# CSC311 Assignment 1

Chen Liang

Sept 30, 2019

## Question 1

**(a)** Knowing that the pdf for uniform distribution on [0, 1] is $f(x) = \frac{1}{1-0} = 1$,
we have $E(X) = E(Y) = \int_0^1 x dx = \frac{1}{2}$,
$E(X^2) = E(Y^2) = \int_0^1 x^2 dx = \frac{1}{3}$
$E(X^3) = E(Y^3) = \int_0^1 x^3 dx = \frac{1}{4}$
$E(X^4) = E(Y^4) = \int_0^1 x^4 dx = \frac{1}{5}$
$Var(X) = E(X^2) - (E(X)^2) = \frac{1}{3} - (\frac{1}{2})^2 = \frac{1}{12}$
$Var(X^2) = E(X^4) - (E(X^2))^2 = \frac{1}{5} - (\frac{1}{3})^2 = \frac{4}{45}$

**Expectation:**
Since $Z = (X - Y)^2$, we could denote $E(Z)$ as:
$E(Z) = E((X - Y)^2) = E(X^2 - 2XY + Y^2)$.
Because of the linearity of expectation, we have,
$E(Z) = E(X^2 - 2XY + Y^2) = E(X^2) + E(Y^2) - 2E(XY)$.
Because $X, Y$ are two independent variables, we have,
$E(Z) = E(X^2) + E(Y^2) - 2E(XY) = E(X^2) + E(Y^2) - 2E(X)E(Y) = \frac{1}{3} + \frac{1}{3} - 2 \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{2}{3} - \frac{1}{2} = \frac{1}{6}$.

**Variance:**
$Var(Z) = E(Z^2) - (E(Z))^2$. Now we calculate $E(Z^2)$ and $E(Z)$ respectively.
$E(Z^2) = E((X - Y)^4) = E(X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4)$,
because of the linearity of expectation, we have,
$E(Z^2) = E(X^4 - 4X^3Y + 6X^2Y^2 - 4XY^3 + Y^4)$
$= E(X^4) - E(4X^3Y) + E(6X^2Y^2) - E(4XY^3) + E(Y^4)$,
and because of the fact that $X, Y$ are two independent variables, we have,
$E(Z^2) = E(X^4) - E(4X^3Y) + E(6X^2Y^2) - E(4XY^3) + E(Y^4)$
$= E(X^4) - 4E(X^3)E(Y) + 6E(X^2)E(Y^2) - 4E(X)E(Y^3) + E(Y^4)$
Plug in the values we calculated above, we have,
$E(Z^2) = \frac{1}{5} - 4 \cdot \frac{1}{4} \cdot \frac{1}{2} + 6 \cdot \frac{1}{3} \cdot \frac{1}{3} - 4 \cdot \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{5}$
$= \frac{2}{5} - 1 + \frac{2}{3} = \frac{1}{15}$
Hence, we could calculate $Var(Z)$ as,
$Var(Z) = E(Z^2) - (E(Z))^2 = \frac{1}{15} - (\frac{1}{6})^2 = \frac{1}{15} - \frac{1}{36} = \frac{12}{180} - \frac{5}{180} = \frac{7}{180}$

**(b)** Since we sample two points independently from a unit cube in d dimension, and random variables $X_1, X_2, ...X_d$, and $Y_1, Y_2, ...Y_d$ are all sampled independently from $[0, 1]$. Knowing that $Z_i = (X_i - Y_i)^2$ for i in range $[1, d]$, we could conclude that each pair of $Z_i, Z_j$ for $i, j$ in range $[1, d]$, and $i \neq j$ are independent from each other.
**Expectation:**
$E(R) = E(Z_1 + Z_2 + ... + Z_d) = E(\sum_{n=1}^{d} Z_i)$,

because of the linearity of expectation, we have,
$E(R) = E(\sum_{i=1}^{d} Z_i) = \sum_{i=1}^{d} E(Z_i)$.
since $X_i$ and $Y_i$ are independent for all $i$ in range of $[1, d]$, we could say that $E(Z_1) = E(Z_2) = ...E(Z_d) = \frac{1}{6}$.
Therefore, $E(R) = \sum_{i=1}^{d} E(Z_i) = d \cdot \frac{1}{6} = \frac{d}{6}$.

