Sven Hauns, Anup Kumar, Simon Bray
Machine Learning in Life Sciences
May 19, 2022

Exercise2: Support Vector Machines

You can find the programming solution to this exercise in the github repro.

Question 1. Breast cancer is one of the most frequently diagnosed cancers in women, in which the cells in the breast grow uncontrollably. This growth is determined by mutations in certain genes. A mutation that creates a fundamental growth advantage is called a driver mutation, and the associated gene is called a driver gene. Passenger mutations, on the other hand, do not necessarily determine the development of cancer. Here we use SVMs to classify candidates into driver and passenger genes. Learn more about the genetic background of breast cancer here.

Download the dataset breast cancer genes. Split the created dataset into training and test datasets to evaluate your model. We have a very unbalanced dataset. What does this mean for the class weights?

Question 2. As we learned in the lecture, SVMs only need to know the relationship between samples to build a classification model. In scikit-learn, the Gram matrix is computed by the support vector machine, so we can pass our data directly. Create a classifier with:

(a) a soft-margin SVM
(b) a hard-margin SVM

Try out different kernel functions! Choose an appropriate evaluation metric. How do you explain the differences between soft-margin SVM and hard-margin SVM?
Perform hyper-parameter optimization using grid search (on the Hyperparameters C and kernel of sklearn) Which values perform best?

(a) create a box plot showing the distribution of a metric for one parameter.
(b) use the best model to find possible driver genes in a dataset of candidates

How many possible driver genes did you identify? Rank candidate genes in the candidate dataset and show top 20 candidates predicted by the model.

Question 3. Explain some advantages and disadvantages of grid search. When would it be advantageous to use a random search instead?

Question 4. Genetic disorders can be caused by mutations in a single or multiple genes. Mutations can either occur spontaneously or be inherited from the parents.

(a) What is the difference between gene prioritization and gene identification?
(b) Huntington's disease is a autosomal dominant disease. What is the difference between dominant and recessive inheritance? Give an example of a disease with recessive inheritance.
(c) What is genetic linkage? Give an example of two genetic disorders that display linkage
(d) Why can a disease such as depression not be analysed in the same way as Huntington's disease? Define GWAS. How can it be used to study such disease? What are the limitations of GWAS?
(e) What do the terms genome and transcriptome mean? What is the advantage of TWAS over GWAS?

Question 5. Answer the following questions regarding SVMs:

(a) Can you briefly explain the main function of an SVM?

(b) Can you explain the terms maximal margin, optimal hyperplane and support vectors in the context of SVMs?

(c) What are Mercer's condition for kernel functions?

(d) What is the kernel-trick? What is the advantage of using it?