Machine Learning I: Foundations
Exercise Sheet 7

Prof. Marius Kloft TA:Billy Joe Franks

05.06.2020

Deadline: 16.06.2020

1) (MANDATORY) 10 Points

Interestingly the linear hard-margin SVM, given by

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 \qquad (1)$$
$$\text{s.t. } 1 - y_i(\mathbf{w}^T\mathbf{x}_i + b) \leq 0, \ \forall i \in \{1,\ldots,n\},$$

requires only two (non-equal) training points (with opposite labels) to find a separating hyperplane. Let $X := \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $Y := \{y_1, \ldots, y_n\}$, with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$, be a dataset. Let $\mathbf{w}^*$ and b^* be the optimal solution to the above optimization problem (1) on X, Y. You may assume $w_1 \neq 0$.

a) Find a minimal dataset (X', Y') with $|X'| = |Y'| = 2$ (consisting only of two data points) with the same hard-margin SVM solution (Eq. (1)) as for the dataset (X, Y) , that is, $\mathbf{w}^*$ and b^*.

b) Prove that, for your choice of X' and Y' in a), $\mathbf{w}^*$ and b^* are optimal solutions of (1).

c) Why would it be advantageous to use (X', Y') instead of X, Y during optimization, assuming we had access to both and knew they are equivalent? (Answer this question with at most 5 sentences.)

d) Assume we train the hard-margin SVM with only two (arbitrary) training points (not the optimal data points as above). Consider $d \to \infty$. What can you state regarding overfitting and underfitting here? Explain your answer. (Answer this question with at most 5 sentences.)

2) Consider the kernel ridge regression optimization problem (Lecture 8, Slide 39). Let $\alpha^* \in \mathbb{R}^d$ be the vector that minimizes the loss function. Show that:

$$\alpha^* = \left(K + \frac{1}{2C}\mathrm{I}_{n\times n}\right)^{-1} y$$

3) In the lecture we found a closed form solution for linear ridge regression and we incorporated b afterwards by simply changing the dataset slightly. This however means that b is regularized during optimization. What would happen if we introduce b in a different way? Consider linear ridge regression with offset

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + C\left\|\mathbf{y} - (X^T\mathbf{w} + \hat{\mathbf{b}})\right\|^2 \tag{2}$$

where $\forall i : \hat{b}_i = b$. $\hat{\mathbf{b}}$ simply copies b into each component. Alternatively the norm could be written as a sum incorporating only b, as follows

$$\min_{\mathbf{w}\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i \left(y_i - (\mathbf{x}_i^T\mathbf{w} + b)\right)^2 \tag{3}$$

(2) and (3) have the same closed-form solution. Find this solution. Thereby choose the version from the above two that you prefer ((2) or (3)).

4) Solve programming task 7.