**Variance:** According to the equation of variance, we have:
$Var(R) = Var(Z_1 + Z_2 + ... + Z_d) = Var(Z_1) + Var(Z_2) + ... + Var(Z_d) + 2Cov(Z_1, Z_2) + ... + 2Cov(Z_1, Z_d) + ...2Cov(Z_{d-1}, Z_d) = \sum_{i=1}^{d} Var(Z_i) + \sum_{k=1}^{d-1} \sum_{j=k+1}^{d} Cov(Z_k Z_j)$.
Since we've shown that $Z_i Z_j$ are independent for every single pair of $i, j$ in $[1, d], i \neq j$.
Then we could draw the conclusion that $\sum_{k=1}^{d-1} \sum_{j=k+1}^{d} Cov(Z_k Z_j) = 0$. In this way,
$Var(R) = \sum_{i=1}^{d} Var(Z_i)$.
Similar to the analysis for expectation, since $X_i$ and $Y_i$ are independent for $i$ in range of $[1, d]$, we could say that $Var(Z_1) = Var(Z_2) = ... = Var(Z_d) = \frac{7}{180}$.
Therefore, $Var(R) = \sum_{i=1}^{d} Var(Z_i) = d \cdot \frac{7}{180} = \frac{7d}{180}$.

**(c)** Here's a detailed explanation for why when $d$ becomes a relatively large number, most points are far away and approximately the same distance:
**Most points are far away:** According to question 2, the expectation for $R$ is: $E(R) = \frac{d}{6}$, so when $d$ approaches infinity, $E(R)$ approaches infinity as well. In this way, we could explain why in high dimensions, most points are far away.
**Most points are approximately the same distance:** Also according to question 2, the variance for $R$ is: $Var(R) = \frac{7d}{180}$, therefore, the standard deviation is $\sigma(R) = \sqrt{d} \cdot \sqrt{\frac{7}{180}}$, so standard increases at rate $\sqrt{d}$, while dimension increases at rate $\sqrt{d}$.

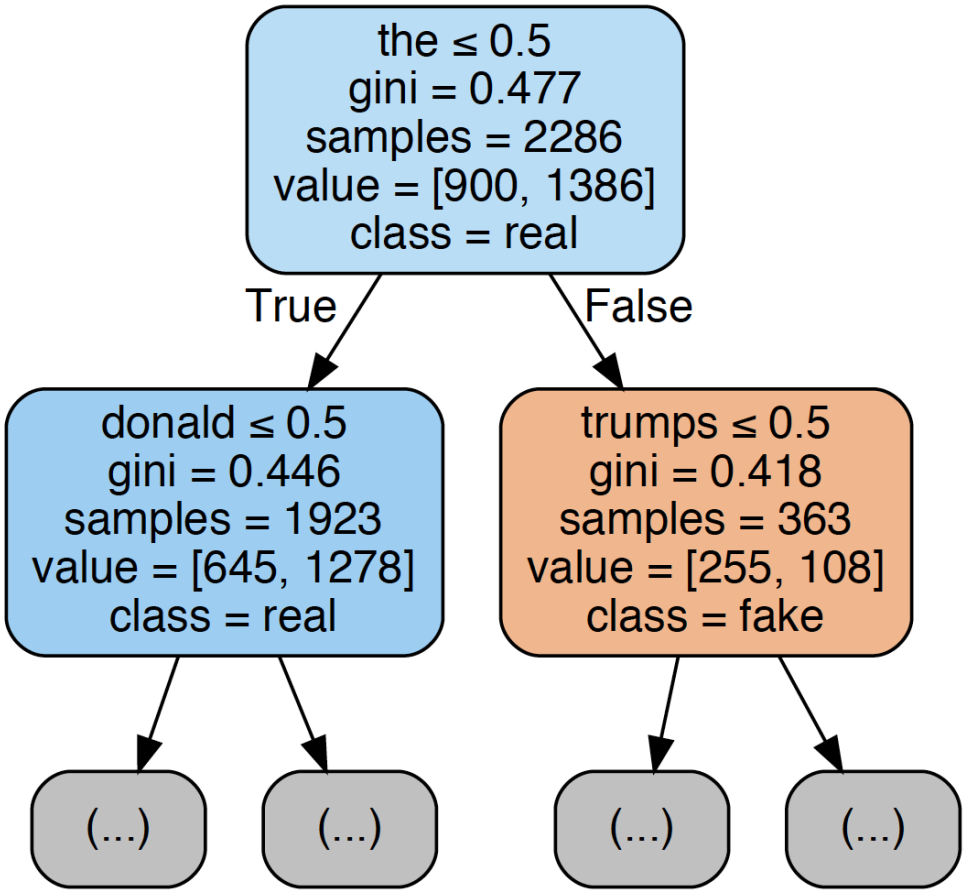
Now we denote the result of dividing the standard deviation by expectation as the percentage change that could happen to the average distances between two points, which is $\frac{\sigma(R)}{E(R)} = \frac{\sqrt{d} \cdot \sqrt{\frac{7}{180}}}{\frac{d}{6}}$, which is linear to $\frac{1}{\sqrt{d}}$. As d becomes relatively large or even approaches infinity, $\frac{1}{\sqrt{d}}$ approaches zero. In this way, for $d$ approaching infinity, $\frac{\sigma(R)}{E(R)}$ approaches zero, which means compare to the mean distance between two points, the standard deviation is so small and could be neglected. In this way, we could say that most points are approximately the same distance.

