Machine learning 2
Exercise sheet 8

FLEISCHMANN Kay, Matrnr: 352247
ROHRMANN Till, Matrnr: 343756

June 5, 2013

1 Weighted Degree Kernels

The weighted degree kernel is defined as:

$$k(x_i, x_j) = \sum_{m=1}^{M} \beta_m \sum_{n=1}^{N-m+1} \boldsymbol{I}\left(u_{m,n}(x_i) = u_{m,n}(x_j)\right) \tag{1}$$

where $u_m, n(x)$ is a string of length m which starts at position n in sequence x. $\boldsymbol{I}(\cdot)$ is the indicator function. $\beta_m = 2\frac{M-m+1}{M(M+1)}$ is used for weighting.

Show that $k(\cdot,\cdot)$ is a positive semi-definite symmetric kernel. First we show that $k(\cdot,\cdot)$ is symmetric: Given two strings x and y of length N.

$$\begin{aligned} k(x,y) &= \sum_{m=1}^{M} \beta_m \sum_{n=1}^{N-m+1} \boldsymbol{I}\left(u_{m,n}(x) = u_{m,n}(y)\right) \\ &= \sum_{m=1}^{M} \beta_m \sum_{n=1}^{N-m+1} \boldsymbol{I}\left(u_{m,n}(y) = u_{m,n}(x)\right) \\ &= k(y,x) \end{aligned}$$

Now we only have to show that $k(\cdot,\cdot)$ is positive semi-definite, that is to say $\forall v \in \mathbb{R}^L$: $v^T K v \geq 0$ where $(K)_{i,j} = k(x_i, x_j)$ for some arbitrary x_i with $1 \leq i \leq L$.

$$\begin{aligned} v^T K v &= \sum_{i,j=1}^{L} v_i k(x_i, x_j) v_j \\ &= \sum_{i,j=1}^{L} v_i \left(\sum_{m=1}^{M} \beta_m \sum_{n=1}^{N-m+1} \boldsymbol{I}\left(u_{m,n}(x_i) = u_{m,n}(x_j)\right) \right) v_j \end{aligned} \tag{2}$$

Since our alphabet $\mathcal{A}$ of the strings is finite, we can reformulate $\boldsymbol{I}\left(u_{m,n}(x_i) = u_{m,n}(x_j)\right)$:

$$\boldsymbol{I}\left(u_{m,n}(x_i) = u_{m,n}(x_j)\right) = \sum_{w \in \mathcal{A}^m} \boldsymbol{I}\left(u_{m,n}(x_i) = w\right) \cdot \boldsymbol{I}\left(u_{m,n}(x_j) = w\right) \tag{3}$$

Inserting equation (3) into equation (2) we obtain:

$$\begin{aligned} (2) &= \sum_{i,j=1}^{L} v_i \left(\sum_{m=1}^{M} \beta_m \sum_{n=1}^{N-m+1} \left(\sum_{w \in \mathcal{A}^m} \boldsymbol{I}\left(u_{m,n}(x_i) = w\right) \cdot \boldsymbol{I}\left(u_{m,n}(x_j) = w\right) \right) \right) v_j \\ &= \sum_{m=1}^{M} \beta_m \sum_{n=1}^{N-m+1} \sum_{w \in \mathcal{A}^m} \sum_{i,j=1}^{L} v_i \boldsymbol{I}(u_{m,n}(x_i) = w) \cdot \boldsymbol{I}(u_{m,n}(x_j) = w) v_j \\ &= \sum_{m=1}^{M} \beta_m \sum_{n=1}^{N-m+1} \sum_{w \in \mathcal{A}^m} \left(\sum_{i=1}^{L} v_i \boldsymbol{I}(u_{m,n}(x_i) = w) \right) \left(\sum_{j=1}^{L} v_j \boldsymbol{I}(u_{m,n}(x_j) = w) \right) \end{aligned}$$

In equation (4) we used the fact that the double sum can be expressed by the product of two sums (by simply factoring out).

