DSC 140A - Homework 01

Due: Wednesday, January 17

Write your solutions to the following problems by either typing them up or handwriting them on another piece of paper. Unless otherwise noted by the problem's instructions, show your work or provide some justification for your answer. Homeworks are due via Gradescope at 11:59 PM.

Problem 1.

In practice, the performance of nearest neighbor predictors often decreases with the number of features. This is often attributed to the so-called "curse of dimensionality". One informal statement of the curse goes: "in high dimensions, almost all points in a randomly-drawn set of points are essentially equidistant from the origin."

In this problem, you'll demonstrate this empirically. For each value in a sequence of increasing d (for example, $d = 2, 4, 8, 16, \ldots$), generate a data set of 1,000 points in $\mathbb{R}^d$, where each coordinate of each point is drawn from the uniform distribution on the interval $[-1, 1]$. That is, for any given d, your data set should consist of $1,000$ draws from the uniform distribution on the d-dimensional hypercube $[-1, 1]^d$.

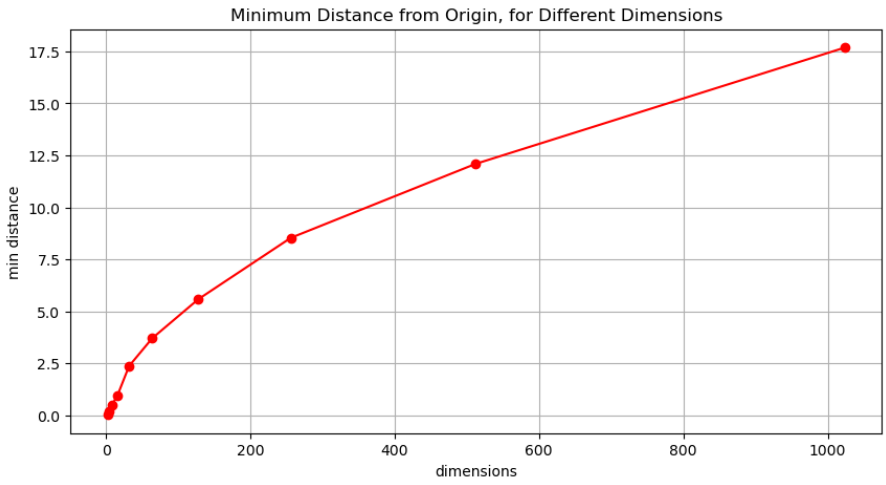
Use your datasets to generate the following plots. You can use whichever programming language you like, but paste your code in your solution. Lastly, it's up to you to decide how large your sequence of d should be – just make sure it's large enough to see the trend.

a) Let $\Delta_0(d)$ be the distance of the **closest** point to the origin in your data set of dimensionality d. Plot $\Delta_0(d)$ as a function of d.

Code

```
import numpy as np
import matplotlib.pyplot as plt

dimensions = [2, 4, 8, 16, 32, 64, 128, 256, 512, 1024]

min_dist = []
max_dist = []

for d in dimensions:
    data = np.random.uniform(-1, 1, (1000, d))

    dist = np.linalg.norm(data, axis=1)

    min_dist.append(np.min(dist))
    max_dist.append(np.max(dist))

plt.figure(figsize=(10, 5))
plt.plot(dimensions, min_dist, label='min', marker='o', color='r')
plt.xlabel('dimensions')
plt.ylabel('min distance')
plt.title('Minimum Distance from Origin, for Different Dimensions')
plt.grid()
plt.show()
```

Solution:



b) Let $\Delta_1(d)$ be the distance from the origin to the **furthest** point in your data set of dimensionality d. Plot the ratio

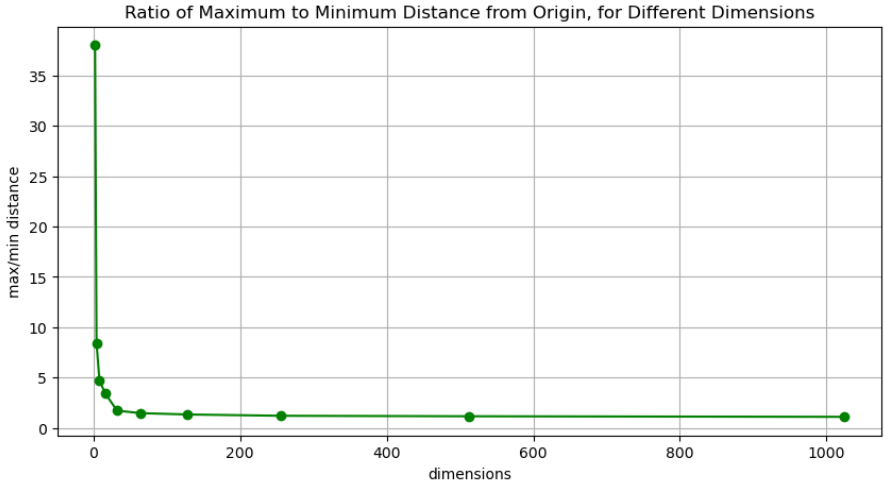
$$\frac{\Delta_1(d)}{\Delta_0(d)}$$

for your sequence of increasing d.

