Datamining and Neural Networks
Exercise Session 3

Moritz Wolter

May 16, 2016

1 Fixed-size LS-SVM

As the number of data points increases working in the dual space becomes harder and harder, because the dimension of the unknown support vectors depends on the number of input data points $\alpha \in \mathcal{R}^N$. The primal problem is better suited for many input data points as the unknown primal weights vector length is determined by the dimension of the input data $\mathbf{w} \in \mathcal{R}^n$. In other words large data set problems should be solved in the primal space, while the dual space should be used if the input data is high dimensional. [1] When SVMs are trained in the primal space, they are called fixed size svms. For the primal space case the kernel trick made it possible to train svms without explicitly knowing which non-linear function mapped to the feature space. When working in the primal space $\varphi(\mathbf{x})$ must be evaluated. This is only simple for linear classifiers, where $\varphi(\mathbf{x}) = \mathbf{x}$. In the non-linear case the Nyström method is required to approximate the nonlinear mapping $\varphi(\mathbf{x})$, the general idea is to choose a a fixed subsampled kernel matrix size M. Typically M is a lot smaller then the true Kernel-matrix size $M \ll N^2$ This smaller kernel matrix is then approximated using a subset of the input data. The computed eigenvalues and eigenvector found from this set are then used as an approximation to the true large version of the matrix. Instead of choosing these support vectors randomly the entropy function,[3]

$$H_r == -\log \int p(\mathbf{x})^2 d\mathbf{x} \tag{1}$$

$$\int p(\mathbf{x})^2 d\mathbf{x} = \frac{1}{N^2} \mathbf{1}^T \Omega \mathbf{1} \tag{2}$$

is used. Starting from a random fixed size pool of support vectors a selected vector is replace with a value from the training set. If the entropy increases the datum is kept in the support vector set. If the entropy function does not increase the new value is rejected and the old one is kept in the set. This procedure is repeated until the entropy function does int increase sufficiently anymore or a maximum number of iteration is reached. The reduced kernel matrix can be determined from the fixed set. After estimating its eigenfunction $\mathbf{w}$ and b are determined. Figure 1 shows the entropy function over hundred iterations of a ten vector subset from a normally distributed data-set. Given this optimized subset the

[1] Support Vector Machines: Methods and Applications, Suykens et al., page 174

[2] Support Vector Machines: Methods and Applications, Suykens et al., page 175

[3] Support Vector Machines: Methods and Applications, Suykens et al., page 181

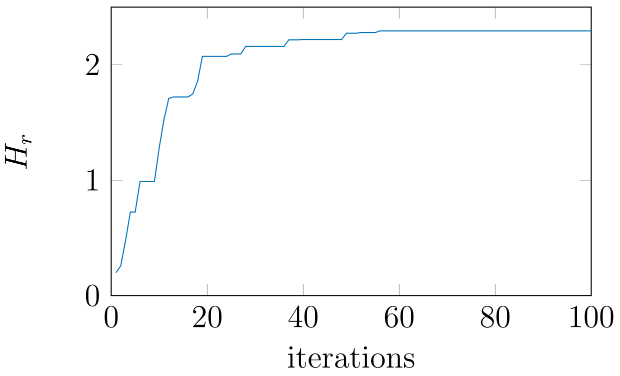

Figure 1: The value of the entropy function for an optimization process of a subset of 10 values drawn from the normal distribution $\mathcal{N}(0, 2^2)$

