EP3260: Fundamentals of Machine Learning over Networks
Computer Assignment 1, Due Date: March 4, 2020
Group 4: H. Hellström, M. Tavana, F. Vannella, and M. Zeng

Computer Assignment 1. *(Closed-form solution vs iterative approaches)*
Let us consider

$$\boldsymbol{w}^* = \min_{\boldsymbol{w}\in\mathbb{R}^d} \frac{1}{N}\sum_{i\in[N]} \left\|\boldsymbol{w}^T\boldsymbol{x}_i - y_i\right\|_2^2 + \lambda\left\|\boldsymbol{w}\right\|_2^2,$$

for a dataset $\{(\boldsymbol{x}_i, y_i)\}$. Then, address the following:

(a) Find a closed-form solution for this problem;

Proof. Lets define

$$f\left(\boldsymbol{w}\right) = \frac{1}{N}\sum_{i\in[N]} \left\|\boldsymbol{w}^T\boldsymbol{x}_i - y_i\right\|_2^2 + \lambda\left\|\boldsymbol{w}\right\|_2^2$$

By taking gradient, we have

$$\nabla f\left(\boldsymbol{w}\right) = \frac{1}{N}\sum_{i\in[N]} \left(2\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{w} - 2y_i\boldsymbol{x}_i\right) + 2\lambda\boldsymbol{w}$$

Lets define $X \triangleq [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N]$ and $\boldsymbol{y} \triangleq [y_1, y_2, \ldots, y_N]^T$. We know that $\nabla f\left(\boldsymbol{w}^*\right) = \boldsymbol{0}$, therefore

$$\frac{1}{N}\sum_{i\in[N]} \left(2\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{w}^* - 2y_i\boldsymbol{x}_i\right) + 2\lambda\boldsymbol{w}^* = \boldsymbol{0},$$

$$\left(\lambda I + \frac{1}{N}\sum_{i\in[N]} \boldsymbol{x}_i\boldsymbol{x}_i^T\right)\boldsymbol{w}^* = \frac{1}{N}\sum_{i\in[N]} y_i\boldsymbol{x}_i,$$

$$\left(XX^T + N\lambda I\right)\boldsymbol{w}^* = X\boldsymbol{y}.$$

Therefore,

$$\boxed{\boldsymbol{w}^* = \left(XX^T + N\lambda I\right)^{-1} X\boldsymbol{y}}. \tag{1}$$

□

(b) Consider "Individual household electric power consumption" dataset ($N = 2075259, d = 9$) and find the optimal linear regressor from the closed-form expression;

Solution: we focus on the regression task that aims at predicting the real variable y = Global_intensity, that represents the household global minute-averaged current intensity, based on the following features:

- Date and Time
- global_active_power and global_reactive_power: household global minute-averaged active and reactive power (in Kilowatt)
- voltage: minute-averaged voltage (in Volt)
- sub_metering_1: energy sub-metering No. 1 (in Watt-hour of active energy), corresponding to the kitchen, containing mainly a dishwasher, an oven and a microwave.
- sub_metering_2: energy sub-metering No. 2 (in Watt-hour of active energy). It corresponds to the laundry room, containing a washing-machine, a tumble-drier, a refrigerator and a light.

- sub_metering_3: energy sub-metering No. 3 (in Watt-hour of active energy). It corresponds to an electric water-heater and an air-conditioner.

We pre-process the date and time features by splitting the date-time information in 5 columns that includes Day, Month, Year, Hour, and Minute. The pre-processed feature space that is be used in the regression problem has dimensionality $d = 13$.

We compute the solutions for the regression problem both in closed form, according to Equation 1, and using the *sklearn.linear_model.Ridge* python package. The results in terms of Mean Squared Error (MSE) and computing time (T) are reported in Table (b). Also, in Figure 1 we show a sample of

	MSE	T
Closed-form solution	$1.72 \cdot 10^{-5}$	0.06188
SkLearn Ridge Regression	$1.98 \cdot 10^{-26}$	0.15460

the Global_intensity variable for the true variable in green, and the regressed variable for both the closed-form solution and the one obtained with the Sklearn ridge package in blue and red respectively.

