KAZAKH-BRITISH TECHNICAL UNIVERSITY

Introduction to Machine Learning
Week 6

Olivier JAYLET

School of Information Technology and Engineering

SIS2 Deadline
24th February 23h50

Recap

So far :

- General understanding of ML
- Understanding of data types (this topic can & should be further explored)
- General understanding of optimization problem and some cost functions
- Some new coding skills (its up to your commitments)
- Linear regression : ability to compute coefficients with your analytical skills

And then

Program for today :

- Study assumptions behind OLS

Rest of the semester :

- Iterative algorithms (next week)
- Logistic regression (for binary outputs)
- Study and understand more ”complex” models, able to capture non-linearity relationships in the data

Gauss-Markov Theorem

The Gauss-Markov theorem states that under some assumptions, OLS is the Best Linear Unbiased Estimator (BLUE).
i.e. the estimator has the smallest variance among other unbiased and linear estimators.

Mean, Variance, Unbiasedness and Efficiency are concepts related to how the estimated values of an estimator behave across repeated samples.

$$\bar{Y} = \frac{1}{N} \sum_{i=1}^{N} Y_i \tag{1}$$

Here, it is computed off of a sample $i = 1, \ldots, N$. But, what if it had been computed off of a different sample ?
What can we say about the distribution of $\hat{Y}$ over repeated samples of size N ?

Unbiasedness

The bias of an estimator is the difference between the expected value of that estimator and the true value of the estimated parameter.

$$\textit{bias}(\hat{\beta}) = \mathbb{E}(\hat{\beta}) - \beta \tag{2}$$

If in repeated samples of a given size the mean value of the estimated parameters $\hat{\beta}$ is equal to β :

$$\mathbb{E}(\hat{\beta}|X) = \beta$$
$$\mathbb{E}(\hat{\beta}) = \beta$$

Then we say that the estimator β^* is unbiased.

Efficiency

Best / most efficient : Among all linear and unbiased estimators, it has the smallest variance, i.e. OLS minimizes the **variance** of the estimator $\hat{\beta}$. OLS produces the most stable estimates with the smallest spread across different samples.
OLS is the best (estimator) because it has the smallest variance:

$$\text{Var}(\hat{\beta}_{OLS}) \leq \text{Var}(\tilde{\beta}_i) \quad \forall\, i$$

Assumptions for OLS

In ordinary least squares (OLS) regression, we need assumptions to ensure that the estimated coefficients are **unbiased**, **efficient**, and **consistent**.

- **Linearity** : The relationship between x and y must be linear. i.e. the model is ”linear in parameters”.
- **Random Sampling** : data are randomly sampled, i.e. they are a representative sample from the population.
- **No Perfect Multicollinearity** : The independent variables should not be perfectly correlated with each other.

Assumptions for OLS

- **Independence of errors** : There is not a relationship between the residuals and the y variable; in other words, y is independent of errors.
- **Homoscedasticity** : The variance of the residuals is the same for all values of x.
- **Zero conditional mean** : The expected value of the residuals is zero for all values of the independent variables.

Assumption OLS.1 (Linear in Parameters)

The model in the population can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \epsilon$$

where the β_j are the population parameters and ϵ is the unobserved error.

- y and the x_k can be nonlinear functions of underlying variables, and so the model is flexible.

Assumption OLS.2 (Random Sampling)

We have a random sample of size n from the population,

$$\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i) : i = 1, ..., n\}$$

- This assumption introduces the data and implies the data are a representative sample from the population.

Assumption OLS.3 (No Perfect Collinearity)

In the sample (and, therefore, in the population), none of the explanatory variables is constant, and there are no exact linear relationships among them.

- The need to rule out cases where $\{x_{ik} : i = 1, ..., n\}$ has no variation for each k is clear from simple regression.

Assumption OLS.4 (Independence of errors)

The independence assumption in linear regression pertains to the independence of errors across observations. It means that the error for one observation is not correlated with the error for another observation.
We express the covariance between errors as:

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = \mathbb{E}[(\varepsilon_i - \mathbb{E}(\varepsilon_i))(\varepsilon_j - \mathbb{E}(\varepsilon_j))] \tag{3}$$

For independence, we require:

$$\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \quad \forall i \neq j \tag{4}$$

Assumption OLS.5 (Homoscedasticity)

The variance of the errors (residuals) should be constant across all levels of the independent variables. This implies that the spread of residuals is the same for all predicted values.

- Heteroskedasticity :

$$\text{Var}(\epsilon_i \mid x_i) = \sigma^2 x_i \tag{5}$$

- Homoskedasticity :

$$\text{Var}(\epsilon_i \mid x_i) = \sigma^2 \tag{6}$$

$$\text{Var}(\epsilon_i) = \sigma^2 \tag{7}$$

Assumption OLS.6 (Zero Conditional Mean)

$$E(u|x_1, x_2, ..., x_k) = 0 \text{ for all } (x_1, ..., x_k)$$

- Remember, the real assumption is $\mathbb{E}(\epsilon|x_1, x_2, ..., x_k) = E(\epsilon)$: the average value of the error does not change across different slices of the population defined by $x_1, ..., x_k$.
- If ϵ is correlated with any of the x_j, OLS.6 is violated. This is usually a good way to think about the problem.
- Mathematically, $\mathbb{E}(\epsilon|X) = 0$ implies $Cov(X_i, \epsilon_{ki}) = 0$
- Literally, we say that independent variables are assumed to be exogenous (not engognous), i.e. they are not influenced by the error term.