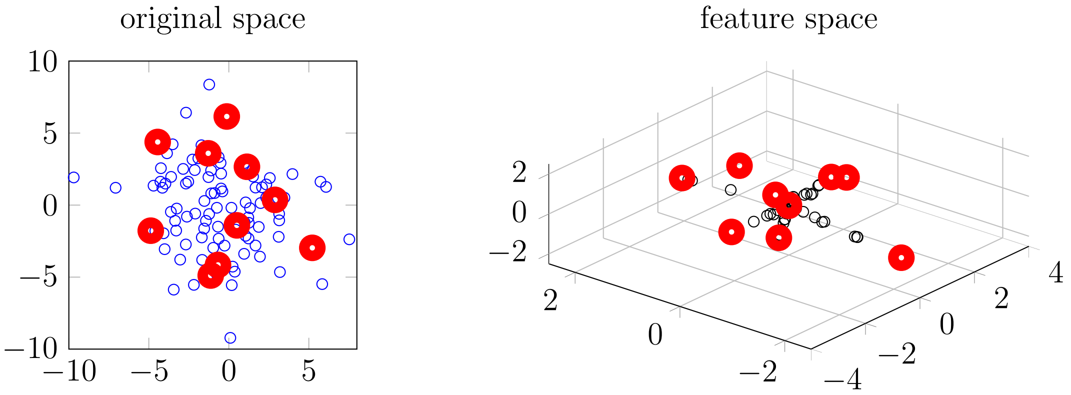

Figure 2: Feature space reconstruction

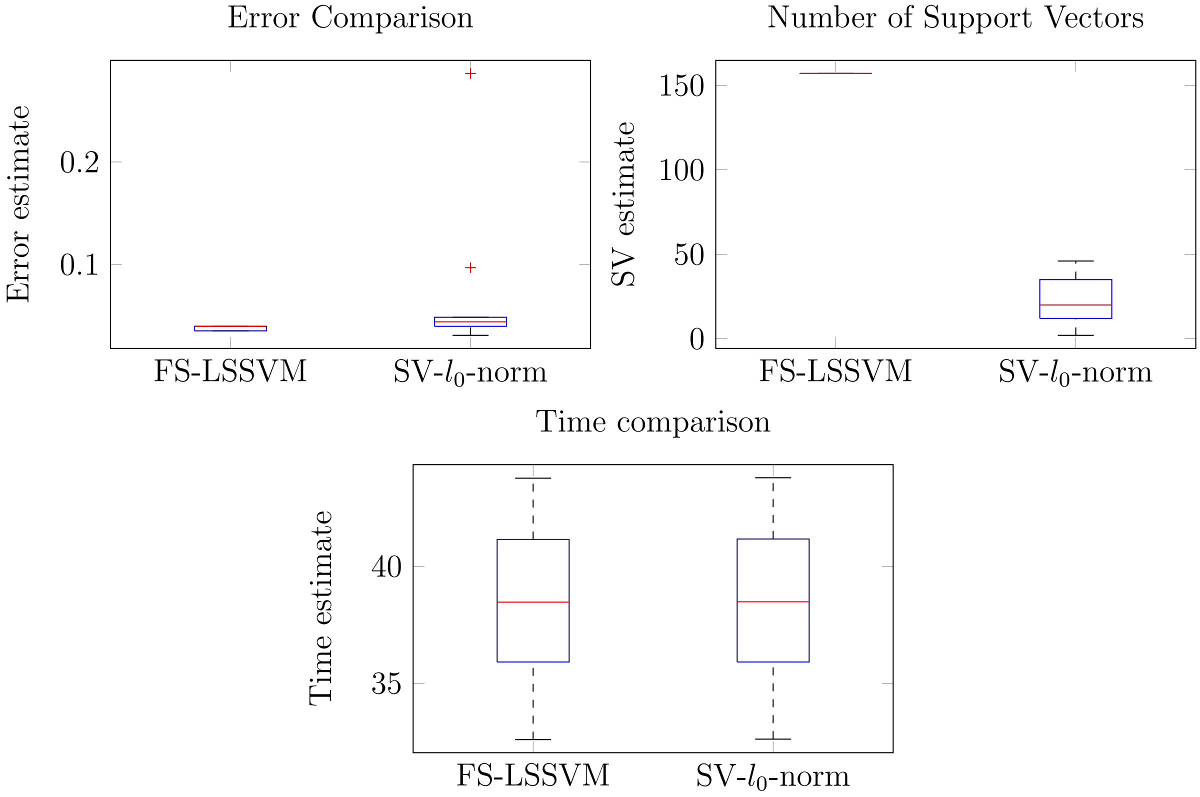

Figure 3: Classical fixed size svm and l_0 reduced version comparison for classification on the Wisconsin breast cancer dataset.

nonlinear mapping can be approximated as shown in figure 2 using the Nyström method, if alongside the selected inputs a kernel function and its parameters are chosen.

