Week 8 Practical: Aerosol Prediction Challenge
SM4 Data Mining and Machine Learning (HT 2017)

Dino Sejdinovic

Reports are due by Tuesday 5pm, March 14th 2016

1 Task

This practical has an associated kaggle-in-class challenge. It concerns the prediction of aerosol optical depth (AOD) based on spectral imaging data collected by *Moderate Resolution Imaging Spectroradiometer (MODIS)* aboard NASA's satellites. This is a regression problem with a real-valued response. However, each input item (to which the labels are associated) is not a single data vector but consists of multiple observations. We will refer to the input items as *bags* and to the observations within each bag as *instances*. The original dataset is a modified version of the MODIS dataset available here and it has been used in [1, 2].

You are free to use any machine learning method and model discussed in the course (and beyond), as long as you describe clearly in the report all the steps and choices you have made. Note that the data you have been given is a shuffled and transformed version of the original data, so looking for true labels for your test data in the original MODIS dataset will not be helpful.

While getting a good prediction performance of your method will be important, remember that you will be assessed based on the quality of your report, so explaining your steps and choices clearly and discussing all the issues you have faced in this challenge will be essential.

The report has a word limit of 2500 words.

2 Data

Data consists of multiple instances per item (bag of instances). Each bag consists of 100 instances representing randomly selected pixels around the AERONET site.

The file `X.csv` corresponds to the input data (both training and testing items):

- Number of bags: 1364
- Total instances: 136400
- Features: 12
- Columns 1-7: instance-level features (7 MODIS reflectances).
- Columns 8-12: bag-level features (5 solar and view zenith angles).
- Column 13 is the bag Id

The file `ytr.csv` containst the training responses and the corresponding bag Ids. Response y is the bag label (AOD measured by the AERONET instrument). You are given labels for the first 980 bags and the goal is to predict labels for the remaining 384 bags, using squared loss.

3 Submission and Evaluation

Each team should use a single kaggle account. You can make up to 3 submissions every day.

Submission Format. Submission files should contain two columns: `Id` and `y`. `Id` refers to the indices of the test bags so these should be $981, ..., 1364$. `y` is your prediction.

The file should contain a header and have the following format:

```
Id,y
981,0
982,0
983,0
984,0
...
```

Evaluation Metric. Accuracy will be assessed in terms of the root mean square error (RMSE) on the held out test set (lower is better):

$$\sqrt{\frac{1}{n_{tst}} \sum_{i=1}^{n_{tst}} (y_i - f(x_i))^2}.$$

When submitting the predictions to kaggle, you will instantly see the RMSE of your method on approximately 30% of the test set. At the end of the competition, the results on the whole test set will be available.

References

[1] Z. Wang, L. Lan, and S. Vucetic. Mixture model for multiple instance regression and applications in remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 50(6):2226–2237, 2012.

[2] Zoltán Szabó, Arthur Gretton, Barnabás Póczos, and Bharath K. Sriperumbudur. Two-stage sampled learning theory on distributions. In *Proc. International Conference on Artificial Intelligence and Statistics, AISTATS*, 2015.