1. What is the general procedure when classifying data with support vector machines?

Given a tool to predict (machine learning model), which is then trained on a set of training examples. Then, we test the model with another set in order to get the accuracy.

2. Define with your own words supervised and unsupervised learning and point out the difference(s). Give 2 example methods for each.

Supervised learning is the process in machine learning, where the data is previosly labelled, which means that the data is classified and to every classified input there will be a desired outpu. In contrast, unsupervised learning is unlabelled, the user does not classify the data before training the model with the data. Additionally, in unsupervised learning, the accuracy cannot be evaluated, which it can be for supervised learning.

supervised learning: linear regression, k-nearest neighbors algorithm

unsupervised learning: usage for protein clustering (k-means), independent component analysis

3. What is cross-validation?

Basically, cross-validation is the process, when there is a limited amount of data and then one can divide the data in two parts and use the first part for training or producing a model and the other part to test the model previously created.

4. What does a line in any of these files correspond to?

Each line corresponds to training examples, in this case each line represents a protein consisting of amino acids.

5. What is the meaning of a -1 in the first column in the file train25_mini?

The number refers to the label, it can be 1 (for positive example) and -1 (for negative example). In this case it means that the following line (after -1) is a negative training example

6. What is the meaning of the 3:1 in the first line of train25_mini?

That means, that in the given protein there are one time the aminoacid aspartic acid.

7. Train an SVM model on train25_mini. Then test the performance of this model on test25. What accuracy did you get?

87.00%

8. Use the svm_model from question 7 and test it on train25_mini. What is the accuracy? Is this a good way of testing an SVM model?

90.00%

9. Train an SVM model on a larger training data set, train25, and then test this model on the set test25. What accuracy did you get?

94.00%

10. Train an SVM model on train100 and test it on test100. What is the accuracy?

75.00%

11. Do you get a better classification by training and testing with the first 25 residues or with the first 100 residues? How would you explain this result?

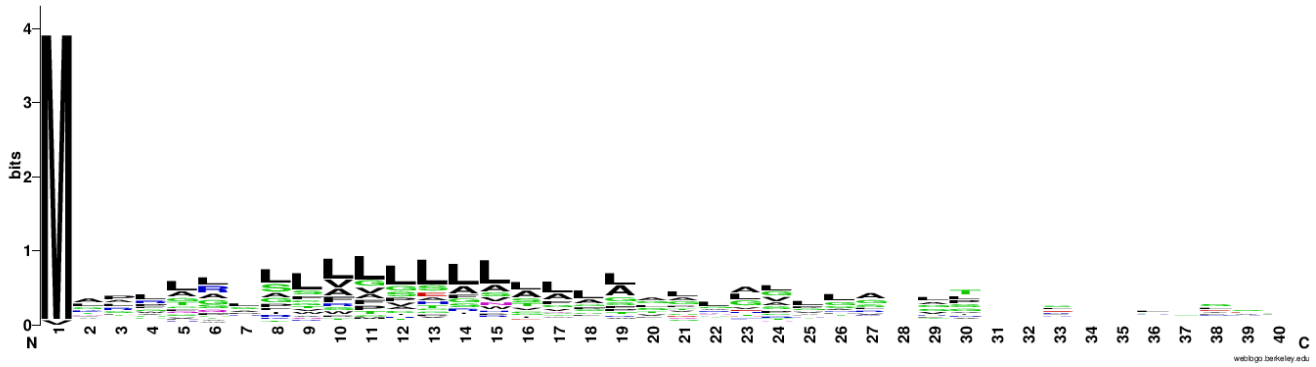
The accuracy for the first 25 residues is significantly higher with 94.00% in comparison to the 100 residues with 75.00%, so the classification is much higher with the first 25 residues. The reason for that is, that the default kernel function (linear regression) fits better for the 25 residues set. The more data points there are, the higher is the chance that the linear regression does not fit anymore, or that the seperation is more difficult, leading to a decreased accuracy.

12. There are different kernels that can be used when creating an SVM model using svm_learn (see different svm_learn options by running svm_learn --help). The svm_learn flag for selecting a kernel is called -t. Which kernel is used by default? Which kernel gives the highest accuracy when using train25 for building an SVM model and test25 for testing the SVM model?

