Assignment 5

Statistical Machine Learning

Moritz Haas / Prof. Ulrike von Luxburg

Summer term 2022 — due on **May 23rd at 12:00**

Exercise 1 (Lagrange multipliers, 2 points)
Use the method of Lagrange multipliers to solve the following problem

$$\begin{aligned} &\max x + y \\ &\text{subject to } x^2 + 2y^2 \leq 5 \end{aligned}$$

Is the constraint active? Do you have a geometrical explanation of why the constraint should be active or inactive?

Exercise 2 (Linear and quadratic programs, 3+1+1=5 points)

(a) The simplest class of convex optimization problems with constraints are linear programs (LP)

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} c^T x \\ &\text{subject to } Ax \leq b \\ &\qquad\qquad x \geq 0, \end{aligned} \tag{2}$$

for $c \in \mathbb{R}^n$, $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$. Show that the dual of a linear program is again a linear program.

Hint: Derive the dual problem. When is it feasible, and what is the attained value?

(b) Another class of convex optimization problems are quadratic programs (QP)

$$\begin{aligned} &\min_{x \in \mathbb{R}^n} \frac{1}{2} x^T Q x + c^T x \\ &\text{subject to } Ax \leq b \end{aligned} \tag{3}$$

Here c, b and A are as in (2), and $Q \in \mathbb{R}^{n \times n}$ is a symmetric positive semi-definite matrix. Why do we require that Q is positive semi-definite? Is the problem in Exercise 1 a quadratic program?

(c) The dual of a quadratic program is again a quadratic program, and strong duality holds for quadratic programs. Briefly explain what this means.

Exercise 3 (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 4 (Linear SVM in action, $1+3+2+3=9$ 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) After centering and normalizing your data, we have split the dataset into training and test parts (70%/30%). Then we have fit the `LinearSVC` model with default hyperparameters on the training part. During fitting, the dual optimization problem is solved. You are not satisfied with this classification error.

Implement an optimization of the hyperparameter C, *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.

(c) Maybe a logistic regression model is a better choice? Repeat all steps in (b) for `LogisticRegression`.

Compare the estimates of the true 0-1-risk *using only the training dataset* between a linear SVM and a logistic regression model. Now train both models on the full train dataset and use the best C found in the previous task for both models. Compute the test error of both models. What do you observe and why?

(d) 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.

Exercise 5 (Feedback Poll, 2 bonus points)
Every week there will be a feedback survey on Ilias about the lecture, tutorials and exercise sheets. If you participate, it helps us in understanding what we can improve.

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