_y p(x,y)\ [\log p(x) p(y|x)] \quad (using\ 8)$$

Using $\log_a uv = \log_a u + \log_a v$

$$= -\sum_x \sum_y p(x,y)\ \log p(x) - \sum_x \sum_y p(x,y) \log p(y|x)$$
$$= -\sum_x p(x)\ \log p(x) - \sum_x \sum_y p(x,y) \log p(y|x) \quad (using \sum_y p(x,y) = p(x))$$
$$= H(X) + H(Y|X) \quad (using\ 9)$$

Hence Proved that $H(X,Y) = H(X) + H(Y|X)$

2.4 Prove that $KL(p||q)$ is non-negative.

Proof :

Relative entropy or the KL-divergence of two distributions p and q , $p(x) > 0, q(x) > 0$ for all x is defined as:

$$= \sum_x p(x) \log_2 \frac{p(x)}{q(x)} \tag{10}$$

let $Z = \frac{q(x)}{p(x)}, \log \frac{q(x)}{p(x)} = \log Z$

Using Jensen's Inequality, log(x) is concave on the set of positive real numbers:

$$E[\log Z] \leq \log E[Z]$$

$$\begin{aligned}
&- KL(p||q) \leq - \sum_x p(x) \log \frac{p(x)}{q(x)} \\
&\leq \log \sum_x p(x) \frac{q(x)}{p(x)} \\
&\leq \log \sum_x q(x) \\
&\leq \log 1 \\
&\leq 0 \\
&- KL(p||q) \leq 0 \\
&.KL(p||q) \geq 0
\end{aligned}$$

Hence Proved that $KL(p||q)$ is non-negative.

2.5 Show that $I(Y;X) = KL(p(x,y)||p(x)p(y))$, where p(x) is the marginal distribution of X and p(y) is the marginal distribution of Y

Proof :

Relative entropy or the KL-divergence of two distributions p and q is given as:

$$KL(p(x,y)||p(x)p(y)) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y))} \quad (using\ 10) \tag{11}$$

The Information Gain or Mutual Information between X and Y is

$$I(Y;X) = H(Y) - H(Y|X)$$

$$= -\sum_y p(y)\ \log p(y) - (-\sum_x \sum_y p(x,y) \log p(y|x)) \qquad (using\ 9)$$

$$= -\sum_y \sum_x p(x,y)\ \log p(y) + \sum_x \sum_y p(x,y) \log p(y|x)$$

$$= \sum_x \sum_y p(x,y)\ [\log p(y|x) - \log p(y)]$$

Using $\log_a \frac{u}{v} = \log_a u - \log_a v$

$$= \sum_x \sum_y p(x,y)\ \log \frac{p(y|x)}{p(y)}$$

$$= \sum_x \sum_y p(x,y)\ \log \frac{p(x,y)}{p(x)p(y)} \qquad (using\ 8)$$

$$= KL(p(x,y)||p(x)p(y)) \qquad (using\ 11)$$

Hence Proved that $I(Y;X) = KL(p(x,y)||p(x)p(y))$

3. Decision Trees and K-Nearest Neighbor

a) Function load_data() performs the following:
   - Read text files
   - Preprocess data and vectorize using **CountVectorizer**
   - Split data using **train_test_split** function of scikit-learn

```
Output of function load_data():

Total Record Count: 3266
% Record Count for Training: 70.0 % =  2286
% Record Count for Test: 15.0 % = 490
% Record Count for Validation: 15.0 % = 490
```

b)

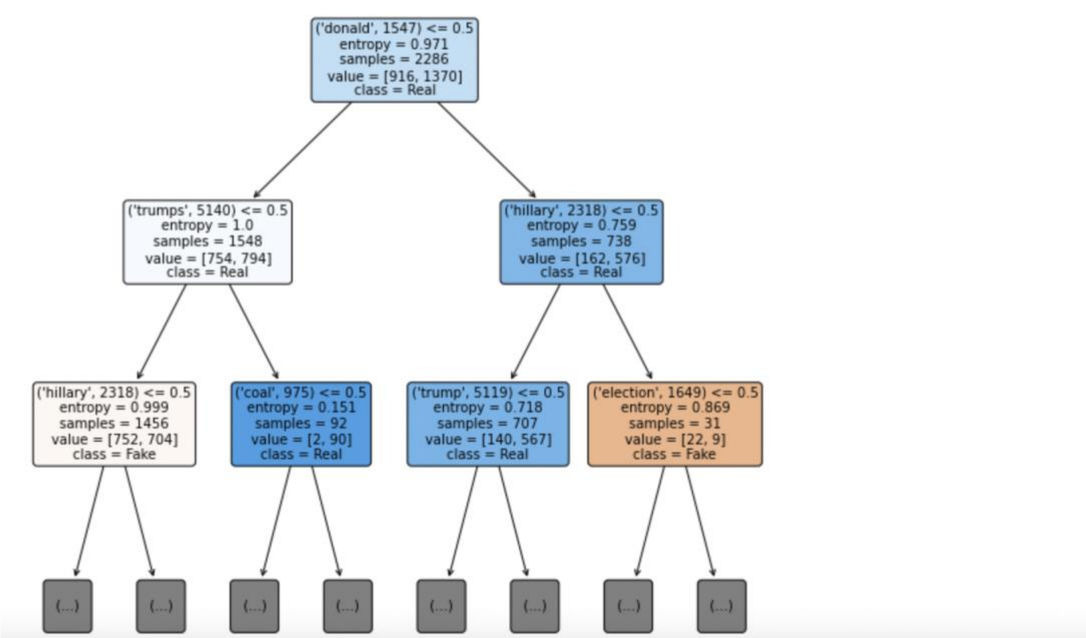
```
Output of function select_tree_model():

Hyperparameters and their corresponding accuracies for Decision Trees:
   Criteria  Max Depth  Accuracy
0  entropy        20      68.6
1     gini        20      68.2
2  entropy        30      68.6
3     gini        30      70.8
4  entropy        50      72.0
5     gini        50      71.6
6  entropy        70      72.7
7     gini        70      71.8
8  entropy        90      71.8
9     gini        90      72.4
```

c)

```
Hyperparameters with highest accuracy for Validation Set: Split Criteria: entropy , Depth:70
Accuracy for DecisionTree, Test Set:  75.9 %

Decision Tree extract of first two layers:
```



d)

```
Output of function compute_information_gain():

Information Gain for the word "donald" is 0.04913422625696717
Information Gain for the word "hillary" is 0.0443445873158429
Information Gain for the word "trumps" is 0.04500636360104682
Information Gain for the word "coal" is 0.00012717419536312224
Information Gain for the word "election" is 0.0013849072273082186
```

e)

```
Output of function select_knn_model():

Max Validation Accuracy --> Best k: 12
Min Validation Error from Graph --> Best k: 12
Accuracy for KNN with best k, Test Set:  71.4 %
```