Learning with Nearest Neighbor

Andrew Kirby andy-kirby@live.com
Kevin Browder browderkevin54@gmail.com
Nathan Stouffer nathanstouffer1999@gmail.com
Eric Kempf erickempf123@gmail.com

Abstract

yeet

1. Problem Statement

Hypothesis

2. Algorithms

2.1 Distance Metrics

The nearest neighbor rule, and therefore K-NN, requires a distance metric to determine the theoretical "distance" between two examples. A common distance metric—and the one used in this paper—is Euclidean distance. The Euclidean distance D between two examples x_1 and x_2 is computed as:

$$D = \sqrt{\sum_{i=0}^{d}(a_1^i - a_2^i)^2}$$

where d is the number of attributes the examples have, a_1^i is the i-th attribute of x_1, and a_2^i is the i-th attribute of x_2.

A Euclidean distance metric assumes that all data in a dataset is continuous. However, in the world of data science and this project, some attributes contain categorical values. One method to compute this distance is the Value Difference Metric (VDM). Note that the VDM relies on classification to compute a distance, so continuous regression values must be discretized by some method.

To compute a distance between two categorical values x_1, x_2 within one attribute a using the VDM, find the following for each value:

$$v_n = \sum_{i=1}^{k} |\frac{N_{1i}}{N_1} - \frac{N_{2i}}{N_2}|$$

where N_{1i}, N_{2i} are the number of examples in a that are of the i^{th} class and have values x_1, x_2 respectively, and N_1, N_2 are the number of examples in a that are of the i^{th} class.

Then compute the difference $v_2 - v_1$. This difference is considered to be the distance v between the two categorical values x_1 and x_2 (Stanfill and Waltz, 1986). The squared difference can now be included in the summation when computing Euclidean distance.

3. Experiment

3.1 Preprocessing Choices

3.2 Tuning

4. Results

5. Summary

References

Miloud-Aouidate Amal and Bab-Ali Ahmed Riadh. Survey of nearest neighbor condensing techniques. *IJACSA International Journal of Advanced Computer Science and Applications*, 2(11), 2011.

T. Cover and P. Hart. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, 13(1):21–27, January 1967. doi: 10.1109/TIT.1967.1053964.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Sahibsingh A Dudani. The distance-weighted k-nearest-neighbor rule. *Institute of Electrical and Electronics Engineers Transactions on Systems, Man, and Cybernetics*, (4):325–327, 1976.

Wendy R. Fox, Lowell Kaufman, and P. J. Rousseeuw. Finding groups in data: An introduction to cluster analysis. 1990.

Matthew D Mullin and Rahul Sukthankar. *Complete Cross-Validation for Nearest Neighbor Classifiers*. 2000.

Craig Stanfill and David L Waltz. Toward memory-based reasoning. *Commun. ACM*, 29 (12):1213–1228, 1986.

K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning*, 2001.

D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *Institute of Electrical and Electronics Engineers Transactions on Systems, Man, and Cybernetics*, SMC-2(3):408–421, July 1972. doi: 10.1109/TSMC.1972.4309137.

Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Debo Cheng. Learning k for knn classification. *Association for Computing Machinery Transactions on Intelligent Systems and Technology (TIST)*, 8(3):43, 2017.