Assignment 7

Machine Learning: Algorithms and Theory

Prof. Ulrike von Luxburg / Diego Fioravanti / Moritz Haas
Tobias Frangen / Siavash Haghiri

Summer term 2018—due **June 12**

Exercise 1 (Reproducing Kernel Hilbert Space, 1+3+3 points)
In this exercise we try to get a better understanding of reproducing kernel Hilbert spaces (RKHS thereafter), namely by looking at the RKHS corresponding to simple kernels. Look at the slides corresponding to RKHS from the lecture to solve this problem.

(a) Define the *linear kernel* $k_\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as $k_\ell(x, y) = \langle x, y \rangle$ for any $x, y \in \mathbb{R}^d$. Use the definition to show that k_ℓ is a positive semi-definite kernel on $\mathbb{R}^d$.

(b) Show that the RKHS associated to k_ℓ consists of all functions of the form $f(x) = \langle \alpha, x \rangle_{\mathbb{R}^d}$ where $\alpha \in \mathbb{R}^d$. We denote this function space by $\mathcal{H}$. Notice that $\mathcal{H}$ is finite dimensional. What are the corresponding inner product and norm defined on $\mathcal{H}$?

(c) We now turn to the *polynomial kernel* $k_P : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ defined as $k_P(x, y) = (\langle x, y \rangle_{\mathbb{R}^2} + 1)^2$ for any $x, y \in \mathbb{R}^2$. Show that k_P is a positive semi-definite kernel on $\mathbb{R}^2$. Show that the associated RKHS consists of all functions of the form

$$f(x) = \alpha_1 x_1^2 + \alpha_2 x_2^2 + \alpha_3 x_1 x_2 + \alpha_4 x_1 + \alpha_5 x_2 + \alpha_6,$$

with $x \in \mathbb{R}^2$ and $\alpha \in \mathbb{R}^6$. Notice that, again, $\mathcal{H}$ is finite-dimensional. What are the inner product and norm on $\mathcal{H}$? *Hint:* Use a feature map.

Exercise 2 (Feature space, positive semi-definite matrices, 2+2 points)

(a) Find a feature space and a feature map such that the data set represented in Figure 1 becomes linearly separable in the feature space. The feature space representation should be a function of the coordinates (x, y).

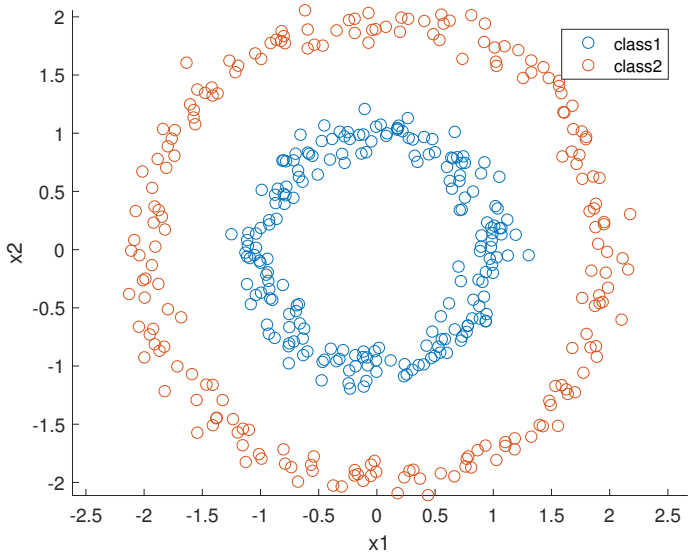

Figure 1: The two classes in this dataset can not be linearly separated in $\mathbb{R}^2$.

(b) Consider a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ defined as $k(x, y) = \langle \Phi(x), \Phi(y) \rangle_{\mathbb{R}^d}$, where $\Phi : \mathcal{X} \to \mathbb{R}^d$ is an arbitrary function. Given a finite set of points $x_1, x_2, \ldots, x_n \in \mathcal{X}$ we define the *Gram matrix* $K \in \mathbb{R}^{n \times n}$ as $K_{ij} = k(x_i, x_j)$. In literature the Gram matrix some times goes under the name of *kernel matrix* . Prove that K is a symmetric matrix and prove that K is in fact a positive semi-definite matrix. *Hint*: A symmetric matrix is positive semi-definite if, and only if, it can be written as $S = XX^\top$ for some matrix X

Exercise 3 (Using kernel SVM, 1+2+1+2+1+2)

(a) In `data` and `labels` you will find the data and labels that we will use for this exercise. Split them into a training set (2/3 of the points) and a testing set (1/3 of the points).

(b) Perform kernel SVM with a Gaussian kernel choosing regularization parameters C and γ via cross-validation. Use `sklearn.svm` to build the SVM model. Then use `sklearn.model_selection .GridSearchCV` to perform the cross-validation with $C \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ and $\gamma \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000, \text{'auto'}\}$. Which values of C and γ give the best results?

(c) Compute the risk on the test set with respect to the 0–1-loss. Then compute the risk with an unbalanced loss define as follows: incur a loss of 0.2 if you wrongly predict 0 whereas the true label is 1 and a loss of 1.8 if you wrongly predict 1 whereas the true label is 0.

(d) For $C \in \{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$, compute the empirical risks with respect to the balanced and unbalanced loss as in task (c) (use the standard value for γ, i.e. don't set γ at all). For both losses plot the risks (y-axis) against C (x-axis) on a log scale. Which C do you prefer for which loss function? Now do the same thing with γ set at the value that you found in (b). does the best C change? What else do you notice?

(e) To further understand the behavior of different kernels we go back to a toy data set where $X_i \in \mathbb{R}^2$ so that `X` is a $n \times 2$ matrix representing X_i in the i-th row. Y_i is the class label in $\{0, 1\}$ stored in `Y`. Apply kernel SVM to the data with (i) a Gaussian kernel, (ii) a linear kernel, (iii) a polynomial kernel of order 2, and (iv) a polynomial kernel of order 3. Choose C and γ via cross validation.

(f) For each kernel in (e) plot the data and the decision boundary. As in sheet 05, use as much code as you like from `http://scikit-learn.org/stable/auto_examples/svm/plot_iris.html` and read the explanation in the beginning.