From Basic Machine Learning models to Advanced Kernel Learning

Home Assignment 1

This homework should be uploaded by **November 22, 2024 at 23:59pm** as a **pdf report** together with a **code file (.py or .ipynb)** on the website

`http://pierre.gaillard.me/teaching/kernel_mosig_2024.php`

The password to upload is `kernel2024`. The results and the figures must be included into the pdf report but not the code. Basic python librairies for linear algebra or sampling may be used but not already fully implemented algorithms for SGD, linear regression, logistic regression or KNN.

1 Regression

In this part, we have a dataset $\{(x_i, y_i)\}_{1\le i\le n}$ of input-output pairs in $\mathbb{R}\times\mathbb{R}$. For the simulations, we have

$$y_i = f(x_i) + \varepsilon_i, \quad \text{where} \quad f : x \mapsto \exp(3x) \text{ and } \varepsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)\,.$$

We aim at estimating the unknown function f.

1. Simulate and plot the data with $n = 40$. Separate the data in two sets of equal size $n_{\text{train}} = n_{\text{test}} = 20$: training and testing sets.
2. Give the equation of the coefficient $\hat\theta_{n_{\text{train}}}$ obtained from linear regression (with features $\varphi(x) = (1, x)$) that minimizes the empirial risk with square loss over functions $f_\theta(x) : x \mapsto \langle \theta, \varphi(x)\rangle$ on training data only. Give the value of $\hat\theta_{n_{\text{train}}}$ and plot the obtained function with the data.
3. Estimate the functions obtained with polynomial features $\varphi(x) = (1, x, x^2, \ldots, x^k)$ for $k = 1, \ldots, 9$ and plot them.
4. Plot the evolution of the train and test errors as a function of the polynom degree. Comment.

2 Classification

In this part, the inputs x_i are in $\mathbb{R}^2$ and the outputs y_i in $\{-1, 1\}$. We simulate the data as a Gaussian mixture:

$$y_i \overset{\text{i.i.d.}}{\sim} \mathcal{B}(1/2) \quad \text{and} \quad x_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_{y_i}, \Sigma_{y_i})$$

with

$$\mu_1 = (1,1) \quad \mu_{-1} = (-1,3) \quad \Sigma_1 = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \quad \Sigma_{-1} = \begin{pmatrix} 1 & -1/2 \\ -1/2 & 5/4 \end{pmatrix}.$$

We aim at predicting the probability of a new input x to belong to class 1.

5. Simulate and plot the data with $n = 200$. Separate the data in two sets of equal size $n_{\text{train}} = n_{\text{test}} = 100$: training and testing sets.

KNN and Cross-validation

6. Implement the function kNN that takes as input the number of neighbors k, the training and testing sets and returns the average number of prediction errors on the testing and training sets. Plot theses errors as a function of $k = 1, \dots, 20$. What seems to be the best value for k?

7. Cross-validation. Separate the training set in 5 groups $D_1, \dots, D_5$ of size 20. For each value of k, run 5 times the function kNN, using for $i = 1, \dots, 5$, D_i as testing set and $\cup_{j \neq i} D_j$ as training set and compute the averaged test error obtained. Plot this error as a function of k. What is the best value for k?

8. Plot the heatmap showing the probability of being class 1 predicted by kNN for the above value of k as a function of the inputs.

Logistic Regression

9. Implement the logistic regression estimator $\hat{\theta}_{n_{\text{train}}}$ using SGD (that you implement from scratch choosing a step-size and a number of iterations that seem reasonable) to minimize the empirical risk

$$\mathcal{R}(\theta) = \frac{1}{n} \sum_{i=1}^{n_{\text{train}}} \ell\big(f_\theta(x_i), y_i\big) \quad \text{where} \quad f_\theta : x \mapsto \langle \theta, \varphi(x) \rangle \quad \text{and} \quad \varphi : x \in \mathbb{R}^2 \mapsto (1, x_1, x_2) \in \mathbb{R}^3$$

where ℓ is the logistic loss for outputs in $\{-1, 1\}$. Write the equation of a gradient update performed by SGD for a step size $\eta > 0$ (by writing explicitely the expression of the gradient here). Plot the obtained classification function $f_{\hat{\theta}_{n_{\text{train}}}}$ with the data. Comment.

10. Estimate the functions $f_{\hat{\theta}_{n_{\text{train}}}}$ obtained with polynomial features $\varphi(x) = \big(\{x_1^a x_2^b\}_{0 \le a+b \le k}\big)$ for $k = 1, \dots, 5$ and plot them.

11. Plot the evolution of the train and test errors as a function of the polynom degree using logistic loss and 0-1 loss. What is the best degree? What should it be in theory?

12. For an estimator $\hat{\theta}_{n_{\text{train}}}$ and a new input x, what is the predicted probability of this point of being class 1? Plot the heatmap showing these predicted probabilities for the value of $k = 2$ as a function of the inputs.