# Active machine learning for classification

Background

We consider a classification task. There are situations where a large amount of unlabeled data is available. It is possible to label some of the observations, but since this is demanding or involves human expert knowledge, only a small part of the data can be labeled. Examples for this situation appear e.g. in text or image classification or in classification of ECG recordings.

One approach would be to select randomly a small subset of the unlabeled datapoints and label them. In contrast, active machine learning selects datapoints which are most informative and requests their labeling. With this, the classification accuracy can be improved in situations where labeling is expensive.

In this Master’s project, you compare different strategies to select the observations to be labeled. You should consider also different classification methods, since the performance of the selection strategy can depend on the classification method used.

Data description

The thesis is aiming for comparison of methods and is initially not connected to a fixed dataset. As a dataset to start with, a bank marketing dataset with around 40000 observations and 20 features might be used with the aim to predict if a client will subscribe a term deposit. However, you are free to use other dataset(s) of your choice.

Research questions

* How do different strategies to select observations perform in the context of a given classification task?
* Does the performance of the selection strategies depend on the classification method used?

Required background

Interest in this field; you need not to have pre-knowledge in active machine learning

Contact details

Frank Miller, [frank.miller@liu.se](mailto:frank.miller@liu.se), Tel: [013-28 14 76](tel:+4613281476)