Computer Lab: Natural Language Processing
SD-TSIA 211

Tamim El Ahmad, Olivier Fercoq, Joël Garde, Iyad Walwil, Bruno Costacèque

6 January 2023

1 Submission and grading information

You can do the computer lab alone or in pairs. Please write a report and post it on `e-campus`. You can do it as a jupyter notebook or a pdf file.

Then, each of you will have to evaluate a couple of other students' reports and give comments.

Only the fact that you produce a report and evaluate your peers count in the final grade,so do not worry if you do not finish everything.
- 1 point for being present on the day of the lab
- 1 point for submitting a report
- 1 point for commenting 2 reports

2 Database

On `e-campus`, download the files `tfidf_matrix_97MB.npz`, `feature_names_97MB.npy` and `train_labels.npy`. These files were generated from `https://granddebat.fr/pages/donnees-ouvertes`, "La transition écologique > jeu de données des propositions" from 21 March 2019. The database consists of propositions on the ecological transition written by French citizens during a consultation led by the government in 2019. The code that we used to pre-process the data is given in `Preprocess_NLP.ipynb`. The main steps of this pre-processing are stemming and computation of the TF-IDF vector representation of the answers. For each word, we compute its frequency over the whole database: its document frequency. Then for each user, we compare the frequency of each term in his vocabulary to the document frequency.

In this computer lab, the objective of the model is to predict the answer to the question "Diriez-vous que votre vie quotidienne est aujourd'hui touchée par le changement climatique ?" using the vocabulary used in the answers to the other questions. Note however, that this is mainly a pretext for you to work on optimization algorithms: we do not claim that the model will perform well.

3 Tikhonov regularization

We would like to solve the following logistic regression problem with ℓ_2 regularization:

$$\min_{w_0\in\mathbb{R},w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\big(1+\exp(-y_i(x_i^\top w + w_0))\big) + \frac{\rho}{2}\|w\|_2^2$$

Question 3.1
Calculate the gradient of $f_1 : (w_0, w) \mapsto \frac{1}{n}\sum_{i=1}^{n} \log\big(1+\exp(-y_i(x_i^\top w + w_0))\big) + \frac{\rho}{2}\|w\|_2^2$ and its Hessian matrix. Is the function convex?

Question 3.2
Code a function that returns the value of f_1, its gradient and Hessian matrix.

Please use the dataset described in Section 1 and $\rho = 1/n$. It may be convenient to add a column of ones to the matrix X.

Test your computations using the function `check_grad` on a small-dimensional problem.

Question 3.3
Code Newton's method and run it with initial point $(w_0^0, w^0) = 0$ and stopping criterion $\|\nabla f_1(w)\|_2 < 10^{-10}$.

Plot the norm of the gradient as a function of iterations in logarithmic scale.

Question 3.4
Run it with initial point $(w_0^0, w^0) = e$ where $e_i = 1$ for all i. What are you observing ?

Question 3.5
The classical solution to this problem is to add a line search step. Code Armijo's line search and justify your parameter choices.

4 Regularization for a sparse model

We are still interested in the logistic regression problem but we change the regularizer.

$$\min_{w_0\in\mathbb{R},w\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \log\big(1+\exp(-y_i(x_i^\top w + w_0))\big) + \rho\|w\|_1$$

Question 4.1
Why can't we use Newton's method to solve this problem?

Question 4.2
Write the objective function as $F_2 = f_2 + g_2$ where f_2 is differentiable and the proximal operator of g_2 is easy to compute. Recall the formula for prox_{g_2}. Calculate the gradient of f_2.

Question 4.3
Code the proximal gradient method with line search. Here, we will take $\rho = 0.02$. What stopping test do you suggest?

5 Choice of the regularization parameter

You may not have time to code this part of the computer lab but it may be worth understanding what can be done in order to choose the regularization parameter.

A natural question when considering a regularized machine learning problem is: what is the best value for the regularization parameter ρ? Its goal is to force the model to choose less complex solutions in order to generalize better.

Hence, to evaluate the generalization performance, we are going to split our data into a training set X_{train}, y_{train} and a validation set X_{valid}, y_{valid}. Then, we solve the logistic regression problem using the training set but test its performance on the validation set. Note that the loss function for the validation set is not necessarily the logistic loss. In our case, we are going to consider the 0-1 loss

$$L_{\text{valid}}(w) = \frac{1}{n_{\text{valid}}} \sum_{i=1}^{n_{\text{valid}}} \delta_{\text{sign}(X_{\text{valid}} w)_i, (y_{\text{valid}})_i}$$

where $\delta_{k,l} = 1$ if $k = l$ and $\delta_{k,l} = 0$ if $k \neq l$.

Gathering everything the problem we are trying to solve is the following bilevel optimization problem

$$\max_{\rho \geq 0} L_{\text{valid}}(\hat{w}^{(\rho)})$$
$$\hat{w}^{(\rho)} \in \arg\min \frac{1}{n} \sum_{i=1}^{n} \log\big(1 + \exp(-y_i(x_i^\top w + w_0))\big) + \rho R(w)$$

where $R(w)$ is either $R(w) = \frac{1}{2}\|w\|_2^2$ or $R(w) = \|w\|_1$.

Since this is a complex nonconvex optimization problem, we are going to evaluate the accuracy on a grid of values for ρ, that is $\rho \in \{\rho_0 a^k : k \in \{0, 1, \dots, K\}\}$ for given $\rho_0 > 0$, $0 < a < 1$ and K. Then, we select the parameter $\hat{w}^{(\rho)}$ that has the smallest 0-1 loss on the validation set.

6 Comparison

Question 6.1
Compare the properties of both optimization problems.

Question 6.2
Compare the solutions obtained by the two types of regularization. It may be useful to compute confusion matrices to evaluate the prediction performance of the models.