$$\begin{aligned} (4) &= \sum_{m=1}^{M} \underbrace{\beta_m}_{\geq 0} \sum_{n=1}^{N-m+1} \sum_{w \in \mathcal{A}^m} \underbrace{\left(\sum_{i=1}^{L} v_i \boldsymbol{I}(u_{m,n}(x_i) = w) \right)^2}_{\geq 0} \\ &\geq 0 \end{aligned}$$

C \ M	1	2	3
0.001	0.9329	0.9434	0.9375
0.01	0.9402	0.9595	0.9646
0.1	0.9264	0.9462	0.9536
1	0.9159	0.9389	0.9513
10	0.9113	0.9389	0.9513

(a)

C \ M	1	2	3
0.001	0.922	0.937	0.93
0.01	0.948	0.967	0.977
0.1	0.97	0.996	0.999
1	0.994	1	1
10	1	1	1

(b)

Table 1: Prediction accuracy (a) on testing data set and (b) on training data set.

Thus we have just proved that $k(\cdot, \cdot)$ is positive semi-definite and together with the fact that it's symmetric the proposition is proven. Due to the Mercer's theorem it further holds that the kernel function $k(\cdot, \cdot)$ is the inner product of two vectors transformed into an inner product space. □

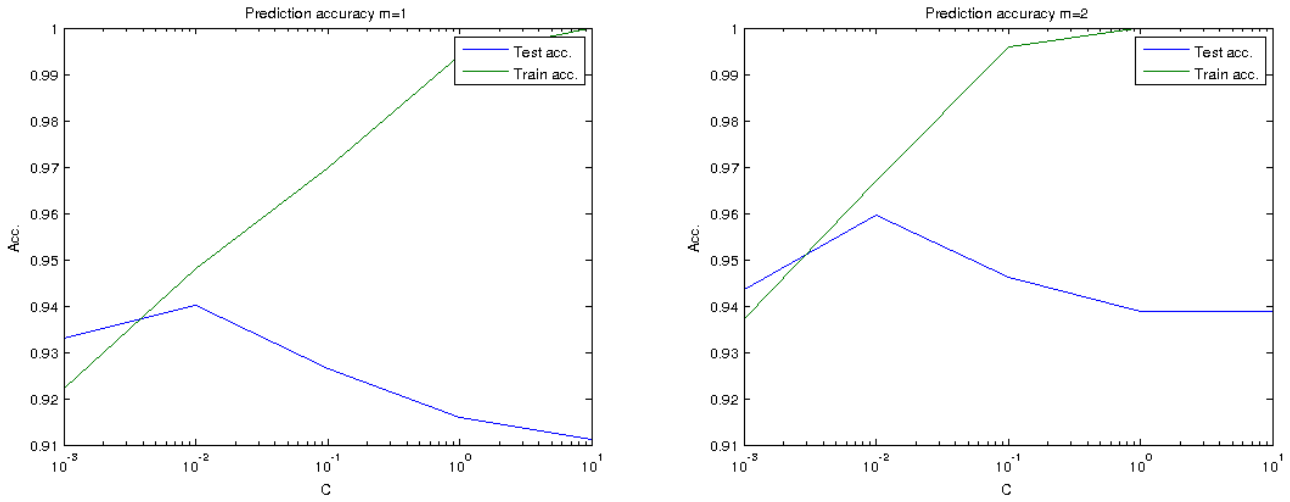
2 SVM light

In this sub task we were supposed to apply SVMlight onto weighted degree kernels and evaluate the performance. We executed SVMlight for the parameters $M = 1, 2, 3$ and the regularization constant $C = 0.001, 0.01, 0.1, 1, 10$.