## Question 2

(b)

```
entropy  with max depth  2  has an accuracy of  0.6489795918367347
entropy  with max depth  3  has an accuracy of  0.6918367346938775
entropy  with max depth  4  has an accuracy of  0.6979591836734694
entropy  with max depth  5  has an accuracy of  0.6959183673469388
entropy  with max depth  6  has an accuracy of  0.6918367346938775
gini  with max depth  2  has an accuracy of  0.6612244897959184
gini  with max depth  3  has an accuracy of  0.6918367346938775
gini  with max depth  4  has an accuracy of  0.6959183673469388
gini  with max depth  5  has an accuracy of  0.7020408163265306
gini  with max depth  6  has an accuracy of  0.7
```

(c)


the ≤ 0.5
gini = 0.477
samples = 2286
value = [900, 1386]
class = real
True
False
donald ≤ 0.5
gini = 0.446
samples = 1923
value = [645, 1278]
class = real
trumps ≤ 0.5
gini = 0.418
samples = 363
value = [255, 108]
class = fake
(...)
(...)
(...)
(...)

# Question 3.1

(a) Because $t|(X,w)$ follows the normal distribution with mean $Xw$ and variance $\sigma^2 I$, then it's PDF is: $p(t_i|(X_{(i)},w)) = \frac{1}{\sqrt{\sigma^2 I}\sqrt{2\pi}} e^{-\frac{(t_i - X_i w)^T (t_i - X_i w)}{2\sigma^2 I}}$, then we could calculate log likelihood $log(f(t_i|(X_i,w)))$ as:
$log(p(t_i|X_i,w)) = log(\frac{1}{\sqrt{\sigma^2 I}\sqrt{2\pi}} e^{-\frac{(t_i - X_i w)^T (t_i - X_i w)}{2\sigma^2 I}}) = -log(\sqrt{\sigma^2 I}\sqrt{2\pi}) - \frac{(t_i - X_i w)^T (t_i - X_i w)}{2\sigma^2 I}$.
Then we have, $\sum_{i=1}^{n} log(p(t_i|X_i,w)) = \sum_{i=1}^{n} -log(\sqrt{\sigma^2 I}\sqrt{2\pi}) - \frac{(t_i - X_i w)^T (t_i - X_i w)}{2\sigma^2 I}$
$= -\sum_{i=1}^{n} log(\sqrt{\sigma^2 I}\sqrt{2\pi}) - \sum_{i=1}^{n} \frac{(t_i - X_i w)^T (t_i - X_i w)}{2\sigma^2 I}$,

$\frac{\partial \sum_{i=1}^{n} log(p(t_i|X_i,w))}{\partial w} = \sum_{i=1}^{n} \frac{-t_i^T X_i - t_i X_i^T + X_i^T X_i w}{2\sigma^2 I} = \frac{-2tX^T + XX^T w}{2\sigma^2 I} = \frac{-tX^T + XX^T w}{\sigma^2 I}$.
If we set $\frac{\partial \sum_{i=1}^{n} log(p(t_i|X_i,w))}{\partial w}$ equals to zero, then
$-tX^T + XX^T w = 0$,
$wXX^T = X^T t$,
$w = (XX^T)^{-1} X^T t$.
Hence, the estimator $\hat{w} = (XX^T)^{-1} X^T t$.

