Tutorial for the Lecture Applied Machine Learning

Freie Universität Berlin, SoS 2016
Annalisa Marsico · Stefan Budach

Sheet 11 · Assignment on 06.07.2016
Please email your solution (pdf + working R script) for the problem set to
budach@molgen.mpg.de until Wednesday 13.07.2016 , 9am.

Introduction and goal of the tutorial

Application of Artificial Neural Networks (ANNs) to cancer location prediction

Micro-RNAs (miRNAs) are evolutionary-conserved non-coding RNAs, approximately 22 nucleotides long, which have been shown to be crucial post-transcriptional regulators of gene expression in metazoans and plants, targeting up to 50% of the protein-coding genes by means of base complementarity. Regulation of messenger RNAs by means of miRNAs induces either mRNA repression (translational inhibition) or degradation, resulting in a fine-tuned down-regulation of miRNA targets. It has been shown that miRNA-mediated regulation is highly time- and tissue-specific, therefore miRNAs have emerged as a class of modulators involved in cancer that might prove to be important predictors of disease risk, location and progression. In addition, different miRNA expression profiles (e.g. from microarray data) can be associated to different cancer sub-types and can be useful as diagnostic features to discriminate such sub-types.

The goal of this tutorial is to perform tumor stratification, i.e. we want to classify patients into multiple classes according to tumor location by using miRNA expression data only. The expression of hundreds of miRNA has been measured for each patient.

We provide new expression data, specific for microRNAs, as well as clinical data and patient labels for human colon and rectal cancer (or cancer location). The data come from the following publication (from the first tutorial):

`Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. Nature. 2012 Jul 18;487(7407):330-7. doi: 10.1038/nature11252. PubMed PMID: 22810696; PubMed Central PMCID: PMC3401966.`

and can be downloaded at the following link:
`http://www.molgen.mpg.de/~mlc/a4-data.tar.gz`

Assignment 1. *Prediction of cancer location prediction based on miRNA expression data*

1. In order to speed up network training, it is better to reduce the number of input features (i.e. the number of miRNAs used for training). Pre-process the data by eliminating low variance features (hint: for each miRNA compute the coefficient of variation among patients) and correlated miRNAs.

2. Decide on a basic architecture of your Neural Network for predicting cancer location. This is a multi-class problem (column "tumor_site"). How many input neurons do you need, how many output neurons?

3. Train the network and use 5-fold cross-validation with different parameter settings, such as initial weight values, number of hidden neurons, number of iterations etc. (hint: use `nnet` R package).

4. Compute the accuracy of the model with the best set of parameters (hint: compute the accuracy simply as the percentage of samples for which their class is predicted correctly, ROC curves are complicated for a multi-class problem).

5. Check if your network is perhaps overfitting by plotting the error function at each iteration step (or every 10 iteration steps) for the training set and for the test set. Hint: as the computation of the error function is not part of the nnet package you will need to implement the formula by yourself. Here some suggestion code (substitute the RSME function with the correct one introduced in the lecture, as this is a classification problem).

   `https://beckmw.wordpress.com/2013/03/19/animating-neural-networks-from-the-nnet-package/`