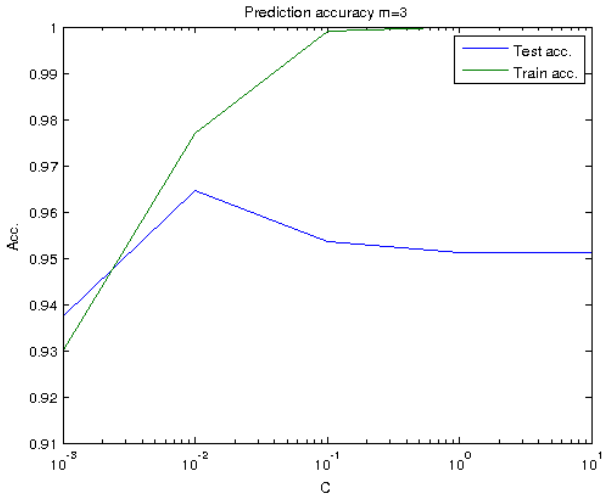
Figure 1 contains the plots of the prediction accuracy on the testing and training data set depending on the regularization parameter C for the different M values. For all values of M we can observe that the prediction accuracy on the training data set converges to 1 with an increasing C. This is due to the fact that the use of slack variables is more highly penalized and thus forcing the optimization procedure to minimize the classification error at the cost of a smaller margin. However, we can also see by observing the prediction accuracy on the testing data set that the accuracy degrades with high C values after reaching a maximum. This indicates that the SVM overfits to the training data set and thus high accuracies on the training data set do not imply necessarily good results on the testing data set. Furthermore, we can observe that a too small C value deteriorates the prediction accuracy on the testing data set as well. The best prediction accuracy is obtained for all M with $C = 0.01$. This shows well that one should not choose the regularization parameter too high because this forces the SVM to overfit to the training data nor should one choose it too small because then the SVM does not learn enough structure from the training data.

Besides, we can see that the prediction accuracies on the testint data set as well as the convergence rate of the prediction accuracies on the training data set are higher with a higher M value. This can be explained by the fact that by increasing the maximum word length M we are giving additional information to the SVM which can be exploited to better classify the data. A comparison of the prediction accuracies on the testing data set of for all values of $M = 1, 2, 3$ is shown in Figure 2. Here we can see again that the higher the value of M is, the higher is the prediction accuracy.

All prediction accuracies used for the plots are shown in Table 1, too.



(a) (b)



(c)

Figure 1: Prediction accuracy on testing and training data set depending on regularization constant C. Figure (a) $M = 1$, figure (b) $M = 2$ and figure (c) $M = 3$.

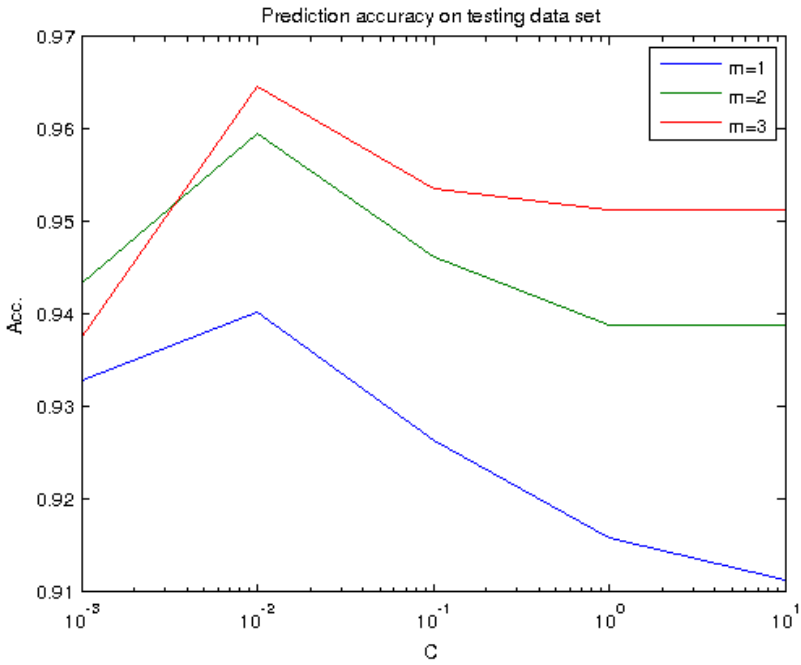


Figure 2: Prediction accuracy on testing data set for all $m = 1, 2, 3$ depending on the regularization constant C.