

**SDSC5001 Statistical Machine Learning I**  
**Assignment #2**

**Deadline: Nov. 12<sup>th</sup>, 2025 Mid-night**

1. 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. We will now investigate this curse.

(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?

(b) Now suppose that we have a set of observations, each with measurements on  $p = 2$  features,  $X_1$  and  $X_2$ . We assume that  $(X_1, X_2)$  are uniformly distributed on  $[0, 1] \times [0, 1]$ . We wish to predict a test observation's response using only observations that are within 10% of the range of  $X_1$  and within 10% of the range of  $X_2$  closest to that test observation. For instance, in order to predict the response for a test observation with  $X_1 = 0.6$  and  $X_2 = 0.35$ , we will use observations in the range  $[0.55, 0.65]$  for  $X_1$  and in the range  $[0.3, 0.4]$  for  $X_2$ . On average, what fraction of the available observations will we use to make the prediction?

(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?

(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.

2. Suppose we collect data for a group of students in a statistics class with variables  $X_1$  = hours studied,  $X_2$  = undergrad GPA, and  $Y$  = receive an A. We fit a logistic regression and produce estimated coefficient,  $\hat{\beta}_0 = -6$ ,  $\hat{\beta}_1 = 0.05$ ,  $\hat{\beta}_2 = 1$ .

(a) Estimate the probability that a student who studies for 40 h and has an undergrad GPA of 3.5 gets an A in the class.

(b) How many hours would the student in part (a) need to study to have a 50% chance of getting an A in the class?