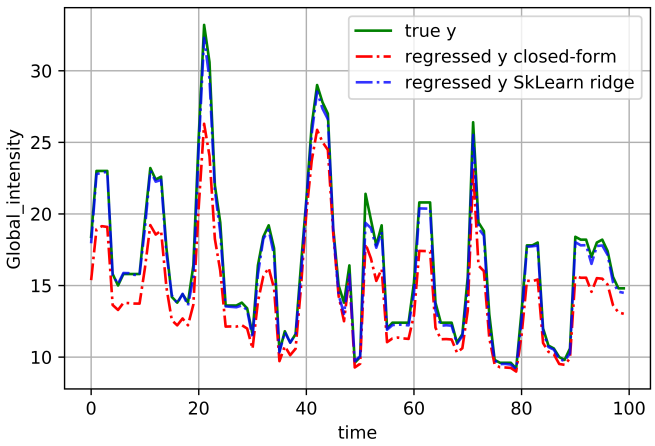

Figure 1: Regression task Individual household dataset

From the results we can observe that the time employed for the closed-form solution is less than half of the one needed for the Sklearn regression. However The Sklearn solution is able to achieve much reliable results in terms of MSE. For these reason we may conclude that there exists a trade-off between computational time to obtain the solution and accuracy in the solution.

(c) Repeat (b) for "Greenhouse gas observing network" dataset ($N = 2921, d = 5232$)) and observe the scalability issue of the closed-form expression;

Solution: we follow the same approach in (b) and we define a regression based-task on the variable that contains the GHG concentrations of synthetic observations, based on the features based on the concentrations of tracers emitted from regions 1-15 as described in the UCI repository. The results for the Closed-form solution and ridge regression MSE and execution time is reported in Table (c).

	MSE	T
Closed-form solution	$5.4 \cdot 10^{6}$	0.048681
SkLearn Ridge Regression	$4.5792 \cdot 10^{-22}$	0.117237

Also, in Figure 2 we show a sample of the predicted variable for the true variable in green, and the regressed variable for both the closed-form solution and the one obtained with the Sklearn ridge package in blue and red respectively.

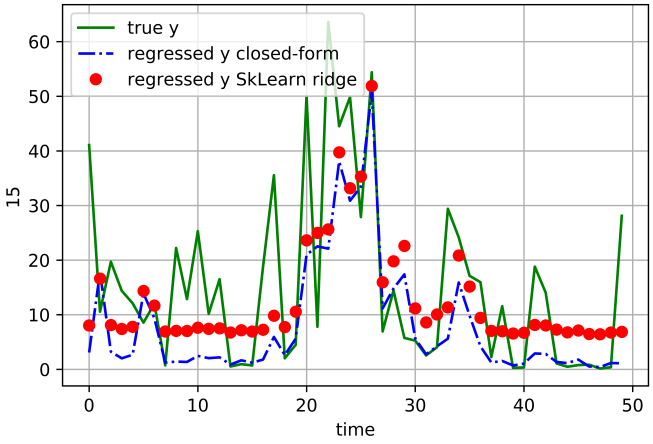

Figure 2: Regression task Greenhouse gas observing network dataset

From the results we can observe that the time employed for the closed-form solution is again less than half of the one needed for the Sklearn regression. WE do not observe a big difference in time with respect to the point (b). However the difference in MSE this time is bigger, since the regressed closed form solution achieves very bad performances.

(d) How would you address even bigger datasets?

Solution: Bigger datasets are computationally unfeasible with closed-form solution because of the fact that needs matrix inversion and multiplications. We may thus rely on methods like Gradient Descent or Stochastic Gradient Descent (SGD) to address these class of problems.