1.5 – K-Nearest Neighbors

1 Comparing Two Variables

In earlier readings, we explored how to describe a single variable, its center, spread, and shape using tools like histograms, mean, median, and standard deviation. These helped us summarize what we know about one quantity at a time. But what if we want to understand how two variables relate to one another?

This is where a *scatter plot* is beneficial. In a scatter plot, each record is represented by a point [1]. Scatter plots can reveal trends or patterns, such as whether one variable tends to increase or decrease as the other increases. It can indicate the form of the relationship whether it's linear, curved, or random. Scatter plots can also help identify outliers, by showing data points that don't follow the overall trend. Scatter plots can also demonstrate whether there is some predictive potential for the data depending on how strong the pattern or trend is.

Definition 1.5.1 (Scatter Plot)
A two-dimensional graph that uses dots to represent the various simultaneous values for two different numerical variables [1].

Let's consider the data in Table 1 we've been using for the last several readings.

Table 1: Sample Cadet Fitness Data

ID	Run Time	Deadlift	Age
A001	14.8	300	20
A002	13.5	275	19
A003	15.2	320	21
A004	13.9	290	20
A005	14.1	310	19
A010	14.0	295	20
A011	10.7	280	
A012	15.5	400	22

Figure 1 shows the scatter plot of cadet run times on the horizontal axis and deadlift weights on the vertical axis. While there isn't a lot of data, we can see that there are two points that seem dissimilar to the rest of the data. Also, it appears that generally, the more weight lifted, the slower the run time. This is a small sample of the total data. But what happens when we include more of the original data?

From Figure 2, where we've plotted the full data set, we can see that the original trend we noticed no longer holds. We can see there isn't any defined trend when we include all of the data. However, we can see that one of the outliers from before still seems to be slightly outside the rest of the data.

Is Figure 2 the best representation of the data? It is accurate, but the scales on each axis are different. Deadlift

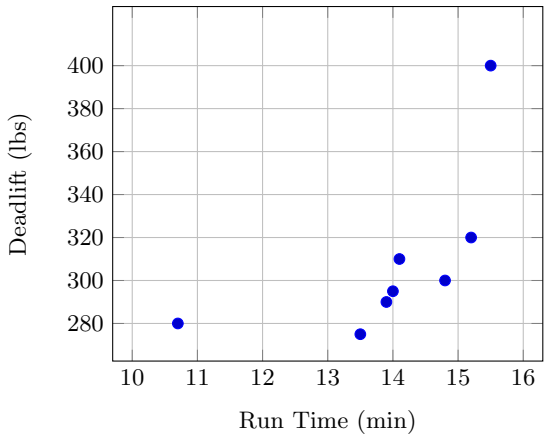

Figure 1: Scatter plot of Cadet Run Time in minutes on the horizontal axis and deadlift in pounds on the vertical axis, limited data.

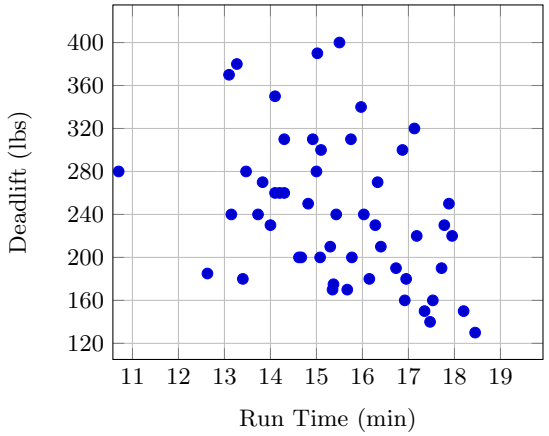

Figure 2: Scatter plot of cadet run time in minutes on the horizontal axis and deadlift in pounds on the vertical axis, full data.

is measured in hundreds of pounds, while most run time are between 10 and 20 minutes. This scaling difference could be detrimental in determining a trend or pattern for the data. If we remember the previous reading, we learned about calculating the z-score for a data point, or calculating the standardized value. Standardizing puts data on the same scale. What happens when we plot the standardized version of the data?

Plotting the standardized data as shown in Figure 3 doesn't appear to change the trend significantly, but it does seem to condense the vertical axis slightly.

