# Class 3 Review Notes

AI & Machine Learning
Fall 2023 - Laurie Ye

## 3 KNN Classification and Regression

### 3.1 Introduction

We can't always have an infinite number of observations (or very large datasets). In this case, we can specify a number k and use the k nearest neighbors. Hence the name k-NN.

### 3.2 k-NN Binary Classification

- For a particular set of x-variables (features), estimate $P(Y = 1 \mid \text{features})$ as the proportion of the k nearest neighbors (to the set of x-variables) that take the value of 1.
- Call this probability $\hat{p}(k, \mathbf{x})$, where $\mathbf{x}$ is the vector of features for which the classification is desired.
- Compare this probability to the threshold probability $p_{\text{threshold}}$.
- Classify as 1 if
$$\hat{p}(k, \mathbf{x}) \geq p_{\text{threshold}}$$
- Note: $p_{\text{threshold}}$ can be varied to produce ROC curves and the AUC.
- Often the criterion is majority rule:
$$p_{\text{threshold}} = 0.5$$

### 3.3 k-NN Classification in General & Regression

- For a particular set of x-variables (features), estimate the probability of each class as the proportion of the k nearest neighbors (to the set of x-variables) that are members of that class.
- Use a similar voting scheme, with the class with the most "votes" determining the predicted class (the classification).
- For a particular set of x-variables (features), estimate the $\hat{Y}$ by the average of the $Y$ values of the $k$ nearest neighbors (to the set of x-variables).

**Note 1. k-NN classification and regression is based on a very simple and intuitive idea. All of the "action" is in determining which observations are close to a particular set of x-variables. So notions of "distance" or "similarity" are important.**

### 3.4 Euclidean Distance (L2 Norm)

The Euclidean distance, also known as the L2 norm, is defined as the length of the difference between two vectors (enforce 0 as coefficient). For two vectors $\mathbf{x}$ and $\mathbf{y}$, the Euclidean distance is given by:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{bmatrix}$$

The distance $d(\mathbf{x}, \mathbf{y})$ is calculated as:

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) = \|\mathbf{x} - \mathbf{y}\| = \|\mathbf{y} - \mathbf{x}\|$$

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

$$= \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

### 3.5 Properties of Distance

The properties of distance are closely related to the properties of length.

- $d(\mathbf{x}, \mathbf{y}) \geq 0$
- $d(\mathbf{x}, \mathbf{x}) = 0$
- $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$
- $d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y})$
- $d(c\mathbf{x}, c\mathbf{y}) = |c| \cdot d(\mathbf{x}, \mathbf{y})$

### 3.6 L1 Norm (Manhattan Distance)

The L1 norm, or Manhattan Distance, between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_p - y_p|$$

$$= \sum_{i=1}^{p} |x_i - y_i|$$

### 3.7 L∞ Norm

The $L\infty$ norm (Chebyshev distance) between two vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as:

$$d(\mathbf{x}, \mathbf{y}) = \max \{|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_p - y_p|\}$$

$$= \max_i |x_i - y_i|$$

### 3.8 Standardize

- Most distance measures are very sensitive to the scale of the variables.
- One approach is to standardize each variable.
- Think $\mathbf{X}$ data matrix like in regression:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

- For each $x$-variable $j$, compute

$$z_{ij} = \frac{x_{ij} - \bar{x}_j}{s_j}$$

where

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$$

and

$$s_j = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2}$$

### 3.9 Collaborative Filtering

- Collaborative filtering is a term used in recommender systems for an idea that is very similar to k-NN classification.
- In this context, "collaborate" is used to refer to the set of similar users.
- To make a prediction for the focal user for a possible item to recommend, all similar users (collaborators) that have rated the item must be found and a prediction based on their ratings made. This is the "filtering" process.

### 3.10 Reference

Goizueta Business School-Emory University: Professor George S. Easton