Code

```
ratio = np.array(max_dist) / np.array(min_dist)
plt.figure(figsize=(10, 5))
plt.plot(dimensions, ratio, label='ratio', marker='o', color='g')
plt.xlabel('dimensions')
plt.ylabel('max/min distance')
plt.title('Ratio of Maximum to Minimum Distance from Origin, for Different Dimensions')
plt.grid()
plt.show()
```

Solution:



Problem 2.

In lecture, we derived the least squares solutions for linear prediction rules $H(x) = w_1 x + w_0$. They were:

$$w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$w_0 = \bar{y} - w_1 \bar{x}$$

Where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

You may see these solutions written in various equivalent forms. In this problem, we'll derive another form that you may find useful in solving other problems.

a) Show that $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$.

Solution: First we expand the sum:

$$\begin{aligned}\sum_{i=1}^{n}(x_i - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + \cdots + (x_n - \bar{x}) \\ &= x_1 + x_2 + \cdots + x_n - n\bar{x} \\ &= (x_1 + x_2 + \cdots + x_n) - \frac{n}{n}\sum_{i=1}^{n} x_i, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \\ &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \\ &= 0\end{aligned}$$

This can also be thought about intuitively, as the sum of the values of x_i is the same as the sum of the mean of x_i. Thus, the difference between the two sums is 0.

b) Use the result of the previous part to show that

$$w_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

is equivalent to the formula for w_1 that was given in lecture.

Solution: Lets equate the two equations for w_1:

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

$$\sum_{i=1}^{n}(x_i - \bar{x})y_i - \sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

We can now use the result from part (a) to simplify the equation:

$$\sum_{i=1}^{n}(x_i - \bar{x})y_i - \sum_{i=1}^{n}(x_i - \bar{x})\bar{y} = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

$$\sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}\sum_{i=1}^{n}(x_i - \bar{x}) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

$$\sum_{i=1}^{n}(x_i - \bar{x})y_i - \bar{y}(0) = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

$$\sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n}(x_i - \bar{x})y_i$$

Thus, we have shown that the two equations are equivalent.

Problem 3.

A *Boolean feature* is one that is either true or false. For example, not the car has an automatic transmission. We can perform least squares regression with Boolean features by "encoding" true and false as numbers: a common choice is to encode true as 1 and false as 0.

In this problem, suppose we have a data set $(x_1, y_1), \ldots, (x_n, y_n)$ of n cars, where the feature x_i is either 1 or 0 (has automatic transmission, or does not) and where y_i is the price of the car. Furthermore, suppose that n_1 of the cars have automatic transmissions, while n_0 do not. Assume for simplicity that the data are sorted so that the first n_0 cars do not have automatic transmissions while the rest do, so that $x_1, \ldots, x_{n_0} = 0$ and $x_{n_0+1}, \ldots, x_n = 1$.

a) Show that $\bar{x} = \frac{n_1}{n}$.

Solution: As we know, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. We can see that the first n_0 cars do not have automatic

transmissions, which means that $x_1, \ldots, x_{n_0} = 0$. Thus, we can rewrite the equation for $\bar{x}$ as:

$$\begin{aligned}
\bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\
&= \frac{1}{n}\left(\sum_{i=1}^{n_0} x_i + \sum_{i=n_0+1}^{n} x_i\right) \\
&= \frac{1}{n}\left(\sum_{i=1}^{n_0} 0 + \sum_{i=n_0+1}^{n} 1\right) \\
&= \frac{1}{n}\left(0 + (n - n_0)\right) \\
&= \frac{n - n_0}{n} \\
&= \frac{n_1}{n}
\end{aligned}$$

b) Show that $\sum_{i=1}^{n}(x_i - \bar{x})y_i = \frac{n_0}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{n}\sum_{i=1}^{n_0} y_i$

Solution:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})y_i &= \sum_{i=1}^{n}(x_i - \frac{n_1}{n})y_i \\
&= \sum_{i=1}^{n_0}(x_i - \frac{n_1}{n})y_i + \sum_{i=n_0+1}^{n}(x_i - \frac{n_1}{n})y_i \\
&= \sum_{i=1}^{n_0}(0 - \frac{n_1}{n})y_i + \sum_{i=n_0+1}^{n}(1 - \frac{n_1}{n})y_i \\
&= (1 - \frac{n_1}{n})\sum_{i=n_0+1}^{n} y_i - (\frac{n_1}{n})\sum_{i=1}^{n_0} y_i \\
&= \frac{n_0}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{n}\sum_{i=1}^{n_0} y_i, \qquad \text{as } \frac{n_1}{n} = \frac{n - n_0}{n} = 1 - \frac{n_0}{n}
\end{aligned}$$

