ASSIGNMENT 1: LINEAR LEAST SQUARES

MA 590 ST: INTRODUCTION TO SCIENTIFIC MACHINE LEARNING

DUE ON SEPTEMBER 11

Given data pairs $(x_i, y_i)_{i=1}^m$, where $y_i = \Gamma(x_i)$ (noise free).

Exercise 1. Use the codes we discussed in class (or develop your own code) to perform linear least squares of polynomial fitting.

Methods to use :

I. Use polynomials (monomials) $\{1, x, x^2, \cdots, x^n\}$ with normalization.

II. Use the Chebyshev polynomials as the basis.

Consider **three sets of data sets**, where you may use the uniform random points or equispaced deterministic points[1] and

(1) $\Gamma(x) = \sin(10x)$ over $[-1, 1]$;

(2) $\Gamma(x) = 1/(1 + 25x^2)$ over $[-5, 5]$.

(3) $\Gamma(x) = 1$ when $x \in [0, 2]$ and otherwise 0 on $[-2, 0)$.

Compare the performance of the above two methods on these three datasets. Report the following results. For each of the following m ($m = 50, m = 200, m = 300$), plot the standard deviations σ and the coefficients of determination with varying n. The horizontal axis should be n (from 1 to 30), and the vertical axis should be the standard deviations or the coefficients of determination.

Deliverables: three pictures for each m (use one data set in each picture where the results from the two methods are compared). Make sure you comment according to your results on which methods perform the best.

(Optional) You may also want to plot the obtained polynomials on the test points versus the exact function for validation. Other validation methods can be applied. Note that $n < m$ holds for most cases. What happens if $m < n$ (This is common in deep learning)?

[1] Changes in training points will change the condition number and the solution.

Exercise 2 (Theory)**.** Consider the following linear regression

$$\min_{\theta\in\mathbb{R}^{n+1}} \sum_{i=1}^{m} r_i^2(\theta) w_i, w_i > 0, \quad r_i^2(\theta) = y_i - M(x_i;\theta).$$

Here we have

$$M(x;\theta) = \sum_{j=0}^{n} \theta_j e_j(x), n \ll m.$$

(1) Show that the loss function can be expressed in the form of $\|(A\theta - y)\|^2$ for an appropriate diagonal matrix D. Present A, D and y.

(2) Derive the normal equation for the minimization problem by setting the gradient with respect to θ equal to zero.

(3) Compute the Hessian matrix of the loss function. What can you conclude about the convexity based on the Hessian matrix and the uniqueness of the solution?

(4) (optional) Can you give a condition on A such that the solution is unique?

(5) (optional) How does the Tikhonov regularization work in the above formulation?

Exercise 3 (Extension to nonlinear regression)**.** Consider the following *nonlinear* regression

$$\min_{\theta\in\mathbb{R}^{n+1}} L(\theta), \quad L(\theta) = \sum_{i=1}^{m} \rho(r_i(\theta)),\ r_i(\theta) = y_i - M(x_i;\theta).$$

When the metric $\rho(\cdot)$ is not $(\cdot)^2$, this becomes a nonlinear regression as we will not get a linear system by setting the gradient to zero. The solution to this minimization problem is to use an iterative method: start with a 'good guess' and then successively improve the results (hopefully). One way to do this is

- Step 1. (initial guess) Start with a solution from linear least squares, i.e., $\min_{\theta\in\mathbb{R}^{n+1}} \sum_{i=1}^{m} r_i^2(\theta)$. Denote the solution as $\theta^{(0)}$.
- Step 2. (iteration) Solve the following problems

$$\min_{\theta\in\mathbb{R}^{n}} \sum_{i=1}^{m} w_i^{(0)} r_i^2(\theta),\ w_i^{(0)} = \frac{\rho(r_i(\theta^{(0)}))}{r_i^2(\theta^{(0)})}.$$

  Denote the obtained solution as $\theta_i^{(1)}$.
- Step 3. Repeat Step 2 many times by setting $\theta^{(1)}$ to $\theta^{(0)}$.

The stopping condition can be one of the following (or a combination of them): (1) $\left\|\theta^{(1)} - \theta^{(0)}\right\| < \varepsilon_1$; (2) $\left|L(\theta^{(1)}) - L(\theta^{(0)})\right| < \varepsilon_2$; (3) $\left\|\nabla_\theta L(\theta^{(1)})\right\| < \varepsilon_3$. Here ε_i's, $i = 1, 2, 3$ can be different and are user-defined. This iteration method allows us to use the algorithms for linear least square problems (normal equations or pseudo inverse). However, the method above may not always work (This is common for iterative methods). Thus, we always use the maximum iteration number (say 1000) as a stopping condition, in addition to the conditions above.

Apply the above iterative methods using the monomial basis for the following loss functions: (1) Huber loss (2) l^p*,* $p = 1.5$ [2] *to the three data sets in Exercise 1.* You will vary the polynomial orders n to see the performance for each loss. You may use any of the stop conditions with $\varepsilon = 10^{-8}$ plus the maximum iteration number 1000.

Deliverables: For each data set, compare the following error

$$\max_{1\leq j\leq 1000} \left|\Gamma(x_j) - M(x_j, \theta)\right|,$$

where $x_j = a + j\frac{b-a}{1000}$**, where** a **and** b **are specified in Exercise 1. Let** $m = 100$**.**

(Optional) Compare the results in this exercise with those in Exercise 1. (Make sure you use the same training points as in Exercise 1) What is your conclusion on what loss/metrics to use on different data sets? (Does Huber or $l^{1.5}$ loss improve the quality of prediction? why?)

(Hint. In each iteration, we need to solve a problem similar to the one stated in Exercise 2. Can we use the pseudo-inverse for that problem?)

[2] To avoid $w^{(0)}$ having zero in the denominator, we can add a small number 1×10^{-8} in the denominator if needed.