

Algorithms in Bioinformatics 2, SoSe2007

Assignment sheet # 6

Christian Rausch

May 21, 2007

1 Calculation of Lagrange Multipliers (2 points)

In the example of the Section “Non-linear SVMs” in the script, the optimal separating hyperplane (OSH) for the three given points is calculated but the details on how the three Lagrange multipliers $\alpha_1 \dots \alpha_3$ were determined are omitted. Please write down the necessary calculations to get from the data points of the example to the Lagrange multipliers.

2 Constrained Optimization (2 points)

The volume and the surface of a cylinder can be calculated as a function of its radius r and its height h . Minimize the surface of a cylinder given its volume $V > 0$ as a constraint. Give your result as a relation between the diameter $d = 2r$ and the height h as a function of the volume V .

3 Soft and Hard Margin Support Vector Machines (1 point)

SVMs that allow for outliers (errors) are frequently called “soft margin SVMs”, as opposed to “hard margin SVMs” which do not allow for errors. Can it make sense to use soft margin SVMs even if the data are linearly separable? Give an example. (Therefore, you may use `svmtoy` (exercise 4) setting c to an extremely high value to obtain a “hard margin SVM”).

4 Support Vector Machines: Toy Examples (3 points)

LIBSVM is a popular SVM implementation. The following exercises will involve the application of LIBSVM on toy examples. Navigate to the LIBSVM web page at <http://www.csie.ntu.edu.tw/~cjlin/libsvm/> and either use the Java Applet in the middle of the page or download the provided zip archive. In the archive you will find a precompiled Windows executable `svmtoy.exe` and the directory `svm-toy` containing subdirectories for `gtk` and `qt`. On a linux machine go to one of these subdirectories, type `make` and you'll get an executable named `svm-toy`.

The toy program allows for the classification of two (or more) classes of data points in 2D. Click on the GUI screen to generate data points and then “change” to change the class. The parameters may be changed as well.

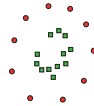
For each of the following tasks, take a screen shot of your result (only the GUI) and comment on it.

4.1 Linear Soft Margin SVMs

Use the `svmtoy` program to generate an example. Try the implementation of linear C-SVMs (kernel `-t0` that means no kernel but “soft margin SVMs”, c = cost factor). Generate an example that is linearly separable and one example that is not. Then, try to solve your non-linearly separable example with a non-linear SVM, i.e. using a kernel. Try different values for cost factors (e.g. 100, 500, 1000, 10000). Please, comment on your results. Take one screenshot of the linearly separable case and one of the other.

4.2 Non-Linear SVMs, Topology 1

Generate something like this:



Try out all three non-linear SVMs (`-t1` through `-t3`). Vary c . Please, comment on your results.

4.3 Non-Linear SVMs, Topology 2

Arrange the points in an s-shaped manner. At some regions, the two classes should intertwine and overlap. Try the RBF and sigmoid kernels. Compare these results with the results obtained with the ν -SVM (`-s0`, see the document “LIBSVM : a library for support vector machines for details”). Try different values for c . Please, comment on your results.

5 Classification of Functional Enzymatic Subtypes with SVMs (2 points)

Have a look at the following article discussed in the lecture:

Rausch, C., Weber, T., Kohlbacher, O., Wohlleben, W., and Huson, D. H. (2005). Specificity prediction of adenylation domains in nonribosomal peptide synthetases (NRPS) using transductive support vector machines (TSVMs). *Nucleic Acids Res*, 33(18):5799-5808.

Describe briefly the strategy that was used to classify homologous enzymes (enzymatic domains) according to their substrate specificity. How did the authors encode the properties of amino acids into numeric features? What special kind of SVMs was employed? Which kernel type(s) was/were most successful in this study?

Assignments due: **Monday, June 4, 10am**