# ECE/CS/ME 539 – Fall 2024 — Activity 9

## 1.

The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

| $X_1$ | $X_2$ | $X_3$ | $Y$ |
|---|---|---|---|
| 0 | 3 | 0 | Red |
| 2 | 0 | 0 | Red |
| 0 | 1 | 3 | Red |
| 0 | 1 | 2 | Green |
| -1 | 0 | 1 | Green |
| 1 | 1 | 1 | Red |

Suppose we wish to use this data set to make a prediction for $Y$ when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

The Euclidean distance formula for two points $p$ and $q$ is: $d(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}$. Therefore, the distance between each observation and the test point $X_1 = X_2 = X_3 = 0$ is:

- For point $X_1 = 0, X_2 = 3, X_3 = 0$: $\sqrt{(0-0)^2 + (3-0)^2 + (0-0)^2} = \sqrt{3^2} = 3$
- For point $X_1 = 2, X_2 = 0, X_3 = 0$: $\sqrt{(2-0)^2 + (0-0)^2 + (0-0)^2} = \sqrt{2^2} = 2$
- For point $X_1 = 0, X_2 = 1, X_3 = 3$: $\sqrt{(0-0)^2 + (1-0)^2 + (3-0)^2} = \sqrt{1^2 + 3^2} = \sqrt{10}$
- For point $X_1 = 0, X_2 = 1, X_3 = 2$: $\sqrt{(0-0)^2 + (1-0)^2 + (2-0)^2} = \sqrt{1^2 + 2^2} = \sqrt{5}$
- For point $X_1 = -1, X_2 = 0, X_3 = 1$: $\sqrt{(-1-0)^2 + (0-0)^2 + (1-0)^2} = \sqrt{(-1)^2 + 1^2} = \sqrt{2}$
- For point $X_1 = 1, X_2 = 1, X_3 = 1$: $\sqrt{(1-0)^2 + (1-0)^2 + (1-0)^2} = \sqrt{1^2 + 1^2 + 1^2} = \sqrt{3}$

(b) What is our prediction with $K = 1$? Why?

When $K = 1$, our prediction is Green, since it is the qualitative response variable $Y$ of the point $X_1 = -1, X_2 = 0, X_3 = 1$, which has the smallest Euclidean distance (i.e, is closer) from the test point $X_1 = X_2 = X_3 = 0$.

(c) What is our prediction with $K = 3$? Why?

When $K = 3$, the 3 points that have the smallest Euclidean distance and thus are closer to the test point $X_1 = X_2 = X_3 = 0$ are:

$X_1 = -1, X_2 = 0, X_3 = 1$ (with distance $\sqrt{2}$), $X_1 = 1, X_2 = 1, X_3 = 1$ (with distance $\sqrt{3}$) and $X_1 = 2, X_2 = 0, X_3 = 0$ (with distance $\sqrt{2^2} = 2$). The respective qualitative response variable $Y$ of these three points are Green, Red and Red. Therefore, since the majority of them are Red, the prediction with $K = 3$ is Red.

## 2.

When the number of features $p$ is large, there tends to be a deterioration in the performance of KNN and other local approaches that perform prediction using only observations that are near the test observation for which a prediction must be made. This phenomenon is known as the curse of dimensionality, and it ties into the fact that non-parametric approaches often perform poorly when $p$ is large.

(a) Suppose that we have a set of observations, each with measurements on $p = 1$ feature, $X$. We assume that $X$ is uniformly (evenly) distributed on [0, 1]. Associated with each observation is a response value. Suppose that we wish to predict a test observation's response using only observations that are within 10% of the range of $X$ closest to that test observation. For instance, in order to predict the response for a test observation with $X = 0.6$, we will use observations in the range [0.55, 0.65]. On average, what fraction of the available observations will we use to make the prediction?

Since $X$ is uniformly distributed on [0,1] and we are only using observations that are within 10% of the range of $X$ closest to the test observation, we expect that the fraction of the available observations that will be used to make the prediction will be $\frac{1}{10}$ or 0.1.

(b) Now suppose that we have a set of observations, each with measurements on $p = 2$ features, $X1$ and $X2$. We assume that predict a test observation's response using only observations that $(X1, X2)$ are uniformly distributed on [0, 1] × [0, 1]. We wish to are within 10% of the range of $X1$ and within 10% of the range of $X2$ closest to that test observation. For instance, in order to predict the response for a test observation with $X1 = 0.6$ and $X2 = 0.35$, we will use observations in the range [0.55, 0.65] for $X1$ and in the range [0.3, 0.4] for $X2$. On average, what fraction of the available observations will we use to make the prediction?

If we plot a unit square in the 2D space, we can view the fraction of the available observations that will be used to make the prediction as the area of the box that is 10% of the width and 10% of the height of the unit square where our data are uniformly distributed. For each dimension ($X1$ and $X2$), we are using 10% (0.10) of the range, so the area of the box is $0.10 \cdot 0.10 = 0.01$ or 1% of the unit square. Thus, we expect that the fraction of the available observations that will be used to make the prediction will be $\frac{1}{100}$ or 0.01.

(c) Now suppose that we have a set of observations on $p = 100$ features. Again the observations are uniformly distributed on each feature, and again each feature ranges in value from 0 to 1. We wish to predict a test observation's response using observations within the 10% of each feature's range that is closest to that test observation. What fraction of the available observations will we use to make the prediction?

Following the reasoning described above, we can infer that the fraction of the available observations that will be used to make the prediction is $\frac{1}{10^p}$. Thus, in the case where $p = 100$, we will use $\frac{1}{10^{100}} = 10^{-100}$ of the available observations.

(d) Using your answers to parts (a)–(c), argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

Since the fraction of the available observations that will be used to make the prediction is $\frac{1}{10^p}$, when p is large $10^p$ tends to infinity and thus $\frac{1}{10^p}$ tends to zero.

(e) Now suppose that we wish to make a prediction for a test observation by creating a $p$-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $p$ =1, 2, and 100, what is the length of each side of the hypercube? Comment what happens to the length of the sides as $\lim_{p\to\infty}$.

Since the $p$-dimensional hypercube that is centered around the test observation contains, on average, 10% of the training observations, that means that the volume of that hypercube should represent that percentage of the total volume. Since each side of the hypercube is of the same length, we need to find the length $l$ of the side of the cube such that the total volume is 10%.

Thus, we want $l^p = 0.10$. Solving for $l$ gives: $l = (0.10)^{1/p}$

For $p = 1 : l = (0.10)^{(1/1)} = 0.10$
For$p = 2 : l = (0.10)^{(1/2)} \approx 0.316$
For $p = 100 : l = (0.10)^{(1/100)} \approx 0.9772$

As $p$ increases, the side length $l$ approaches 1. Thus, in high-dimensional data, to encompass just 10% of observations, nearly the full range of each feature is needed. This diminishes the concept of "closeness," problematic for KNN, which relies on a meaningful proximity.