2 Making a Prediction

What if we add a variable to our dataset? Could we make a prediction to determine the value of that variable? In this example, could we predict a new cadet's sprint-drag-carry (SDC) score if we only knew their deadlift and two

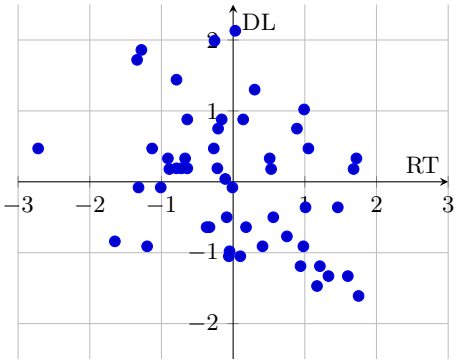

Figure 3: Scatter plot of standardized run time (RT) against standardized deadlift (DL).

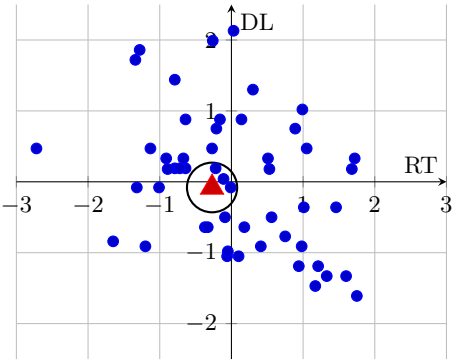

Figure 4: Scatter plot of standardized run time (RT) against standardized deadlift (DL) with the predictor record plotted as a red triangle and a circle identifying the closest records.

mile run time?

We can! In this section we start the transition into predictive modeling. Predictive modeling uses data to make a prediction of what is likely to happen [1]. We will discuss more about how the modeling process applies to predictive modeling in the next reading, but as with all types of modeling, defining variables is a key component. In predictive modeling we are concerned with two different variables, explanatory variables and response variables.

Definition 1.5.2 (**Explanatory Variable**)
An *explanatory* variable is the variable used to predict or explain the values of another variable. Also called an independent variable. [3]

Definition 1.5.3 (**Response Variable**)
A *response* variable is the variable whose value is predicted by the explanatory variable. Also called a dependent variable. [3]

2.1 K-Nearest Neighbors Algorithm

In our example, the explanatory variables are the deadlift weight and the two-mile run time. The response variable is the sprint, drag, carry (SDC) time in minutes and seconds. With two or more numerical explanatory variables, one of the methods we can use to make a prediction is called **K-Nearest Neighbors (KNN)**. KNN is a popular algorithm that makes a prediction based on a specific number (k) of closest neighbors to a predictor record [1]. A predictor record is the record, or data entry, for which we are making a prediction [1].

Let's say there is another cadet, and we know that they run a 15 minute two-mile and lifts 240 lbs. What would we predict their SDC score to be? In Figure 4 we've plotted the new record as a red triangle against the previous data. The idea behind KNN is that we find the k-nearest data points to our new record, then the average of those response variable values is the prediction.

If we let $k = 3$ then the circle around the new record in Figure 4 shows us visually the three records we should expect to be the closest to the new one. However, we can do better than visually inspecting the scatter plot. We can find the distance from this new point to every point in the data set and then find the closest three. We calculate the distance between two points using the Euclidean distance formula.

Definition 1.5.4 (**Euclidean Distance** [1])
The distance between point (x_1, y_1) and (x_2, y_2) is

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Once we have found the three closest records to the new one, we can average the values of the response variable (SDC score) to come up with a predicted score. We can also use the KNN algorithm to make predictions on categorical response variables [1]. For example, what if we were trying to predict whether this new cadet passes or fails a timed ruck march? As long as we have the historical data associated with passing and failing for the data set, we can apply KNN to make a prediction. What is different is that we can't take the average of categorical data. So instead of finding the average, we find the most often repeated category and use that as our prediction. For example, if $k = 3$ and the closest points have two passes and one fail, the prediction is that the cadet will pass.

Problem 1 below provides a worked example of the KNN algorithm for categorical variables.

If we think about KNN and how it applies in our modeling triangle, the output of our transform step is the distance formula, the scatter plot, and a proposed k value. The solve step encompasses all steps in running the algorithm: calculating the distances, determining the k-closest points, and making the prediction. Finally, our

interpret step is making sense of the prediction and analyzing the sensitivity of k.

