Exercises for the Applied Machine Learning Lecture Summer 2016

Problem Set 6

Bernhard Renard - renardB@rki.de

May 23, 2016

This week's exercise focuses on implementing and evaluating a simple active learning scheme based on random forest classification and applied to an imaging mass spectrometry tumor data set.

Step 1
Load all data from http://www.renard.it/mstat/tumor.zip and unzip. Load the imaging mass spectrometry tumor dataset of patient 1. This dataset contains measurements of 118 different mass channels for 440 spatial positions. Further, load the labeling data; a label of 0 represents connective tissue, a label of 1 tumor tissue.

Step 2
Randomly subsample 3 to 100 data points from patient 1 for training and visualize the correlation of training and test error.

Step 3
For this dataset, develop an active learning classifier that only regards insecurity of classification. Thus, always choose that data point that is most undecided (randomly select for ties). Retrain after each addition of a datapoint (start with as little as one positive and one negative example) and visualize the development of training and test error. Also visualize the spatial positions of the chosen data points.

Step 4
Extend your classification strategy by including spatial information and ensure that no spatially close data points are considered. Choose a suitable method (e.g. clustering information, spatial distance between data points).

Step 5
Visually represent your results. Comment on the following aspects: Starting from which number of data points are you more successful with your active learning approach? Starting from which number is the difference negligible?

Please email your solution to the problem set to RenardB@rki.de until Wednesday, **June 1st, 2016, 9am**. You cannot get credit for solutions turned in late and there are no exceptions from this rule. Feel free to work in groups of two and hand in a single joint solution with both names clearly stated. Feel free to discuss with other groups, but each group has to hand in their own solution in their own words (and please do not plagiarize or copy from the internet). Please note that we cannot give an active participation passing grade to those cheating.