DSC 140A - Homework 01

Due: Wednesday, January 18

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 p.m.

Problem 1.

In practice, the performance of nearest neighbor predictors is often seen to decrease 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 an 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 provide your code.

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.

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.

Solution:

Problem 2.

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

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

$$w_0 = \bar{y} - w_1 \bar{x}$$

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: We begin by breaking the sum apart:

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

Since $\bar{x}$ does not depend upon i, we can pull it out in front of its summation:

$$\begin{aligned} &= \sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} 1 \\ &= \sum_{i=1}^{n} x_i - n\bar{x} \end{aligned}$$

Using the definition of $\bar{x} = \frac{1}{n}\sum x_i$:

$$\begin{aligned} &= \sum_{i=1}^{n} x_i - n \cdot \frac{1}{n} \cdot \sum_{i=1}^{n} x_i \\ &= \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i \\ &= 0 \end{aligned}$$

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

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

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

Solution: The only difference between this new formula and the familiar one is that $\sum(x_i - \bar{x})(y_i - \bar{y})$ is replaced by $\sum(x_i - \bar{x})y_i$, so we'll show that these two are equal.

We'll start with $\sum(x_i - \bar{x})(y_i - \bar{y})$. We want to use the result of the previous part, which requires us to get $\sum(x_i - \bar{x})$ by itself. To do so, we'll try expanding the product in the summand:

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

We recognize the first term in the summand, $(x_i - \bar{x})y_i$, as the one we want to be left with; can we get rid of the second term somehow? We'll split the summand:

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

Now, $\bar{y}$ is a constant as far as the summation is concerned, so we can move it in front:

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

And now we've isolated $\sum (x_i - \bar{x})$ as we wanted. We can get rid of the entire second term, since it is zero:

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

Since $\sum (x_i - \bar{x})(y_i - \bar{y}) = \sum (x_i - \bar{x}) y_i$, we have our result:

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

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: We know that $\bar{x}$ is the average of the x_i's

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

We know that the first n_0 of the x_i's are zero, and last n_1 are one. So let's break the sum into two sums: one over the first n_0 terms, and the second over the remaining:

$$= \frac{1}{n} \left(\sum_{i=1}^{n_0} x_i + \sum_{i=n_0+1}^{n} x_i \right)$$

Each term in the first sum is zero, and so the sum is zero. Each term in the second sum is one:

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

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

Solution: We start by using the fact that $\bar{x} = n_1/n$. So:

$$\sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n}(x_i - n_1/n)y_i$$

I also know that the first n_0 of the x_i are zero, while the rest are one. So we'll once again split the summation into two summations:

$$
\begin{aligned}
&= \sum_{i=1}^{n_0}(x_i - n_1/n)y_i + \sum_{i=n_0+1}^{n}(x_i - n_1/n)y_i \\
&= \sum_{i=1}^{n_0}(0 - n_1/n)y_i + \sum_{i=n_0+1}^{n}(1 - n_1/n)y_i \\
&= \sum_{i=1}^{n_0}(-n_1/n)y_i + \sum_{i=n_0+1}^{n}(1 - n_1/n)y_i
\end{aligned}
$$

Switching the order of the result to match the target expression:

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

This looks very similar to the target expression, but is $(1 - n_1/n) = n_0/n$? If we rewrite 1 and n/n, we get $(1 - n_1/n) = (n - n_1)/n = n_0/n$, so:

$$= \frac{n_0}{n}\sum_{i=n_0+1}^{n} y_i - \frac{n_1}{n}\sum_{i=1}^{n_0} y_i$$

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: In order to make predictions, we need to first find the slope w_1 and the intercept w_0. Recognize that the expression we found in the last step is the numerator of

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

(This is an alternative formula for w_1 that you derived in another problem).

We'll now compute the denominator and simplify to find an expression for w_1. We have:

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

Substituting $\bar{x} = n_1/n$:

$$\begin{aligned}&= \sum_{i=1}^{n_0} (0 - n_1/n)^2 + \sum_{i=n_0+1}^{n_1} (1 - n_1/n)^2 \\ &= \sum_{i=1}^{n_0} (n_1/n)^2 + \sum_{i=n_0+1}^{n_1} (1 - n_1/n)^2\end{aligned}$$

We can simplify $1 - n_1/n$ by noting that it is $(n - n_1)/n = n_0/n$:

$$\begin{aligned}&= \sum_{i=1}^{n_0} (n_1/n)^2 + \sum_{i=n_0+1}^{n_1} (n_0/n)^2 \\ &= n_0 (n_1/n)^2 + n_1 (n_0/n)^2 \\ &= \frac{n_0 n_1^2}{n^2} + \frac{n_0^2 n_1}{n^2} \\ &= \frac{n_0 n_1 (n_1 + n_0)}{n^2} \\ &= \frac{n_0 n_1 n}{n^2} \\ &= \frac{n_0 n_1}{n}\end{aligned}$$

That gives us:

$$
\begin{aligned}
w_1 &= \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \frac{\frac{n_0}{n} \sum_{i=n_0+1}^n y_i - \frac{n_1}{n} \sum_{i=1}^{n_0} y_i}{\frac{n_0 n_1}{n}} \\
&= \frac{\dfrac{n_0}{n} \displaystyle\sum_{i=n_0+1}^n y_i - \dfrac{n_1}{n} \displaystyle\sum_{i=1}^{n_0} y_i}{\dfrac{n_0 n_1}{n}} \\
&= \left(\frac{n_0}{n} \sum_{i=n_0+1}^n y_i - \frac{n_1}{n} \sum_{i=1}^{n_0} y_i \right) \cdot \frac{n}{n_0 n_1} \\
&= \frac{1}{n_1} \sum_{i=n_0+1}^n y_i - \frac{1}{n_0} \sum_{i=1}^{n_0} y_i
\end{aligned}
$$

So the slope is just the mean price of cars with automatic transmissions, minus the mean price of cars without automatic transmissions.

Recall that $w_0 = \bar{y} - w_1 \bar{x}$. So

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

Since $H(0) = w_0$, this is our predicted price for cars without automatic transmissions (it is the average price of cars without automatic transmissions).

Now we'll compute $H(1)$:

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

This is just the average price of cars with automatic transmissions.