Data Mining and Machine Learning in Bioinformatics

Dr. Holger Fröhlich, Dr. Martin Vogt
Sabyasachi Patjoshi
sabyasachi2k13@gmail.com, martin.vogt@bit.uni-bonn.de

Due: Jul 15, 10:30 (by the end of the lecture)

Exercise Series 11

General: Exercises are to be solved and submitted in fixed groups by at most 3 students. Every member of a group should help solving each task and thus be able to answer questions to each task. No late submissions are accepted. Copying solutions will automatically lead to a point reduction of at least 50%. N – 1 homework assignments and N – 2 programming tasks have to be submitted in total.
A group can gain additional bonus points, if it presents its solution for a particular task during the tutorials. Accordingly, each task has a defined number of points as well as bonus points.

1. ***k* nearest neighbor (*k*-NN) classification** is a very simple classification method where a label for a test instance is determined by calculating its distance to all training instances and considering only the *k* closest instances. The test instance is then predicted to have the label of the majority of the *k* closest instances. *k* is a constant parameter, e.g., 1 (only consider the nearest neighbor), 3, or 5. (Choosing an odd parameter *k* will avoid the possibility of ties in binary classification assuming no distances are tied.)
   a) Implement a k-NN classifier in R. **Tip**: use function dist to calculate a distance matrix. (3 points + 1 bonus point)
   b) Take the colonCA dataset introduced in exercise 6. Randomly take a subset of 50 persons for training and the rest for testing (make sure that not all belong to the same class). Which **sensitivity, specificity, precision, balanced accuracy, MCC and F1** does a 1-NN, 3-NN and 5-NN classifier achieve on (i) the training and (ii) the test set. (3 points + 1 bonus point)
   c) In b) all genes were used for classification. Now, based on the training data, select only the 20 genes showing the largest difference in their average expression between cancer and normal. Which changes in the result from b) do you observe? Interpret your findings. (3 points + 1 bonus point)
   d) Implement a **10-fold cross-validation** procedure (repeated 10 times) and evaluate your 3-NN classifier using the gene filtering strategy from c). Do this in two ways:
      i) Gene selection prior to cross-validation
      ii) Gene selection within the cross-validation procedure (i.e. as part of classifier training)

Which differences in terms of sensitivity, specificity, precision, balanced accuracy, MCC and F1 do you observe? Make box plots to visualize the distribution of each of these performance measures and explain the reason for your finding. **Hint**: In order to make these plots comparable you should ensure that the value range on the y-axis is the same for each plot (argument ylim). (6 points + 1 bonus point)

2. Consider the **LASSO regression** introduced in the lecture.

a) In which way are **bias** and **variance** of the LASSO model influenced by the **regularization** parameter λ? (4 points + 1 bonus point)

b) Suppose you have the choice to fit either a LASSO or a ordinary least squares regression model. Which of both models will have

i) less biased coefficient estimates?

ii) a larger variance of model predictions?

Give reasons for your answer (4 points + 1 bonus point)

c) Now compare the LASSO to a ridge regression. Which of both models yields higher sparsity (i.e. more coefficients being set to 0)? Give reasons for your answer. (4 points + 1 bonus point)