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.

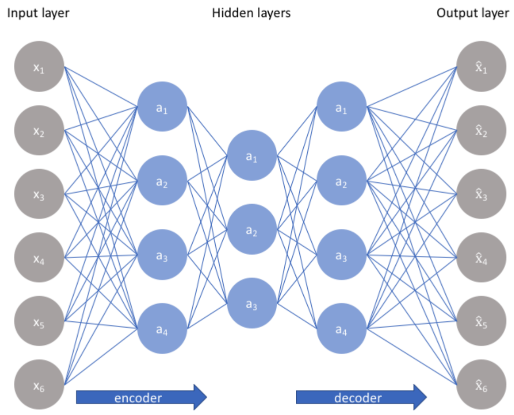
ML-kNN is used in a variety of problems, such as, text categorization [6], where each document may belong to several topics, such as the use case for our project. Apart from this, it can also be useful in areas such as functional genomics where each gene may be associated with a set of functional classes [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 (see section 2.4 for an introduction to ANNs). 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 [5].

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 [4]. 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)

Approaches utilizing different artificial neural networks are also utilized in our project. Inspired by biological nervous systems, neural networks date back to the first half of the 20^{th} century with works such as those by McCulloch and Pitts, which could model simple logical operations [8]. Since most subsequent work in the following two decades centered around single layer networks, the power of neural networks was restricted to linearly separable problems. This excluded the possibility of learning even simple functions like XOR, which required a second layer [3]. In the early 1980s, research on neural networks resurged largely due to successful learning algorithms for multi-layer neural networks and are used today for various tasks such as computer vision, associative memory, representation learning, NLP, etc..

For our project, we tried implementing both feed forward and recurrent neural network architectures. The feed forward architectures were trained on 2094 dimensional tf-idf vectors. The recurrent networks were trained on padded sequences of integers corresponding to sequences of the processed tokens from the paragraphs; urls, punctuation, and stop words were all removed.

Training was performed, using two different loss functions. The standard binary cross entropy loss function was used, in addition, to the more novel BPMLL (back prop for multilabel learning) loss. The BPMLL loss function requires instances to have at least one label. Thus, on the one hand, we trained both cross entropy and BPMLL models on such a reduced dataset in addition to training just cross entropy models on the full dataset.

The BPMLL loss function aims to leverage correlations between labels by evaluating the error as a function of pairwise errors between labels. Namely, the loss function is given by:

$$E = \sum_{i=1}^{m} E_i = \sum_{i=1}^{m} \frac{1}{|Y_i||\overline{Y}_i|} \sum_{(k,l)\in Y_i\times\overline{Y}_i} \exp(-(c_k^i - c_l^i))$$

where $c_j^i = c_j(x_i)$ is the output of the network on x_i on the j^{th} class. The back propagation algorithm is derived in the same manner is for a cross entropy or MSE loss. Details are ommitted here, but can be found in [7].

All of our networks utilize the same optimization algorithm (Adam optimization). All of our feed forward networks utilize the same architectures; the first hidden layer is of size 32 with ReLU activation, dropout regularization with drop probability of 0.5, and an output layer of size 13 (the number of labels) with sigmoid activations. On the reduced dataset, all of our RNN models utilize the same bidirectional LSTM architecture. On the full dataset, all of our models utilize the same single-directional LSTM architecture.

2.5 Threshold Function Learning

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] Simon Haykin. *Neural Networks and Learning Machines*. Pearson, 2009.

[4] 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.

[5] 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.

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

[7] Min-Ling Zhang and Zhi-Hua Zhou. “Multilabel Neural Networks with Applications to Functional Genomics and Text Categorization”. In: *IEEE Transactions on Knowledge and Data Engineering* 18.10 (2006), pp. 1338–1351. DOI: doi:10.1109/TKDE.2006.162.

[8] Gualtiero Piccinini. “The First Computational Theory of Mind and Brain: A Close Look at McCulloch and Pitts' *Logical Calculus of Ideas Immanent in Nervous Activity*”. In: *Synthese* 141.2 (2004), pp. 175–215. DOI: 10.1023/B:SYNT.0000043018.52445.3e.