Problem Set 01

Due: Thursday, 9/15 at 11:59pm

Introduction

For the K-Nearest Neighbors classifier, after choosing K we need to somehow find the nearest neighbors. That is, for a new point x_0, we need to determine the neighboring set $\mathcal{N}_0$.

Here, "nearest" gives off a notion of closeness or distance. Therefore, we want to somehow measure distance between features/predictors. How might we do that?

Euclidean distance

In Mathematics, the Euclidean distance is defined as the shortest possible path through space between two points.

On a number line (one-dimension), the distance between two points a and b is simply the absolute value of their difference: letting $d(a, b)$ denote the Euclidean distance between points a and b, $d(a, b) = |a - b|$. This is equivalent to $d(a, b) = \sqrt{(a-b)^2}$

In two-dimensions (think x - y coordinate system, latitude-longitude), the two points are $\mathbf{a} = (a_1, a_2)$ and $\mathbf{b} = (b_1, b_2)$. The euclidean distance between points $\mathbf{a}$ and $\mathbf{b}$ is given by

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

Importantly, "two-dimensions" refers to the number of coordinates in each point; not the fact that we are calculating a distance between two points.

This easily generalizes to p-dimensions! If our two points are p-dimensional (i.e. $\mathbf{a} = (a_1, a_2, \ldots, a_p)$ and $\mathbf{b} = (b_1, b_2, \ldots, b_p)$), then

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots + (a_p - b_p)^2} = \sqrt{\sum_{j=1}^{p} (a_j - b_j)^2}$$

For example: let $\mathbf{a} = (1, 2)$ and $\mathbf{b} = (3, 5)$. Then the Euclidean distance between them is $d(\mathbf{a}, \mathbf{b}) = \sqrt{(3-1)^2 + (5-2)^2} = \sqrt{13}$.

Manhattan distance

Another possible way to define distance is using the Manhattan distance. This is named after the grid-system of Manhattan's roads. An analogy is the number of blocks (in the north, south, east or west directions) a taxicab must travel on, in order to reach its destination on the grid of streets in parts of New York City.

The Manhattan distance d_m between two p-dimensional points $\mathbf{a}$ and $\mathbf{b}$ is defined as

$$d_m(\mathbf{a}, \mathbf{b}) = \sum_{j=1}^{p} |a_j - b_j|$$

Distances for categorical features

Thus far, we have assumed that all our predictors are quantitative, but oftentimes we have qualitative features. How does one measure the distance between categorical levels?

One approach is to assign numeric values for each possible category. For example, if the feature is "Experience level" with levels (Beginner, Average, Professional), maybe we assign the values (0, 5, 10). However, this assumes an implicit ordering in the levels.

Alternatively, consider that all our p features are binary in that they each take one of two values. Example: X_1 = smoker status (non-smoker/smoker) and X_2 = drinks alcohol (no/yes). The Hamming distance is the number (or proportion, depending on the context) of features for which the two values do not match.

Exercise 1

Suppose we have 6 training data points as follows:

X1	X2	X3	Y
0	3	1	red
1	-1	2	red
1	1	-1	blue
2	0	0	red
-1	1	0	blue
0	1	0	red

Suppose we want to use this data set to make a prediction for Y when (X1, X2, X3) = (0,0,0) using K-nearest neighbors.

a) Compute the Euclidean distance between each observation and the test point.

b) Using the Euclidean distance metric, what is our prediction with $K = 1$? Why?

c) Using the Euclidean distance metric, what is our prediction with $K = 3$? Why?

d) Compute the Manhattan distance between each observation and the test point.

e) Using the Manhattan distance metric, what is our prediction with $K = 1$? Why?

f) Using the Manhattan distance metric, what is our prediction with $K = 3$? Why?

g) Can you think of certain scenarios where we might prefer one distance metric over another?