Polynomial Regression

- In fact, in most of the model and use case, data aren't linearly related.
- When the linear relationship assumption is violated, we can turn to polynomial regression models such as

$$y = \beta_0 + \beta_1 x + \cdots + \beta_k x^k + \epsilon$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon$$

  to represent the nonlinear patterns:

- The top model is called a kth-order polynomial model in one variable;
- The bottom model is called a quadratic model in two variables.

Polynomial Regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2. \tag{8}$$

Now the model has three parameters $\mathbf{w} = (\beta_0, \beta_1, \beta_2)$, which could be also fitted by optimizing of MSE:

$$L(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i - \beta_2 x_i^2)^2 \to \min_{\beta}. \tag{9}$$

Polynomial regression

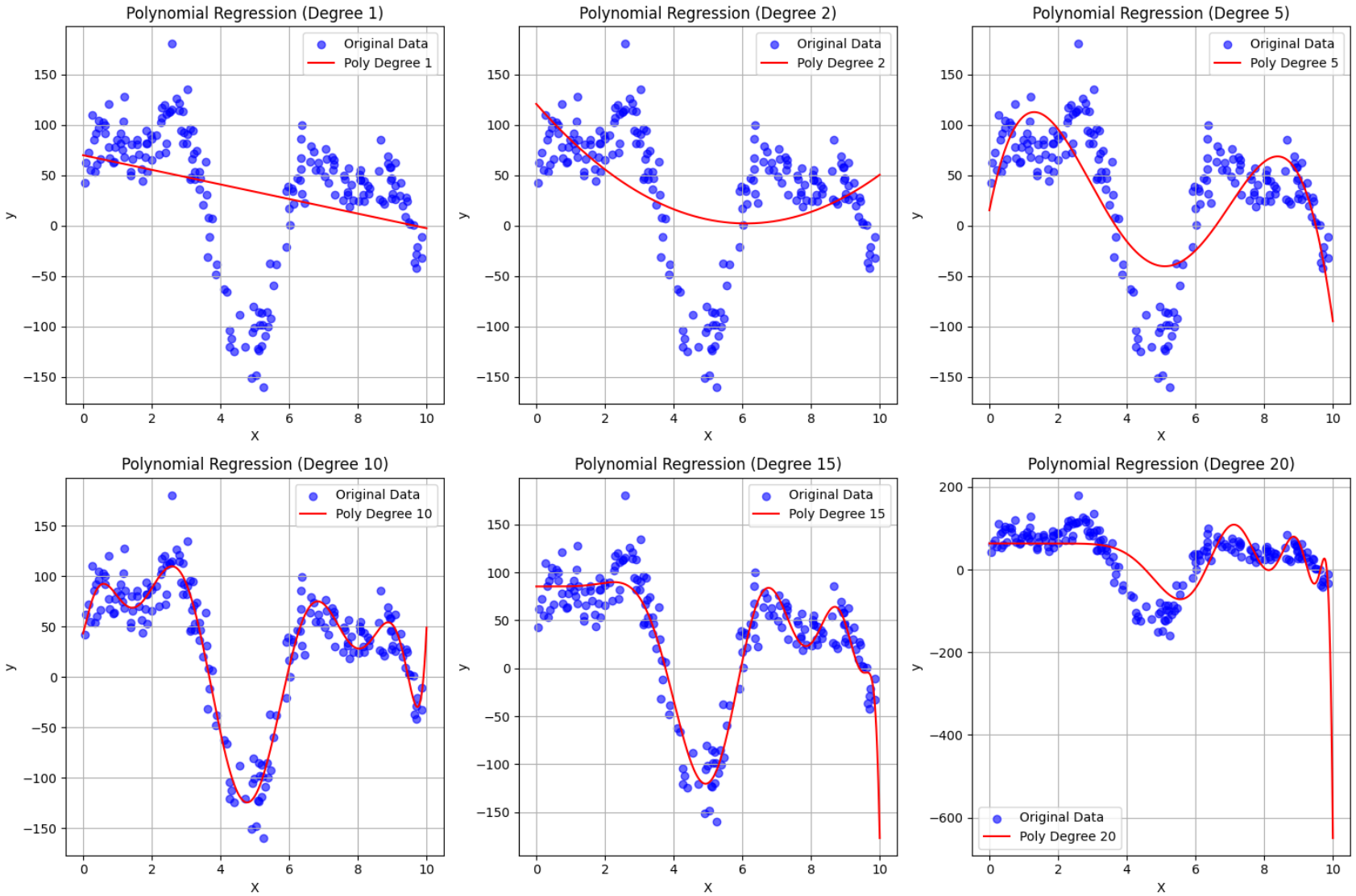
Of course, the degree of the polynomial can be any number $m \in \mathbb{N}$. The model of the polynomial regression is

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m = \sum_{k=0}^{m} \beta_k x^k. \quad (16) \qquad (10)$$

In case of MSE loss the model is fitted via minimizing the function

$$L(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left(y_i - \sum_{k=0}^{m} \beta_k x_i^k \right)^2 \to \min_{\beta}. \quad (17) \qquad (11)$$

Finding the optimal number of polynomials



Thank you for your attention !