Assignment18

John Bute

2024-10-15

R Markdown

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.

1. 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?

If X , then we use the observations found in the range X , which correspond to 10% of the available observations, as X is uniformly distributed on [0,1]. However, if X > 0.95, the range is [0.9, 1], which is not centered around X. Likewise, if X <0.05, then X is centered around [0, 0.1]. In both exceptions, we still use 10% of the available observations. Thus, we would expect 10% of available observations we will use to make the prediction

2. Now suppose that we have a set of observations, each with measurements on p = 2 features, X1 and X2. We assume that (X1,X2) are uniformly distributed on [0,1] ×[0,1]. We wish to predict a test observation's response using only observations that 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 (X_1, X_2) fall within the range of [0.05, 0.95], then they would use a range of $X_1 * X_2$, which relates to 10% of the range of X1 and 10% of the range of X2, but 10% * 10% of the range X1 x X2. If X_i <0.05 the correspodnding range is [0, 0.1]. Likewise, if X_i > 0.95, the corresponding range will be [0.9, 1]. In all cases, we will have an area of observations that correspond to about 1% of the total square [0,1] x [0,1], granded that all features are uniformly distributed

3. 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?

As we continue, we decrease the % of total observations used for our prediction. Say for [0,1]^100, each feature will correspond to 10% of a range [0, 1], provided that each feature is uniformly distributed within this range. Then since we focus on the area of our observations, which in this case corresponds to an area with 100 sides, then we use about 0.1^100 of the observatios, which is very small.

4. Using your answers to parts 1. to 3., argue that a drawback of KNN when p is large is that there are very few training observations "near" any given test observation.

For results from 1 to 3, we see that as we increase the number of features p, the fraction of observations within a local region decreases. For example:

p = 1, we use 10% of observations p = 2, we use 1% of observations p = 100, we use almost none of the obseravtions.

This demonstrates the curse of dimensionality, as when the number of p dimensions increases, the volume of the space for observations grows exponentially, thus the data becomes harder to find. Thus, for KNN, which relies on neighbourhoods, performs very poorly as there are few training observations that are local to a test point. For a high dimensional space, all points are far away from each other, so KNN's ability to find meaningful neighbours deteriorates, leading to poor performance.

5. 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 on your answer. Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When p = 1, a hypercube is simply a line segment, when p = 2 it is a square, and when p = 100 it is a 100−dimensional cube.

The hypercube's volume contains on average, 10% of all observations. To calculate the volume of a cube, we must:

$$s^p = 0.1$$

where s is the side length, and p is the number of features.

To find the side length, we merely solve for s, therefore

$$s = 0.1^{\frac{1}{p}}$$

For p = 1:

$$s = 0.1$$

For p = 2:

$$s = 0.1^{\frac{1}{2}} = 0.316$$

For p = 100:

$$s = 0.1^{\frac{1}{100}} = 0.977$$