Exercise 2

a) Suppose we have 6 training data points as follows

Smoke.status	Alcohol	Athlete
smoker	yes	no
non-smoker	no	yes
smoker	no	no
smoker	yes	yes
smoker	yes	no

Smoke.status	Alcohol	Athlete
non-smoker	no	yes

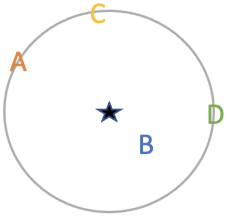
Calculate the Hamming distance between a new test point where the individual does not smoke, does not consume alcohol, and does not play a sport.

b) Hamming distance is commonly used to measure distance between words or phrases. What is the Hamming distance between the words "MIDDLEBURY" and "SWARTHMORE"?

c) For points with p features, are the minimum and maximum possible Hamming distances?

Exercise 3

Suppose we are still classification setting, but we have 4 different classes: $\{A, B, C, D\}$. We wish to classify the starred observation using the KNN classification method with $K = 2$. After calculating distances, we have 4 nearest neighbors:



Part 1 (a-c) Unfortunately, besides one obvious neighbor belonging to class B, the three other points are equidistant to our unclassified star observation. How do we solve this?

Option 1: *choose a different* K.

Option 2: *randomly choose between the tied neighbors.*

Option 3: *allow at least* K *neighbors until a natural stopping point.* In this example, the idea is to choose the smallest number such that K is greater than or equal to 2, and that no ties exist.

a) What are some issues with Option 1?

b) If we go with Option 2, what are the possible K-neighbor sets?

c) If we go with Option 3, what is our resulting neighbor set?

Part 2 (d-e)

d) Suppose we go with Option 3 to choose our neighbors. Now we need to classify the starred point. However, what issue arises if we apply the usual "majority vote" rule?

e) Suggest at least two methods for addressing the concern in (d). For each suggestion, describe what you would classify the starred point as. Of your suggestions, which would you prefer and why?

f) Although the issues in Part 1 and Part 2 are similar, clearly state the differences between the two.

Exercise 4

In class, we introduced the notion of mean squared error (MSE):

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where y_i is the true value, and $\hat{y}_i$ is the predicted value. We can use MSE to compare different models' quality of predictions.

Another metric that is commonly used is the mean absolute error (MAE), which simply uses the absolute value of the difference, rather than the squared difference:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

In the following table, the first column `y_true` denotes the true values y_i. The second column `yhat1` holds the predicted values obtained from a model $\mathcal{M}_1$, and the third column `yhat2` holds the predicted values obtained from a different model $\mathcal{M}_2$.

y_true	yhat1	yhat2
1	0	2
3	1	3
1	0	2
-1	0	2
0	-1	1
5	3	4

a) Calculate the MSE for both models.

b) Calculate the MAE for both models.

c) Based on your results in (a) and (b), which model performs better predictions?

Exercise 5

Now, let's consider some new data with just one model for predictions:

y_true	yhat
1	0
3	2
1	0
-1	-3
0	-1
2	1

a) Calculate the MSE and MAE.

b) Now, suppose we measure a 7th point, where $y_7 = 5$ and $\hat{y}_7 = 15$. Clearly our prediction is way off! Calculate the MSE and MAE including this new observation, and comment on how they compare to the errors obtained on the original six data points in part (a).

c) Are there scenarios where we might prefer one error metric over the other?

Exercise 6

For (a) and (b), explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Also provide n and p.

a) We are foraging for mushrooms and happen to find a mushroom we have never seen before. How exciting! We would like to determine if it is poisonous based on thirty mushrooms we have previously collected. For each of these thirty mushrooms, we know its color, shape, where it was found, spore dispersal method, and poison status. **While statistical learning methods are powerful, please do not trust them to rely on identifying poisonous mushrooms!**

b) We collect data on 300 national universities, recording the average school expenditures per student, average SAT scores, tuition cost, and average family income. We would like to understand what factors might influence the average salary of alumni from these universities.

c) What are some other variables/features that may be useful for the problem described in (b)?

Exercise 7

For each of the following, indicate whether we would generally expect the performance of a flexible statistical learning methods to be better or worse than an inflexible method. Justify your answer.

a) The relationship between the predictors and response is highly complex.

b) The variance of the errors terms ($\sigma^2 = \text{Var}(\epsilon)$) is extremely high.

c) The number of predictors p is extremely large relative to the number of observations n.

d) You are confident in your assumptions about the underlying function.

Submission

Please upload your finished assignment to Canvas as a PDF (either scanned or converted document).