Übung zur Vorlesung

Maschinelles Lernen 2

Sommersemester 2013

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 8

Due **June 10**, 9am local time on ISIS

Weighted Degree Kernels (30 + 70 points)

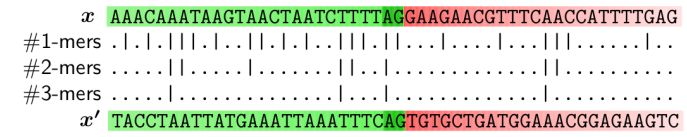
In the lecture, machine learning methods in bioinformatics based on the example *mGene* have been discussed. *mGene* acts as a two-layer approach. The first layer learns biological signals using support vector machines (SVM). The outputs of the SVMs are then used in the second layer as inputs for structural learning, to achieve accurate segmentations into *gene* and *not-gene* (introns, exons, ...).

1. In this exercise, the weighted degree kernel (WDK) shall be used for the detection of splicing locations. For that we use SVM^{light} (`http://svmlight.joachims.org/`) as an implementation of the support vector machine.

   The weighted degree kernel is defined as :

   $$\mathrm{k}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sum_{m=1}^{M} \beta_m \sum_{n=1}^{N-m+1} \mathbf{I}(\boldsymbol{u}_{m,n}(\boldsymbol{x}_i) = \boldsymbol{u}_{m,n}(\boldsymbol{x}_j)). \tag{1}$$

   where $\boldsymbol{u}_{m,n}(\boldsymbol{x})$ is a string of length m which starts at position n in sequence $\boldsymbol{x}$. The symbol $\mathbf{I}(.)$ denotes the indicator function which returns 1 if the input argument is true and 0 otherwise. We use $\beta_m = 2\frac{M-m+1}{M(M+1)}$ for weighting.

   Show that k is a positive definite symmetric kernel (= Mercer Kernel).

2. As SVM^{light} does not support the WDK as a built-in function we will calculate its explicit feature representation

   $$\phi_{w,m,n} = \sqrt{\beta_m}\mathbf{I}(\boldsymbol{u}_{m,n}(\boldsymbol{x}) = w).$$

   for all $w \in \{A, C, G, T\}^m$, $m \in \{1, \ldots, M\}$, $n \in \{1, \ldots |\boldsymbol{x}| - m\}$ which can then be used with a linear kernel.

   a) Complete the function `write_wdk_features`, which writes the WDK features into an SVM^{light} file as described above.

   b) Train SVM^{light} using the splicing data set with parameters $M = \{1, 2, 3\}$ and regularization constant $C = \{0.001, 0.01, 0.1, 1, 10\}$. Measure the prediction accuracy on the training *and* evaluation data set. For each value m create a plot visualizing those prediction accuracies against the value of C and hand them in additionally as tables showing the results of each C versus m. Attach these results to your submitted solution. Describe and interpret your results.

Please ask questions in the ISIS discussion forums for Machine Learning 2:
`https://www.isis.tu-berlin.de/course/view.php?id=8005`