Course: Machine learning

By: Pavlo Mozharovskyi

Tutorial to Lecture 4: Support vector machines

Task 1: Comparative study of different classification algorithms.

Conduct a simulation study in spirit of Task 2 of Tutorial 3, where you include support vector machine (with an arbitrary choice of tuning parameters) and at least four other classifiers. For each of at least three distributional settings from Table 1 (Appendix from Tutorial 2), plot a diagram with boxplots reflecting classification performance of the chosen classifiers averaged over 100 random draws.

Task 2: Classification of written digits.

1. Read about the MNIST data set at http://yann.lecun.com/exdb/mnist. Download it and read it in `R` in a format acceptable for supervised learning.

2. Train support vector machine on this data set (use, *e.g.*, `R`-function `svm` from `R`-package `e1071`). For speed reasons, skip tuning, after re-scaling data to $[-1, 1]^d$ use Gaussian kernel $K(\boldsymbol{u}, \boldsymbol{v}) = e^{-\gamma\|\boldsymbol{u}-\boldsymbol{v}\|^2}$ with $\gamma = 0.01$ and take value 1 for the box constraint. Further, again for time reasons, restrict to $2 < Q < 10$ digits, and during the training phase use a sub-sample consisting approximately of Q thousand observations.

3. Test the performance of the trained support vector machine on the complete test sample (restricted to Q classes) containing approximately Q thousand observations. Print the confusion $Q \times Q$ table, where each row stands for a digit (of the test sample) and represents the percentages of times it has been assigned to each of the available digits (in columns).