News Article Classification

Lauren Contard, Archit Datar
Yue Li, Robert Lumpkin, Haihang Wu

1 Introduction and Problem Statement

2 Methods

2.1 ML-kNN

ML-kNN (Multi-label k nearest neighbors) is derived from the traditional k nearest neighbors (kNN), except for the multi-label case. While the goal of the traditional kNN algorithm is to predict whether class of the test sample based on the classes of its k nearest neighbors, the goal of ML-kNN is to predict multiple classes based on the classes of the k nearest neighbors of the test point. For the unseen data point, its nearest neighbors are identified. Then, based on the number of neighboring instances belonging to each possible class, maximum a posteriori (MAP) principle is utilized to determine the label set for the unseen instance.

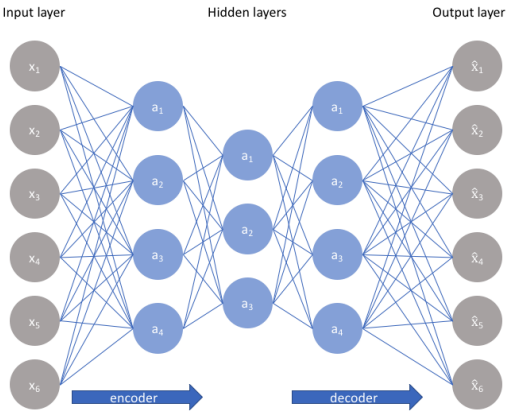
ML-kNN is used in a variety of problems, such as, text categorization [5], where each document may belong to several topics, such as the use case for our project. Apart from this, it can also be useful in areas such as functional genomics where each gene may be associated with a set of functional classes [2], and in image classification, where each image could have multiple genres.[1]

2.2 Linear Dimension Reduction (PCA)

2.3 Nonlinear Dimension Reduction (ANN Autoencoder)

In addition to reductions in dimension due to PCA, we also implement an ANN autoencoder. Autoencoders can learn data projections with suitable dimensionality and sparsity limitations that are more useful than other fundamental methods such as PCA, which only allow for linear data representations [4].

This nonlinear dimension reduction is done by by training a feed forward neural network to perform the identity mapping, where the network inputs are reproduced at the output layer. The network contains an internal "bottleneck" layer (containing fewer nodes than input or output layers), which forces the network to develop a compact representation of the input data, and two additional hidden layers [3]. Look to the diagram to the right, for a visualisation of this set-up.



The particular network that we trained had three hidden layers, as in the diagram. The first, second and third hidden layers are of dimensions 128, 64, and 128, and use the activations tanh, ReLu, and sigmoid, respectively. Training was performed using Adam optimization, MSE loss, and over 400 epochs. After training, the generated encodings were used to repeat our model fitting procedures for both the binary-relevance KNN and ML-KNN algorithms.

2.4 Artificial Neural Networks (Feed-Forward & Recurrent)

3 Results

3.1 ML-KNN Results

3.2 Artificial Neural Network Results

4 Discussion & Conclusions

References

[1] Matthew R. Boutell et al. *Learning multi-label scene classification*. 2004.

[2] André Elisseeff and Jason Weston. "A Kernel Method for Multi-Labelled Classification". In: *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*. NIPS'01. Vancouver, British Columbia, Canada: MIT Press, 2001, pp. 681–687.

[3] Mark A. Kramer. "Nonlinear Principal Component Analysis Using Autoassociative Neural Networks". In: *AIChE Journal* 37.2 (1991), pp. 233–243. DOI: https://doi.org/10.1002/aic.690370209.

[4] Mohamad Aljnidi Kadan Aljoumaa Maha Alkhayrat. "A comparative dimensionality reduction study in telecom customer segmentation using deep learning and PCA". In: *Journal of Big Data* 7.9 (2020). ISSN: 2196-1115. DOI: https://doi.org/10.1186/s40537-020-0286-0.

[5] Andrew Kachites McCallum. "Multi-label text classification with a mixture model trained by EM". In: *AAAI 99 Workshop on Text Learning*. 1999.