1.1 Sparsity and the l_0-norm

Sparsity is a desirable property of trained support vector machines. In the dual space an ls-svm classifier is sparse if many support vectors are zero. [4] When evaluating the classifier only the non-zero vectors have to be taken into account, making the computation more efficient. For fixed size ls-svms the primal problem is considered, the notion of sparsity translates into a smaller representation of the basis given by:

$$\mathbf{w} = \sum_{N}^{k=1} \alpha_k \varphi(\mathbf{x}_k) \tag{3}$$

A good method of determining suitable subset of the input space $\{\varphi(\mathbf{x}_k)\}_{k=1}^{N}$, could for example be the entropy selection method discussed earlier. The determined subset could then serve as a way to approximate a sparse $\mathbf{w}$. This goal is formulated using the l_0 norm

$$\min_{\mathbf{w}} \|\mathbf{w}\|_0 = \|w_1\|^0 + \|w_2\|^0 + \cdots + \|w_N\|^0 \tag{4}$$

[4] Support Vector Machines: Methods and Applications, Suykens et al., page 33

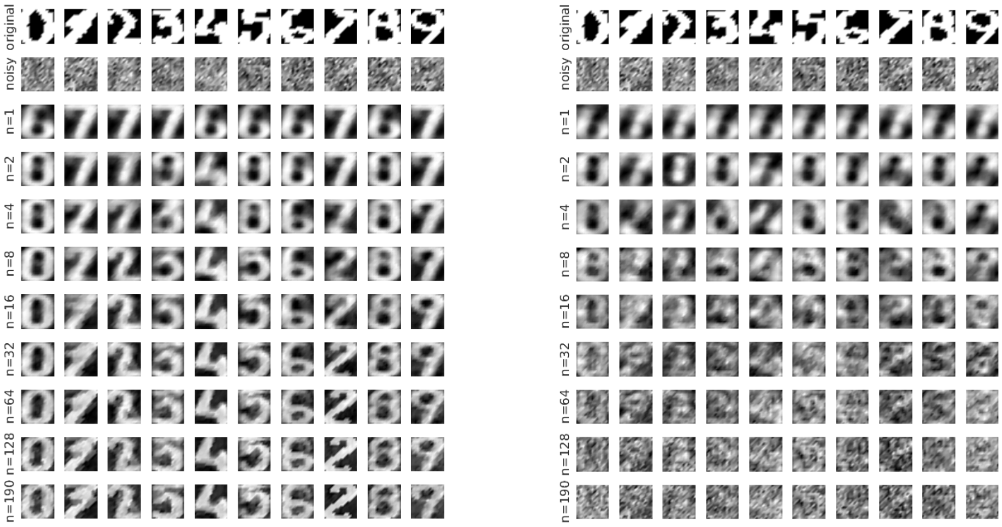

Figure 4: Kernel based and linear pca de-noising of very noise data.

which is equivalent to minimizing the count of non-zero elements in the vector.[5] A first experiment using this method of producing sparsity is use for classification of the comparatively small Wisconsin data set. Results from running `fslssvm_script` are shown in figure 3. The error comparison shows a slightly elevated median for the sparse version, probably due to the random nature of the input set reduction process bad outliers in terms of error exist. The figure also reveals, that the l_0 norm minimization process significatly reduced the number of support vectors, while it does not lead to a significant increase in training time.

2 Kernel and linear PCA-de-noising on high noise data

In this section an extreme noise example is considered, where the human eye has trouble identifying the characters correctly. Figure 4 shows the input data as well as the performance of a low sigma kernel-pca in comparison to a linear one. The trade off that came with choosing the kernel width σ^2 was observed earlier. A too large σ^2 led to correct predictions but little noise reduction. Smaller width sometimes ended up clear but incorrect letter representations. In the case shown in figure 4 a small σ^2 had to be chosen in order to deal with the very noisy input. Given the low quality of the input the kernel-PCA does an incredibly good job at de-noising. It does confuse 3 with 5 and 6 with 2 however. In the linear case this does not happen but it does not come close in terms of noise reduction.

[5] David Wipf and Bhaskar Rao, l_0-norm Minimization for Basis Selection

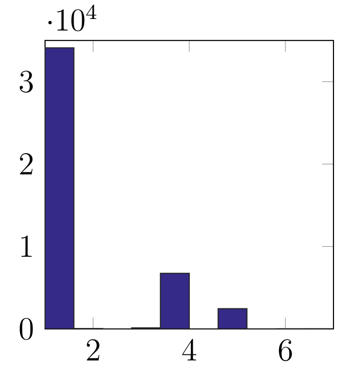

Figure 5: Histogram of the statlog dataset's training data class distribution.

