Regression approaches

1 Preliminaries

The goal of regression analysis is to model the expected value of a dependent variable y in terms of the value of an independent variable (or vector of independent variables) x. In polynomial regression, the model reads :

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_d x^d + \epsilon,$$

with $d > 0$ the sought degree of the polynomial, $(\beta_j)_{0 \leq j \leq d}$ are the regression parameters that are to estimate, and ϵ is a noise term accounting for possible modeling errors.

1. Express the polynomial regression model as a system of linear equations $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, considering n data samples.
2. Express and implement the least-squared estimator.
3. Compare visually the observed data and the estimated model. What do you observe when d is too small ? too large ?

2 Regularized regression

An approach to reduce the over-fitting phenomenon arising with too large number of basis polynomial functions is to add a regularization term in the model.

1. Recall the definition of the ridge regression strategy. Give the expression of the ridge estimator, for a given penalty parameter $\alpha > 0$, and implement it. Discuss the influence of α on the visual fitting results.
2. Recall the definition of the lasso strategy. Propose an iterative scheme, to compute the lasso estimator, for a given $\alpha > 0$, and implement it. Discuss the influence of α on the visual fitting results.

3 Robust regression

In the presence of many outliers, or when there is a mismatch between the fitting model and the data, it can be useful to rely on a more robust estimator. This can be done by modifying the least-squared term as follows :

$$F(\beta) = \sum_{i=1}^{n} \rho\left(y_i - [\mathbf{X}\beta]_i\right)$$

The minimization of F can be performed efficiently using the Iterative Least Squares algorithm (see course).

- Express the derivative $\cdot\rho$ and the associated weight function ω, for the following potential functions, parameterized by $\delta > 0$:
  - Huber potential :

$$\rho(e) = \begin{cases} \frac{e^2}{2} \quad \text{if} \quad |e| \leq \delta \\ \delta|e| - \frac{\delta^2}{2} \quad \text{if} \quad |e| > \delta \end{cases}$$

  - Bisquare potential :

$$\rho(e) = \begin{cases} \frac{\delta^2}{6}\left(1 - (1 - \frac{e^2}{\delta^2})^3\right) \quad \text{if} \quad |e| \leq \delta \\ \frac{\delta^2}{6} \quad \text{if} \quad |e| > \delta \end{cases}$$

- Implement the IRLS algorithm, and test it for both potential functions.
- Comment the obtained results.

4 Bonus

Load the bicycle dataset, and apply the above regression methods to it.