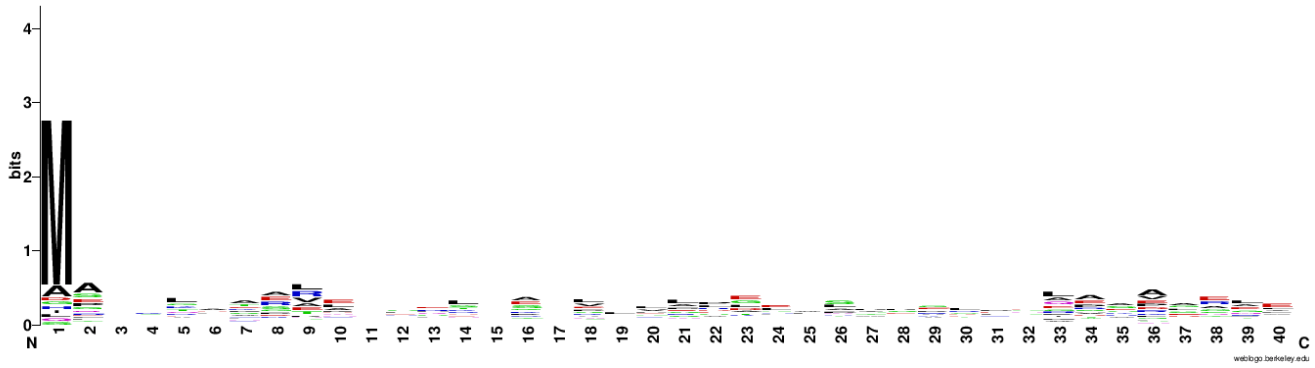
-t type of kernel function: 0 (default) linear
-t 1 -> polinomial 94.00%
-t 2 -> radial basis function 52.00%
-t 3 -> sigmoid tanh 51.00%
-t 4 -> new invented function 49.00%

13. A sequence LOGO is generally created to compare the different positions in a multiple sequence alignment in terms of information content. The higher the letters at a certain position in the LOGO, the more informative or conserved this position is in terms of sequence evolution. You can create sequence LOGOs using the online tool WebLogo. Create two LOGOs, one for the sequences in the attached file signal_sequences.txt and one for the sequences in nonsignal_seqs.txt. Submit these images together with the report. Compare these two LOGOs, can you observe any differences?