3 Shuttle dataset-analysis

The statlog/shuttle dataset, contains 58000 ten dimensional row data vectors. With the last entry in each row indicating the class to which the data point belongs. Traditionally the first 43500 data rows are used for training and the last 14500 data points are retained for testing. Figure 5 shows the shuttle-dataset's class distribution. About eighty percent of the data belongs to the first class. A modified version of the `fslssvm_script` has been used with the shuttle data set as an input. The aim here is again classification, but its must be noted that the shuttle dataset is much larger than the Wisconsin-cancer set, it has 58000 data points while the cancer set only contains 682 data points. The results of the machine architecture comparison are shown in figure 6. On this larger data set the difference in terms of the error estimate become negligible. Quite a significance difference in terms of sparsity persist however, while the training time does not increase significantly.

4 California dataset-analysis

The California housing regression problem.

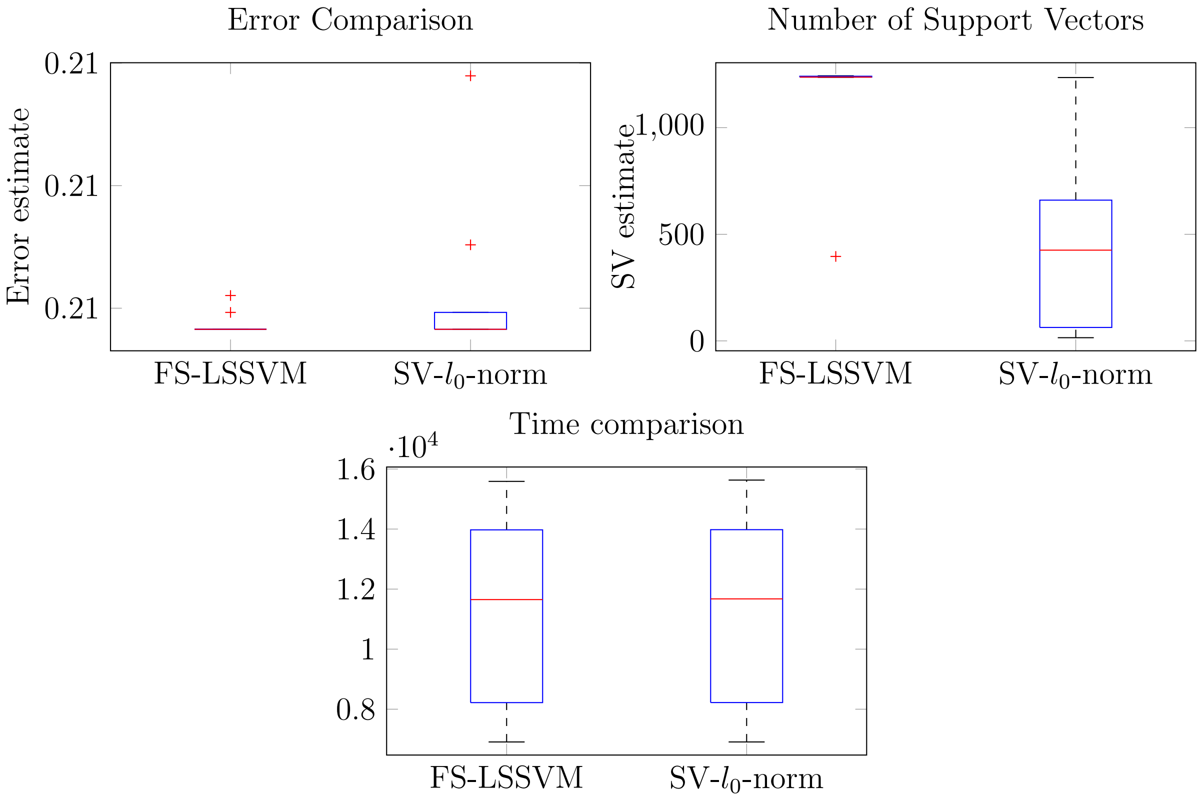

Figure 6: Fixed size svm and l_0 svm comparison for classification on the nasa shuttle dataset.