Assignment 4

Statistical Machine Learning

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

Summer term 2020 — due on **May 21th at 14:00**

Exercise 1 (Lagrange multipliers, 2+2=4 points)

(a) Use the method of Lagrange multipliers to solve the following problem

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

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

(b) Use the method of Lagrange multipliers to solve the following problem

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

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 any of the problems 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 (Duality gap, 1+1+1=3 points)
Consider the following optimization problem:

$$\underset{x \in \mathbb{R}}{\text{minimize}} f(x) = x^4 - 10x^2 + x$$
$$\text{subject to } g(x) = x^2 - 2x - 3 \leq 0$$

(a) Write down the feasible set of this optimization problem and solve its primal. You do not have to solve the primal by hand and you can use a solver of your choice, for example WolframAlpha.

(b) Compute the Lagrangian and the dual problem. Derive a lower bound on the primal using the optimal solution of the dual. You can use the fact that the dual function is optimal when the lagrange multiplier equals 0.5. You do not have to solve the dual by hand and you can use a solver of your choice, for example WolframAlpha.

(c) Does strong duality hold in this case? Explain why and quantify the duality gap if any.

Exercise 4 (Logistic regression, 2+2+2+2=8 points)

In this exercise we are going to use the candy dataset from kaggle. It is available from our website as `candy-data.csv`. With this dataset we are going to do two things. First we will apply logistic regression to find out if a candy contains chocolate or not. Second, we will study the effects that regularization has on model training and predictions.

We are also going to use the popular machine learning framework `scikit-learn`. Instead of implementing the machine learning algorithms ourselves, as we did for kNN and ridge regression in the last two assignments, we are now going to use the implementation of logistic regression that comes with `scikit-learn`. We will continue to use this framework in future assignments.

Note: You might get a `FutureWarning` when using `LogisticRegression`. Just ignore it.

(a) In the variable `candies` you can find the dataset that we loaded for you. Now, divide the dataset in three parts: `names` that contains the names of the candies, `ys` that contains if a candy has chocolate or not and `xs` everything else. Now, divide `xs` and `ys` in `xs_train`, `ys_train` and `xs_test`, `ys_test`. The first $\frac{2}{3}$ (rounded down) of the dataset goes into train, the rest into test.

(b) Use scikit-learn `LogisticRegression` with the default parameters to predict if a candy in `xs_test` contains chocolate or not. Compute the accuracy for your prediction and store it in a variable `acc`.

(c) Use scikit-learn `GridSearchCV` to do a 10-fold cross validation for $C \in \{0.01, 0.1, 1, 10, 100, 1000\}$. Plot on the same graph the cross validation error and the test error. Please use a log scale when plotting C.

Hint: The key that you are interested in is `mean_test_score`.

(d) Does cross validation choose the best C? Comment in terms of over/underfitting. What happens if you re-run the code for this exercise? Does the best C change?

Hint: `LogisticRegression` has a method called `score(X, Y)` that computes the accuracy between the predictions and the true `Y`. You cannot use it for b) but you can use it here.