Hence the result is shown.

c) Suppose least squares regression is used to fit a linear prediction rule $H(x) = w_1 x + w_0$ to this data. Show that the prediction $H(0)$ is the mean price of cars without automatic transmissions ($\frac{1}{n_0}\sum_{i=1}^{n_0} y_i$) and the prediction $H(1)$ is the mean price of cars with automatic transmissions ($\frac{1}{n_1}\sum_{i=n_0+1}^{n} y_i$).

Hint: use the result from the previous part, combined with the result from Problem 2, part (b).

Solution: Using the other least squares regression result, we and the answer from part (b), we

can see that:

$$\begin{aligned}
H(0) &= w_1(0) + w_0 \\
&= w_0 \\
&= \bar{y} - w_1\bar{x} \\
&= \bar{y} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x} \\
&= \bar{y} - \frac{\frac{n_0}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{n}\sum_{i=1}^{n_0} y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x}
\end{aligned}$$

Now lets simplify the denominator:

$$\begin{aligned}
\sum_{i=1}^{n}(x_i - \bar{x})^2 &= \sum_{i=1}^{n_0}(0 - \frac{n_1}{n})^2 + \sum_{i=n_0+1}^{n}(1 - \frac{n_1}{n})^2 \\
&= \sum_{i=1}^{n_0}(\frac{n_1}{n})^2 + \sum_{i=n_0+1}^{n}(\frac{n_0}{n})^2 \\
&= \frac{n_1^2}{n^2}n_0 + \frac{n_0^2}{n^2}(n - n_0) \\
&= \frac{n_1^2}{n^2}n_0 + \frac{n_0^2}{n^2}(n_1) \\
&= \frac{n_0 n_1(n_0 + n_1)}{n^2}, \quad (n_0 + n_1 = n) \\
&= \frac{n_0 n_1}{n}
\end{aligned}$$

Now we can plug this back into the original equation with $\bar{x} = \frac{n_1}{n}$:

$$\begin{aligned}
w_1\bar{x} &= \frac{\frac{n_0}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{n}\sum_{i=1}^{n_0} y_i}{\frac{n_0 n_1}{n}}\frac{n_1}{n} \\
&= (\frac{n_0}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{n}\sum_{i=1}^{n_0} y_i)\frac{1}{n_0} \\
&= \frac{1}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{nn_0}\sum_{i=1}^{n_0} y_i
\end{aligned}$$

Now we can plug this back into the original equation:

$$\begin{aligned}
w_0 &= \bar{y} - w_1\bar{x} \\
&= \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{1}{n}\sum_{i=n_0+1}^{n} y_i + \frac{n_1}{nn_0}\sum_{i=1}^{n_0} y_i \\
&= \frac{1}{n}\sum_{i=n_0+1}^{n} y_i - \frac{1}{n}\sum_{i=1}^{n_0} y_i + \frac{1}{n}\sum_{i=i}^{n_0} y_i + \frac{n_1}{nn_0} y_i \\
&= \sum_{i=1}^{n_0} y_i(\frac{1}{n} + \frac{n_1}{nn_0}) \\
&= \sum_{i=1}^{n_0} y_i(\frac{n_0 + n_1}{nn_0}) \\
&= \sum_{i=1}^{n_0} y_i(\frac{n}{nn_0}) \\
&= \sum_{i=1}^{n_0} y_i(\frac{1}{n_0}) \\
&= \frac{1}{n_0}\sum_{i=1}^{n_0} y_i
\end{aligned}$$

Which is the mean price of cars without automatic transmissions and is equal to the $H(0)$ given. Now using this we can find the $H(1)$:

$$\begin{aligned}
H(1) &= w_1(1) + w_0 \\
&= (\frac{n_0}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{n}\sum_{i=1}^{n_0} y_i)(\frac{n}{n_1 n_0}) + \frac{1}{n_0}\sum_{i=1}^{n_0} y_i \\
&= \frac{1}{n_1}\sum_{i=n_0+1}^{n} y_i - \frac{1}{n_0}\sum_{i=1}^{n_0} y_i + \frac{1}{n_0}\sum_{i=1}^{n_0} y_i \\
&= \frac{1}{n_1}\sum_{i=n_0+1}^{n} y_i
\end{aligned}$$

Which is the mean price of cars with automatic transmissions and is equal to the $H(1)$ given.