Paper Template for COMP30027 Report

Anonymous

1 Introduction and background

2 Support vector machine approaches

We investigated three main types of support vector machine (SVM) classifier: a SVM with a linear kernel and one-verses-rest multi-class classification (model A), a SVM model with a radial basis function (RBF) kernel and one-verses-rest multi-class classification (model B), and a SVM with a linear kernel separating the positive and negative sentiment classes, in which instances close to the boundary are classified as having neutral sentiment (model C).

In model C, rather than using the raw distance to the hyperplane to determine marginal instances, we opted to use a variant of SVM which estimates the probability of an instance belonging to its class described in (Platt, 1999).

In all three cases paragraph vectors were used as features. Link to next section

2.1 Selection of main models

Link to previous section For support vector machine classifiers we identified three broad hyperparameters which would result in different models: (1) The features selected, (2) The choice of kernel, and (3) How to approach the multi-class problem.

For (1) we chose to use the paragraph vector encoding (Le and Mikolov, 2014) since paragraph vectors have meaningful geometric relationships to one another. This makes them well suited to classification by a SVM model, which attempts to exploit the geometry of the feature space by fitting a hyperplane that separates classes.

For (2), given our choice of paragraph vectors as features, we expected that a linear kernel would be best suited to this task. This is because paragraph vectors are designed to transfer meaning to linear relationships between vectors: "king" - "man" + "woman" = "queen" (Le and Mikolov, 2014). If we suppose that 1 and 5 star reviews have opposite meaning (with 3 star reviews somewhere in between) we would expect these classes to be linearly separable in the paragraph vector encoding. Preliminary testing of SVM models with linear, radial basis function (RBF), polynomial and sigmoid kernels showed that the linear and RBF kernels had the best performance across all metrics.

For (3) we considered using the standard multi-class approaches for SVM models (classifying classes one-verses-one or one-verses-rest) in contrast to using a binary classifier to separate the 1 and 5 star classes (the negative and positive sentiment classes respectively), and classifying instances close to the boundary as neutral sentiment (3 star). While the latter approach is tempting, especially considering the purported ability of paragraph vectors to capture meaning through linear relationships, results in the literature suggest that this will always result in worse results than a standard multi-class approach (Koppel and Schler, 2006). Nevertheless, we decided to pursue this approach as well. We opted to use a probabilistic variant of SVM to determine marginal instances (rather than the distance to the hyperplane) primarily because it makes the threshold hyperparameter less dependent on the dimension of the feature space. If this threshold were distance-based we would not expect distances in one feature space to correspond to distances in another feature space, making tuning these parameters much more difficult.

2.2 Tuning hyperparameters

We identified the following hyperparameters for each model.

Model A The dimension of the feature vector encoding, the degree of regularisation.

Model B The dimension of the feature vector encoding, the degree of regularisation, the

kernel coefficient.

Model C The dimension of feature encoding, the degree of regularisation, and the kernel coefficient.

The dimension of the feature space can broadly be interpreted as a measure of the complexity of our model. As the dimension of the paragraph vector encoding a review increases, the degree to which the paragraph vector can capture the information in the review will increase too. Not all of the information in a review is likely to be useful for this classification task however[1], so there is a point at which increasing the dimension of the feature space will lead to over-fitting and degrade the performance of the classifier.

Due to the computational expense of producing paragraph vector encodings (producing cross-validation splits for many dimensions is out of the question) we opted to treat this hyperparameter somewhat differently to the others. For dimensions 25 to 300 in steps of 25, we computed the feature vectors of a 80:20 random holdout, only using the training set in each case to find the feature vector encoding. We then plotted the learning curve for each model (with other hyperparameters unturned) and identified the optimal dimension for that model. We then computed paragraph vector encodings for a 5-fold cross validation split at that dimension, again only using the training set of each split to produce the encoding. This cross-validation split was used to tune the remaining hyperparameters. these splits were all stratified

The remaining hyperparameters were found using the precomputed cross-validation split using a grid search. Since each cross validation run took a long time parameter values were initially adjusted in large increments, and then a finer full grid search was done on regions of interest. The results of this can be seen in Figure insert grid search figure.

References

Moshe Koppel and Jonathan Schler. 2006. THE IMPORTANCE OF NEUTRAL EXAMPLES FOR LEARNING SENTIMENT. 22(2):100–109.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents.

Arjun Mukherjee, Vivek Venkataraman, Bing Liu, and Natalie Glance. 2013. What yelp fake review filter might be doing? In *7th International AAAI Conference on Weblogs and Social Media*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in python. 12:2825–2830.

John Platt. 1999. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. 10(3):61–74. Publisher: Cambridge, MA.

Shebuti Rayana and Leman Akoglu. 2015. Collective opinion spam detection: Bridging review networks and metadata. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 985–994.

Radim Řehůřek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50. ELRA.

[1]Review text also may contain information such as the cuisine of the restaurant, the time-of-day of the visit and the gender of the wait-staff, much of which is likely irrelevant to the task at hand.