(b) **Expectation:**
$E(\hat{w}) = E((XX^T)^{-1} X^T t)$,
since $(XX^T)^{-1} X^T$ is a matrix, then we could rewrite the expectation as,
$E(\hat{w}) = (XX^T)^{-1} X^T E(t)$.
Given that $t|(X,w) \sim \mathcal{N}(Xw, \sigma^2 I)$ we could state that $E(t) = Xw$.
In this way, $E(\hat{w}) = (XX^T)^{-1} X^T Xw = w$.

**Covariance matrix:**
$Var(\hat{w}) = Var((XX^T)^{-1} X^T t)$
According to the property of multivariate Gaussian random vectors,
$Var(\hat{w}) = (XX^T)^{-1} X^T Var(t)((XX^T)^{-1} X^T)^T$
Given that $t|(X,w) \sim \mathcal{N}(Xw, \sigma^2 I)$ we could state that $Var(t) = \sigma^2 I$.
In this way, $Var(\hat{w}) = (XX^T)^{-1} X^T \sigma^2 I((XX^T)^{-1} X^T)^T = \sigma^2 (XX^T)^{-1} X^T ((XX^T)^{-1} X^T)^T$
$= \sigma^2 (XX^T)^{-1} X^T X (XX^T)^{-1} = \sigma^2 (XX^T)^{-1}$,

As we know that $\hat{(w)}$ follows normal distribution with expectation as $E(\hat{w}) = w$, and variance as $Var(\hat{w}) = \sigma^2 (XX^T)^{-1}$, we could conclude that $\hat{w} \sim \mathcal{N}(w, \sigma^2 (XX^T)^{-1})$

## Question 3.2

Since $\hat{w} = argmax\{p(w|X,t) \propto p(t|X,w)p(w|X)\}$, to find such $w$ to make $p(w|X,t)$ maximum equals to make $p(t|X,w)p(w|X)$ maximum, since $p(w|X,t) \propto p(t|X,w)p(w|X)$, $p(w|X,t) = k \cdot p(t|X,w)p(w|X)$, where $k$ is a constant number. In this way, we find take log-likelihood on $p(w|X,t) = kp(t_i|X_i,w)p(w|X_i)$:
$log(p(w|X,t)) = log(k \cdot p(t_i|X_i,w)p(w|X_i)) = log(k \cdot \frac{1}{\sqrt{\sigma^2 I}\sqrt{2\pi}} e^{-\frac{(t_i - X_i w)^T (t_i - X_i w)}{2\sigma^2 I}} \cdot \frac{1}{\sqrt{\tau^2 I}\sqrt{2\pi}} e^{-\frac{w^T w}{2\tau^2 I}})$
$= \frac{1}{2\sigma^2}(t_i - w^T x_i)^2 - \frac{1}{2\tau^2} w_i^2 + C$, where $C$ is constant.
Therefore, $log(p(w|X,t)) = \frac{1}{2\sigma^2}\sum_{i=1}^{n}(t_i - w^T x_i)^2 - \frac{1}{2\tau^2}\sum_{i=1}^{n} w_i^2 + C = \frac{1}{2\sigma^2}(t_- Xw)^T(t - Xw) - \frac{1}{2\tau^2} w^T w + C$, where $C$ is constant.
In this way, $\frac{\partial log(p(w|X,t))}{\partial \hat{w}} = \frac{1}{\sigma^2} X^T(t - Xw) - \frac{1}{\tau^2} w = 0$, then we have, $\frac{\sigma^2}{\tau^2} w = X^T(t - Xw)$,
$\frac{\sigma^2}{\tau^2} w + X^T X w = X^T t$, then
$\hat{w_{MAP}} = (X^T X + \frac{\sigma^2}{\tau^2} I)^{-1} X^T y$
Hence, we have estimator $\hat{w_{MAP}} = (X^T X + \frac{\sigma^2}{\tau^2} I)^{-1} X^T y$

## Question 3.3

Anlysis: According to the graph, we know that the optimal $\lambda$ for both 10 fold cross validation and 5 fold cross validation would be 0.322.
Comment on shapes: For $\lambda << 0.32$ the error scale is large, because the function is unfitting. When $\lambda >> 0.32$, the error scale is relatively high because the model loses flexibility and cannot fit the training data as we want, which would eventually decrease its accuracy.