Homework 3 - Introduction to Machine Learning for Engineers

kipngeno koech - bkoech

March 12, 2025

1 Dimensionality of K-Nearest Neighbors

When the number of features d is large, the performance of k-nearest neighbors, which makes predictions using only observations that are near the test observation, tends to degrade. This phenomenon is known as the *curse of dimensionality*, and it ties into the fact that non-parametric approaches often perform poorly when d is large.

1. Suppose that we have a set of training observations, each corresponding to a one-dimensional ($d = 1$) feature, X. We assume that X is uniformly (evenly) distributed on $[0, 1]$. Associated with each training observation is a response value.Suppose that we wish to predict a test observation x's response using only training observations that are within 10% of the range of x closest to that test observation. In other words, if $x \in [0.05, 0.95]$ then we will use training observations in the range $[x - 0.05, x + 0.05]$, as shown in Figure 1 when $x = 0.6$. When $x \in [0, 0.05)$ we use the range $[0, 0.1]$, and when $x \in (0.95, 1]$ we use training observations in the range $[0.9, 1]$. Figure 1 shows this range for $x = 0.02$.On average (assuming x is uniformly distributed on $[0, 1]$), what fraction of the available observations will we use to make the prediction?

2. Now suppose that we have a set of observations, each corresponding to two features, X_1 and X_2 (i.e., $d = 2$). We assume that (X_1, X_2) are uniformly distributed on $[0, 1] \times [0, 1]$. We wish to predict the response of a test observation (x_1, x_2) using only training 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 $x_2 = 0.04$, we will use training observations (X_1, X_2) such that $X_1 \in [0.55, 0.65]$ and $X_2 \in [0, 0.1]$.

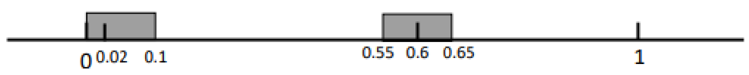

Figure 1: Range of observation, $d = 1$

Figure 1: Primal Formulation of SVM

3. On average, assuming x_1 and x_2 are each uniformly distributed on $[0, 1]$, what fraction of the available observations will we use to make the prediction?

4. Now suppose that we have a set of training observations on $d = 100$ 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?

5. Using your answers to parts a–c, argue that a drawback of k-nearest neighbors when d is large is that there are very few training observations "near" any given test observation.

6. Now suppose that we wish to make a prediction for a test observation by creating a d-dimensional hypercube centered around the test observation that contains, on average, 10% of the training observations. For $d = 1$, 2, and 100, what is the length of each side of the hypercube? How does your answer change as d increases, and what does this imply for the accuracy of k-nearest neighbors when d is large?

   Note: A hypercube is a generalization of a cube to an arbitrary number of dimensions. When $d = 1$, a hypercube is simply a line segment; when $d = 2$, it is a square; and when $d = 100$, it is a 100-dimensional cube.

2 Decision Trees

You obtained the following data from interviewing 15 people on the street. Based on a person's relationship status, age, education level and income, you can now build a decision tree to predict a person's phone usage.

Relationship Status	Age	Education	Income	Phone Usage
Single	>25	University	≤50K	Low
In a relationship	≤25	College	≤50K	Medium
Single	>25	University	>50K	Low
Married	≤25	University	≤50K	High
Single	>25	University	>50K	Low
Married	>25	College	≤50K	Medium
In a relationship	≤25	College	>50K	Medium
In a relationship	>25	High School	≤50K	Low
Married	>25	University	≤50K	High
Single	>25	High School	>50K	Low
In a relationship	≤25	College	>50K	Medium
In a relationship	>25	High School	≤50K	Low
Married	>25	University	≤50K	High
Single	≤25	High School	>50K	Low
In a relationship	≤25	College	>50K	Medium

Figure 2: Primal Formulation of SVM

1. What is the entropy of Phone Usage? (Calculate entropy in log2 base and round to 4 decimal places)
2. Find the information gain (IG) from each feature (relationship status, age, education levarianceel and income). Which feature should be chosen at the root of the tree? Show your calculations for the information gain (IG) and explain your choice in a sentence. (Calculate entropy in log2 base and round to 4 decimal places)
3. Use the root you found in part (b) to determine the rest of the nodes in the decision tree for the abovariancee data. Draw the full decision tree, i.e., keep splitting the nodes until further splits do not lead to any information gain (you may not need all the features for this). For each split, show your working on why you chose this feature based on information gain.