Assignment 5

Statistical Machine Learning

Sebastian Bordt / David Künstle / Martina Contisciani / Nicolò Ruggeri /
Rabanus Derr / Prof. Ulrike von Luxburg

Summer term 2020 — due on **May 28th at 14:00**

Exercise 1 (SVM by hand, $2+3+1+1=7$ points)
Consider a dataset containing the following three data points in $\mathbb{R}^2$: $x_1 = (0,0)^T$, $x_2 = (1,2)^T$, $x_3 = (-1,2)^T$ and their corresponding labels $y_1 = -1$, $y_2 = 1$, $y_3 = 1$.

(a) Write down the primal problem of the hard margin SVM on this dataset. Compute the Lagrangian $L(w, b, \alpha)$.

(b) What is the dimension of w? What is the dimension of α? Compute the saddle point according the saddle point conditions.

(c) Explain what a support vector is and determine the support vectors of this problem.

(d) Apply `sklearn.svm.SVC`, with the correct parameters, on this dataset to compute the SVM model. Check its attributes `coef_`, that gives w, `dual_coef_`, that gives the products $y_i\alpha_i$, and `intercept_`, that gives b. Are the values equal to your analytic solution? Why?

Exercise 2 (Primal hard margin SVM problem, 3 points) Given training data $(X_i, Y_i)_{i=1,\dots,n} \in (\mathbb{R}^d \times \{-1,+1\})^n$, the primal hard margin SVM problem is given as

$$\begin{aligned} &\min_{w\in\mathbb{R}^d, b\in\mathbb{R}} \frac{1}{2}\|w\|^2 \\ &\text{subject to } Y_i\left(w^T X_i + b\right) \geq 1, \forall\, i = 1,\dots,n. \end{aligned} \tag{1}$$

Recall the meaning of a hyperplane in canonical representation. Show that any solution of (1) gives rise to a hyperplane in canonical representation.

Exercise 3 (Linear SVM in action, $1+2+3+1+3=10$ points)

For the diagnosis of cancer doctors use medical images as an important indicator. The doctors usually need to go through an intense training to interpret the images correctly. This is a situation in which we can support them by machine learning. Let us build a first prototype for breast cancer detection.
Note: Do not `import` additional library functionality.

(a) First, you should inspect the provided dataset[1] of your prototype. Describe in your own words, what is represented by the point features, and class labels.

(b) Let's try linear SVMs. Split the dataset into training and test parts (70%/30%). Fit `LinearSVC` model with default hyperparameter on the training part. During fitting, the dual optimization problem is solved. Analyse how the test error varies according to different random initialization (`random_state` parameter of `LinearSVC`).

(c) You are not satisfied with this classification error. Implement an optimization of the hyperparameter, *using only the training dataset*. State the best hyperparameter configuration and an estimate of the corresponding true risk based on the 0-1-loss *using only the training dataset*. Make sure, that the hyperparameter optimization takes less than 30 seconds wall clock time on a normal computer.

[1] `https://scikit-learn.org/stable/datasets/index.html#breast-cancer-dataset`

(d) Maybe a logistic regression model is a better choice? Compare the 0-1-loss on the test dataset between a linear SVM and a logistic regression model. Train both models on the full train dataset and use the best C found in the previous task for the SVM. *Note:* For a fair comparison, one has to optimize the hyperparameter of the logistic regression model, too. We ignore this here and use the default.

(e) Now we finished the prototype, but hang on. State three concerns about your approach. The concerns should include one ethical and one technical or statistical problem.