Machine Learning

Exercise Sheet 10

Winter Term 2023/2024
Prof. Dr. Niels Landwehr
Dr. Ujjwal

Available: 25.01.2024
Hand in until: 01.02.2024 11:59am
Exercise sessions: 05.02.2024/07.02.2024

Task 1 – Distance of a point from a hyperplane 5 points

Show that the distance of a point $\mathbf{x}_o$ from a hyperplane given by the equation $\mathbf{x}^T\ \boldsymbol{\theta}+b=0$ is equal to $\frac{\mathbf{x}_o^T\ \boldsymbol{\theta}+b}{||\boldsymbol{\theta}||}$. You may use the following formula for vector projection: The projection of a vector $\mathbf{a}$ onto a vector $\mathbf{b}$ is given by $\mathbf{a}_1 = a_1 \frac{\mathbf{b}}{||\mathbf{b}||}$, where $a_1 = \mathbf{a}^\mathsf{T} \frac{\mathbf{b}}{||\mathbf{b}||}$.

Task 2 – Hard margin SVM: primal and dual [5+10+5 points]

Let $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), (\mathbf{x}_3, y_3)\}$ denote a training data set, with $\mathbf{x}_1 = (1,1)^\mathsf{T}$, $\mathbf{x}_2 = (1,3)^\mathsf{T}$, $\mathbf{x}_3 = (3,2)^\mathsf{T}$ and $y_1 = 1$, $y_2 = 1$, $y_3 = -1$. In this task, we study hard-margin support vector machines, that is, linear models of the form $f_{\boldsymbol{\theta}} : \mathbb{R}^2 \to \mathbb{R}$ with $f_{\boldsymbol{\theta}}(\mathbf{x}) = \mathbf{x}^\mathsf{T}\boldsymbol{\theta} + b$.

a) Manually plot the data set in the 2D-space with axes x_1, x_2. Also plot the maximum margin decision boundary. Based on the geometric visualization, write down the maximum margin linear model with parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)^\mathsf{T} \in \mathbb{R}^2$ and $b \in \mathbb{R}$. As discussed in the lecture (Slide 11), normalize your model such that $||\boldsymbol{\theta}|| = 1/C$.

b) Prove that the model you found is indeed the solution of the primal hard-margin SVM problem given on Slide 12 of the lecture. Hint: first show that your model fulfills the linear inequality constraints. Then derive from the linear inequality constraints that any other solution of the linear inequality constraints would have a larger norm $||\boldsymbol{\theta}||$.

c) Derive a dual version of the same model, that is, find a vector $\boldsymbol{\alpha} \in \mathbb{R}^3$ of dual variables that satisfies the optimization problem given on Slide 14 of the lecture. Hint: exploit that you already know the primal solution.

Task 3 – Kernels in SVMs 5 points

Prove or disprove the statement – " $K(\mathbf{x}, \mathbf{z}) = -\frac{<\mathbf{x},\mathbf{z}>}{||\mathbf{x}||_2\ ||\mathbf{z}||_2}$ is a valid kernel function". Here, $\mathbf{x}, \mathbf{z} \in \mathbb{R}^M$, $< ., . >$ denotes the inner product and $||.||_2$ denotes the $L2$ norm.

Task 4 – Programming 4+2+4 points

You are given an IPython notebook *svm_programming_task.ipynb*. Please read the following description carefully and perform the tasks that follow:

The notebook generates a set of random data points spanning 2 classes in 2D space using *make_blobs()* function in scikit-learn. Please refer to the documentation for further details.

This data is then classified using a SVM for which a *fit_data()* function is provided in the notebook. The noteboook also contains functions to plot the decision boundary of the learnt SVM. Please answer the following questions

a) Vary the value of the parameter C in the notebook and for each value study the number and class of support vectors. Complete the function *plot_decision_boundary()* function to plot the support vectors as well as the decision boundary of the SVM. Specify your observations about the number of support vectors as you change the value of C.

Hint: The decision boundary C an be plotted by predicting all the points in the 2D space and plotting different classes in different colors.

b) There is a normalize option in the *fit_data()* function which normalizes the features by whitening them before training. Experiment with it and mention your observations about how this affects the accuracy on the test data. Give plausible reasons for your observations.
c) Change the *kernel* parameter in the *svm.SVC()* function and study the impact on the accuracy. Give reasons for your observations