Steps for KNN Algorithm

1. Calculate mean and standard deviation of explanatory variables, without the new record included.
2. Standardize the explanatory variables, if necessary, including those for the new record.
3. Calculate distance from new record to every other record in the data set.
4. Select k value
5. Identify k-nearest neighbors
6. Make a prediction:
   - Most common category if response variable is categorical
   - Average of response variable if numerical

2.2 Choosing k

The value of k for KNN is essential to getting a reasonable prediction. k is the number of records to include to make the prediction [1]. It can take on any positive integer from 1 to n, the total number of data points in the sample. The value selected for k can change the result, so it is important to measure the sensitivity of k to try and select the best one.

If we were to let $k = 1$ then our prediction is based solely on the single closest neighbor [1]. If the nearest neighbor is not representative of other points close to it, we could make a bad prediction. Similarly, if we let $k = n$ or the total number of data points in our sample, our prediction is not different than the overall trend of the data set. For our example, it would be equivalent to finding the average final score for the fitness tests. We don't need to run the algorithm to find that value. Unfortunately, there is no single method for selecting an appropriate k-value. However, there are several rules of thumb. One rule of thumb is to select $k = \sqrt{n}$ and round to the closest odd integer [2]. Another rule of thumb is to select an odd value for k [2]. Selecting an odd value can be beneficial especially when trying to make a prediction for a categorical variable as it decreases the possibility of having a tie.

We can also use data partitioning to help determine the best value for k. If we split our data into a training set and a test set, we use the data points in our test set to identify a value of k that reduces the prediction error [1]. This works because we know the results for the response variable in our test set. We can try different values for k and determine which k-value produces the most accurate predictions.

A simplified test for the sensitivity of k is to identify how selecting a different value for k changes your resulting prediction [1]. For example, if we originally set $k = 3$ and we predict that a cadet will pass their fitness test, then we change k and each time we change k we get a different prediction, we can say that our k-value is sensitive. Our goal with selecting k is to reduce its sensitivity.

Problem 1.5.1: K-Nearest Neighbors

Given the standardized data for five cadets below, predict if a cadet who runs a 15 minute two-mile and deadlifts 240 lbs passes a timed ruck using $k = 3$. The mean and standard deviation for each variable is provided.

ID	Std. RT	Std. DL	Ruck Pass/Fail
A013	-0.51	-0.83	1
A022	-0.43	-0.36	1
A035	-0.22	0.19	1
A042	-0.35	-0.36	1
A043	-0.01	-0.08	0

$$\bar{x}_{DL} = 245.96, \qquad s_{DL} = 72.23$$
$$\bar{x}_{RT} = 15.46, \qquad s_{RT} = 1.71$$

Solution:

Standardize new data: $z_i = \frac{x_i - \bar{x}}{s}$

$$\text{Run Time: } z = \frac{15 - 15.46}{1.71} = -0.27$$
$$\text{Deadlift: } z = \frac{240 - 245.96}{72.23} = -0.08$$

Calculate distance from new record to every other record.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$d_{A013} = \sqrt{(-0.51 - (-0.27))^2 + (-0.83 - (-0.08))^2}$$
$$d_{A013} = 0.24$$

$$d_{A022} = \sqrt{(-0.43 - (-0.27))^2 + (-0.36 - (-0.08))^2}$$
$$d_{A022} = 0.32$$

$$d_{A035} = \sqrt{(-0.22 - (-0.27))^2 + (0.19 - (-0.08))^2}$$
$$d_{A035} = 0.28$$

$$d_{A042} = \sqrt{(-0.35 - (-0.27))^2 + (-0.36 - (-0.08))^2}$$
$$d_{A042} = 0.29$$

$$d_{A043} = \sqrt{(-0.01 - (-0.27))^2 + (-0.08 - (-0.08))^2}$$

$$d_{A043} = 0.25$$

The three closest records are `A013` (pass), `A043` (fail), and `A035` (pass). Two out of the three closest records passed the timed ruck.

We predict the new cadet will pass the timed ruck event.

References

[1] Frederick Hillier et al. *MA103 Mathematical Modeling: Introduction to Management, Science, & Business Analytics with Connect*. McGraw-Hill, 2024.

[2] Geetha Mattaparthi. *K-Nearest Neighbors(KNN): A comprehensive guide*. Medium, Dec. 2023. URL: https://medium.com/%40geethasreemattaparthi/k-nearest-neighbors-knn-a-comprehensive-guide-a83857ab2666 (visited on 05/19/2025).

[3] W. Paul Vogt. *Dictionary of Statistics & Methodology: A Nontechnical Guide for the Social Sciences*. Sage Publications, 2005.