Milestone 3: Traditional statistical and machine learning methods, due Wednesday, April 19, 2017

Think about how you would address the genre prediction problem with traditional statistical or machine learning methods. This includes everything you learned about modeling in this course before the deep learning part. Implement your ideas and compare different classifiers. Report your results and discuss what challenges you faced and how you overcame them. What works and what does not? If there are parts that do not work as expected, make sure to discuss briefly what you think is the cause and how you would address this if you would have more time and resources.

You do not necessarily need to use the movie posters for this step, but even without a background in computer vision, there are very simple features you can extract from the posters to help guide a traditional machine learning model. Think about the PCA lecture for example, or how to use clustering to extract color information. In addition to considering the movie posters it would be worthwhile to have a look at the metadata that IMDb provides.

You could use Spark and the ML library (https://spark.apache.org/docs/latest/ml-features.html#word2vec) to build your model features from the data. This may be especially beneficial if you use additional data, e.g., in text form.

You also need to think about how you are going to evaluate your classifier. Which metrics or scores will you report to show how good the performance is?

Detailed description and implementation of two different models

Baseline models:

We will use One vs Rest Classification with various sets of movie data:

- independently training one binary classifier for each label.
- predict all labels for a new sample for which the respective classifiers predict a positive result

We will use Support Vector Machines and Random Forest Classifiers for the OneVsRestClassification.

Model 1: SVM/RF on the IMDB/TDMB movie metadata

Movie features including release data, rating, stunts etc will be used as the input to the OneVsRestClassifier with both SVM (linear and radial kernels) and RF. The parameters of SVM and RF will be tuned using cross validation.

We will also experiment with using a multi layer perceptron with 1-2 hidden layers

- output layer has n sigmoid activation units where n is the number of unique genres
- output is a binary vector with 1s indicating the movie has the movie genre given by that position in the label vector

Model 2: RF on the IMDB keyword data

We will first create a one hot encoded matrix for all the unique keywords (~12,000) and use a separate RF classifier for each genre. We can then use the RF feature importances to determine which keywords are most relevant in classifying each genre. This can subsequently be used as a dimensionality reduction technique - we can select the top 10 keywords from the feature importances for each genre and add them to the movie metadata.

In addition we can also use a OneVsRestClassifier on the one hot encoded keyword vectors to predict a multilabel output.

Model 3: PCA on the IMDB movie posters

We also plan to use the movie posters in a classification model. We will first apply dimensionality reduction using PCA, retaining the components which contribute to 90% of the variance.

The training and testing data can then be projected onto the top principal components and this can be used as the input into a OneVsRestClassifier using SVM/RF.

Description of your performance metrics

We will use three metrics to evaluate the performance of our classifiers:

- Hamming loss: the fraction of the wrong labels to the total number of labels
- Percent at least one match: proportion of movies for which at least one genre was predicted correctly
- Percent exact match: percentage of movies that have all their genres correctly predicted

Careful performance evaluations for both models

iteration	hamming loss	percent exact match	percent at least one match
Tuned Radial SVM on movie metadata	0.082	63.05%	88.8%
Tuned RF on movie metadata	0.088	61.5%	95.3%
ANN on movie metadata	0.248	13.1%	96.8%
RF on keywords	0.194	27.5%	75.2%.
PCA on posters + untuned SVM	0.212	13.4%	100%

Discussion of the differences between the models, their strengths, weaknesses, etc.

Comparing the metrics from the different models the following observations can be made:

- Tuned Radial SVM has the lowest Hamming loss and highest percentage of exact matches. Both these metrics indicate low error rate.
- Tuned RF also has similar performace to the radial SVM with higher percent of at least one label matching without much decrease in the percent exact match.
- Untuned ANN on the movie metadata has high percent of at least one label matching however it has very low percent exact match and the highest hamming loss out of all the models.
- RF on the keywords has a higher hamming loss indicating higher fraction of mislabelling and also has low percent exact match.
- PCA overpredicts some of the genres, result in 100% of a least one label matching but a very low number of exact matches (this could also be due to the smaller data set used for PCA due to