Signal

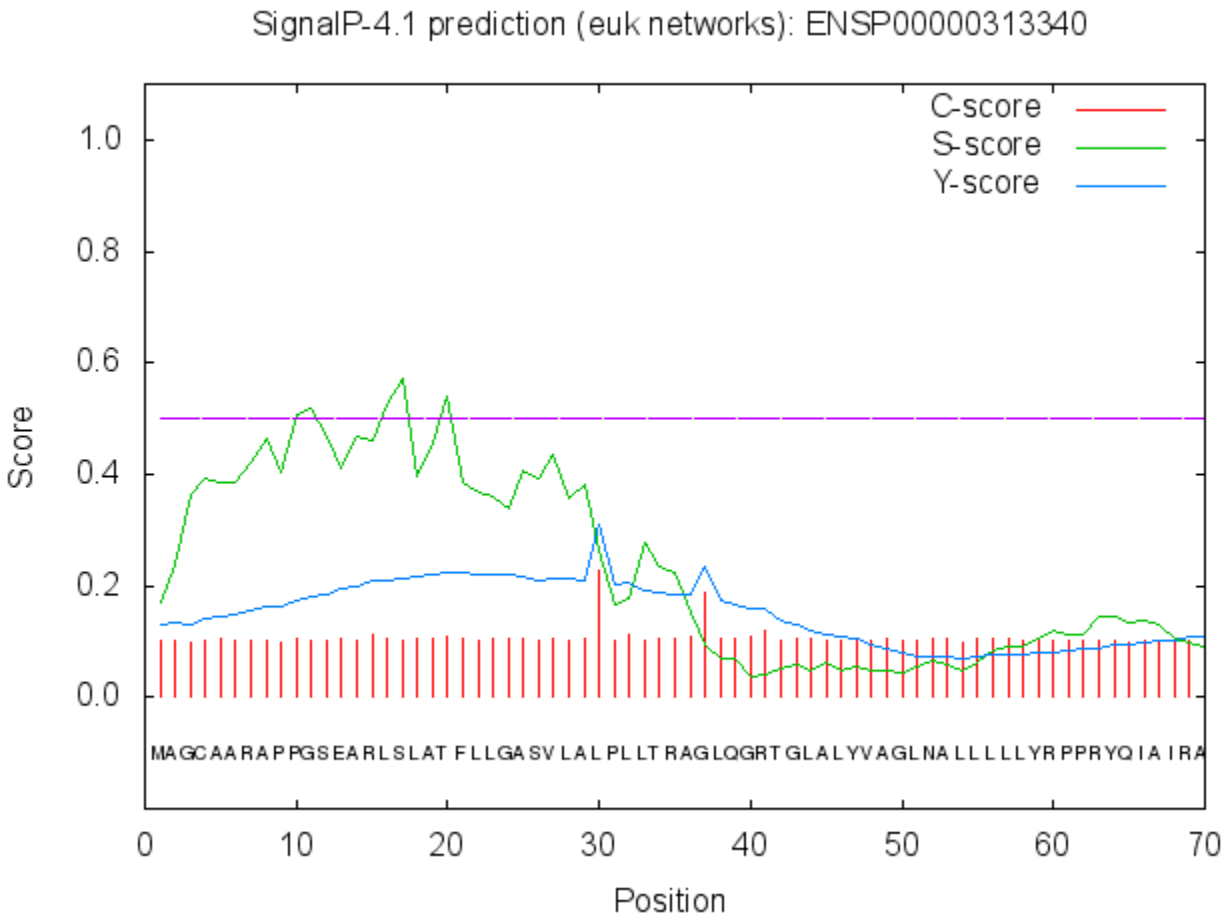


Nonsignal

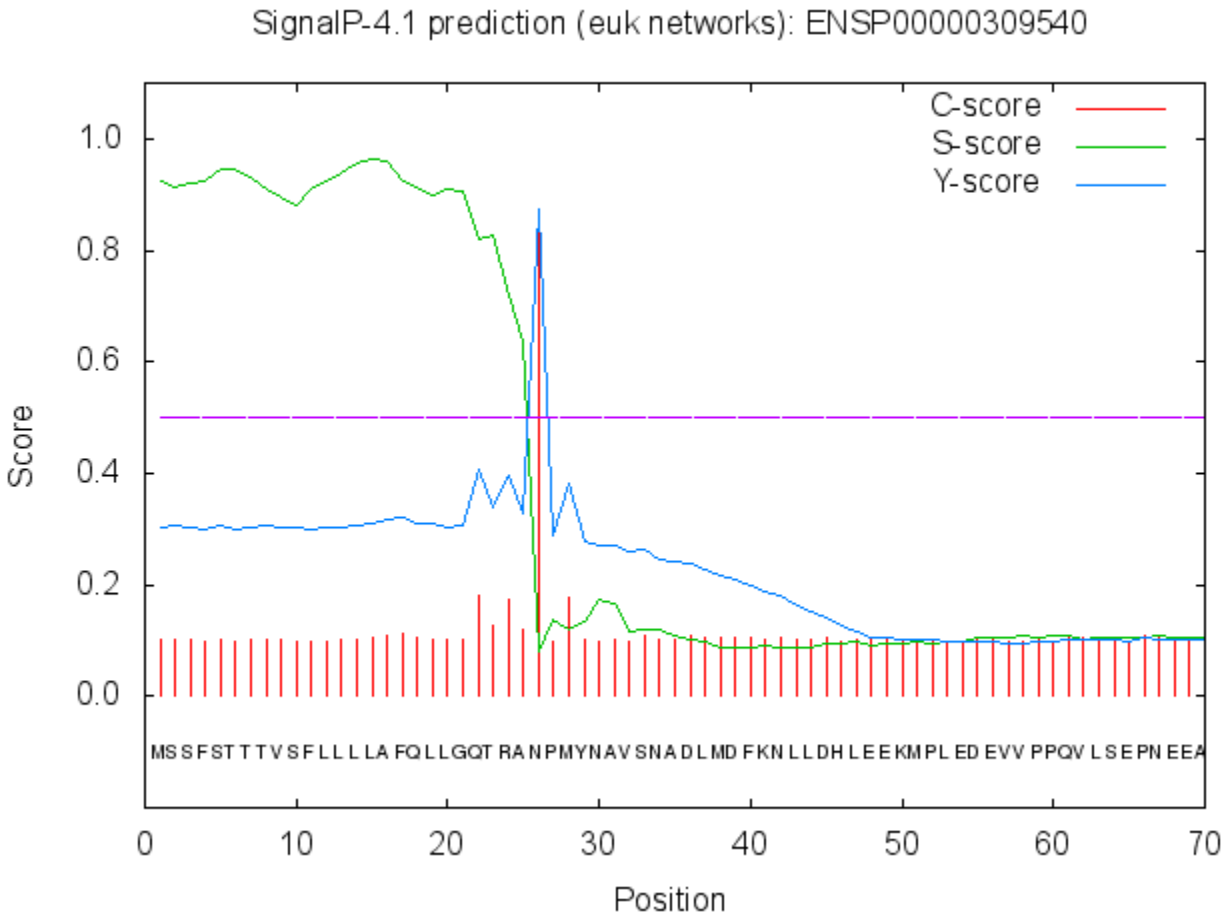


The signal peptide shows a high conversation of leucine at positions 8-19 and a significant higher conservation concerning the Methionin at position 1. In Contrast, the Methionin in the non signal peptide surely is high conserved as well, but not to the same amount

14. There are various methods for predicting signal peptides in protein sequences. These tools have been trained on known signal peptide data and some of them perform really well. The tool SignalP is available online. Use it on the first two human proteins available in the file sequences.txt. Save the result plots of the SignalIP-NN and submit it together with the report.
First



Second



15. What do high and low S-scores indicate in the SignalP-NN result?

A high S-score indicates a higher chance, that a signaling peptide is existent at the specific place.

16. What are the D-scores for the two sequences?

First: 0.351
Second: 0.885

17. The results from SignalP include something called cleavage sites. How do you interpret the term cleavage site in the context of signal peptides?

The cleavage site is the position on which the signal peptide gets cleaved.