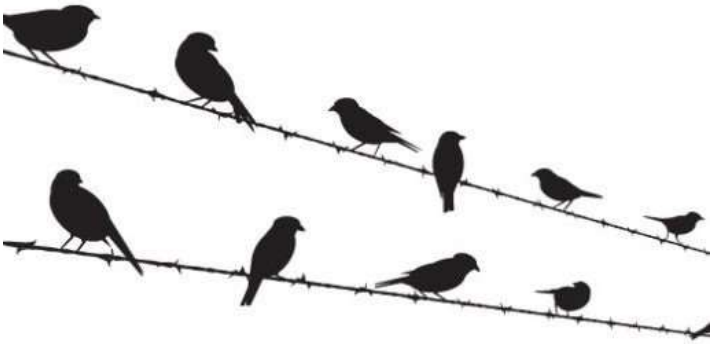

Breaking Ties in K-NN Classification

Nicholas Pylypiw
Chief Data Officer at Cape Fear Collective
4 artikler + Følg

7. juli 2017

Most of us are familiar with the expression "Birds of a feather flock together", which means that a person can, in a way, be classified by those who surround them. Memaw used this expression to warn us about the people we chose to spend time with, lest we be lumped into the same category. For example, if you choose to hang around with people with criminal records, others would assume you shared similar extracurricular interests. Some people attempt to use this to their favor, thinking that if they hang around people more successful, attractive, or hip than they are, others will project these traits onto them.

This expression is essentially the methodology at the core of K-Nearest Neighbor classification. KNN uses the status of nearby observations to determine the status of a previously unclassified observation. The parameter k represents *how many* neighbors are taken into consideration. Continuing our analogy, imagine your friends are discussing "Sue", who you have not met. You may be able to make some assumptions about Sue based on the people she is close to (you may also be *very* wrong... Memaw also warned us about judging a book by its cover). Who are the people I should consider? Sue's best friend? Her cousins? Siblings? Parents? Co-workers? This

exposes one of the problems in KNN classification. Suppose I can only choose two people to ask about Sue. I choose her best friend, and then…hmmm…who would the second be? There might be a tie between mom, dad, brother, and sister. Depending on which of these people I choose, I might get very different impressions of Sue.

Since it is easier for our brains to think and see in two dimensions, let's consider a simple two-nearest neighbor classification example with two explanatory variables (x and y) and two values of status (A and B). For every new observation, I look at the two nearest neighbors (based on Euclidean distance in the xy plane), and use them to determine the new observation's classification. There are two areas where 'ties' can create problems (Geez, how many times can I type 'two' in a paragraph?).

Neighbor Selection

A tie can occur when two or more points are equidistant from an unclassified observation, thereby making it difficult to choose which neighbors are included. In *figure i* below, it is obvious which two neighbors should be selected. However, things are more complicated in the other two scenarios. In *figure ii*, there are three neighbors which are equidistant from the unclassified observation M. Similarly, in *figure iii*, one of the nearest neighbors is obvious, but the other three have an equal chance of being selected. There are three prevailing theories on how to address this complication:

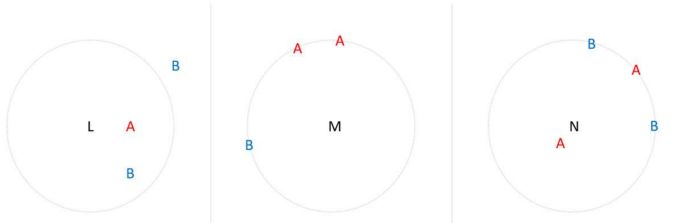

Figure i. Of the three neighbors shown, the two nearest are obvious.

Figure ii. Three neighbors are equidistant. Which two should be chosen?

Figure iii. Of the three neighbors shown, one is obvious, but which of the other two should be chosen?

1. *Choose a different k.* Although a three-nearest neighbor classification method would solve the issue of neighbor selection in *figures i* and *ii*, it does not solve the problem in *figure iii*. In fact, there still exists the possibility that selection ties occur on the greater data set regardless of k value. Additionally, I have a philosophical objection to arbitrarily changing the optimal k value simply to make life easier.

2. *Randomly choose between the tied values.* Applying this approach to *figure ii*, each of the three observations would have an equal chance of being selected as one of the two neighbors. In *figure iii*, the A closest to N would be selected, and one of the remaining three observed points would be randomly selected. Depending which of the tied values are allowed into the model could have a significant effect in how N is classified. Though I am not a fan of this method, it seems (slightly) more reasonable than the previous method.

3. *Allow observations in until natural stop point.* This, personally, is the method I prefer. The idea here is to choose the smallest number such that k is greater than or equal to two, and that no ties exist. For *figure i*, the two nearest observations would be selected (just as the previous two methods did). For *figure ii*, since there is a three-way tie, all three neighbors are considered. The constraint on only two neighbors is lifted for this particular observation. Similarly, all *four* observations are selected in *figure iii*. Note that this method would choose as many values as necessary to avoid the tie. If figure ii instead had eight equidistant points, all eight would be allowed into the model.

Classification

Let's assume we settled on the third method above, and now have our selected neighbors (*table i*). Now, we can begin classifying our unclassified points. The default classification method is majority vote, which works well for M. Since there are two As nearby and only one B, M is predicted to be an A. However, things are more complicated for L and N. Again, there are several theories of how to approach this.

Unclassified	Neighbors
L	A B
M	A A B
N	A A B B

Table i. The selected nearest neighbors for each point.

1. *Choose an odd k.* Some suggest to simply choose an odd value for k. As before, I'm not fond of the idea of choosing a non-optimal k to simply address downstream classification issues. Furthermore, this method does not always work, since the classification statuses could be odd as well (A, B, C).

2. *Randomly select between tied neighbors.* The ol' standby for statisticians everywhere, "Just pick a random one!" Using this method, L and N both have an equal chance of being classified as A or B. But is random *fair?* Does it make sense that N, who is very close to an observed A, has an equal chance of being a B?

3. *Weighted by distance.* Addressing the concern created by the previous method, it is possible to weight the neighbors so that those nearest to the unobserved point have a "greater vote". This approach would result in both L and N being classified as A, since the A neighbors are closer than the others by comparison. This method seems to address most scenarios, although it is still possible it won't solve *every issue* (Consider a four-way equidistant tie…in this case, it probably makes sense to just randomly select one of the four).

I hope this article has given you some things to think about when dealing with ties in KNN classification. Remember, only you know your data, so choose the method that is most appropriate for your use case. Please comment with questions, corrections, additions, disagreements, etc. Also, check **this** out if you're interested in a more code-based exploration. Thanks for reading!

Tilføj en kommentar ...

Nicholas Pylypiw

Chief Data Officer at Cape Fear Collective

+ Følg

Mere fra Nicholas Pylypiw

Modeling Attack Probability for the Game of Risk - An R Shiny Project

Nicholas Pylypiw på LinkedIn

Five Common Mistakes Made by Data Scientists

Nicholas Pylypiw på LinkedIn

Are You Hiring Problem Solvers? An Exercise in Pizza and the Atomic Bomb

Nicholas Pylypiw på LinkedIn