Classical Data Analysis
Day 3

Master in Big Data & A.I. Solutions

BARCELONA TECHNOLOGY SCHOOL

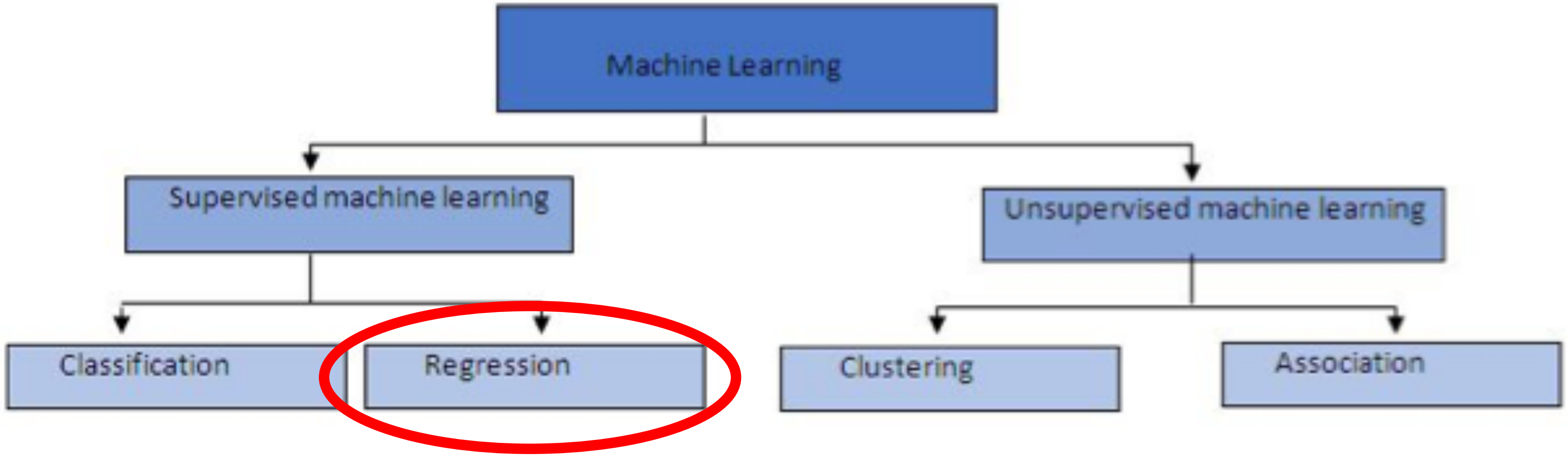
Trainer´s name: Miguel Sanz

BTS Email address: miguel.sanz@bts.tech

Today

- Introduction to data analysis
- Linear Regression
- Logistic Regression
- **Regression analysis (Polynomial, Ridge, Lasso, etc.)**
- Neural Networks (MLPs in depth)
- Support Vector Machines (SVM)
- Decision Trees
- Ensemble Methods (Random Forests, Bagging, etc.)
- Other Classifier (KNN, Naïve Bayes)
- K-Means
- PCA
- Hierarchical Clustering

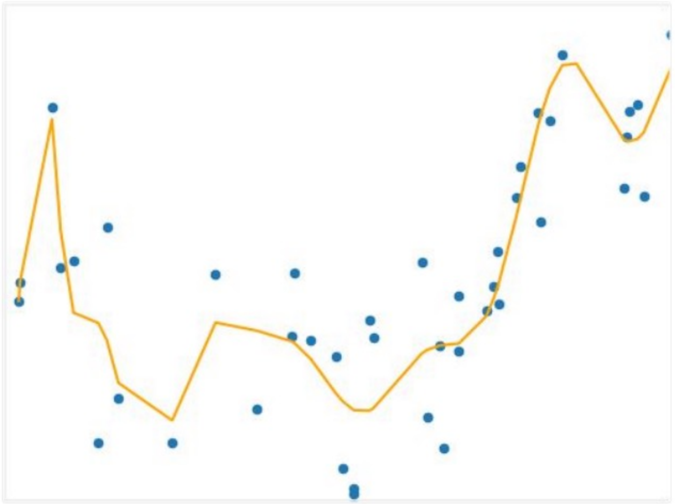
Supervised learning: Regression



Polynomial regression

Limitation of linear Regression

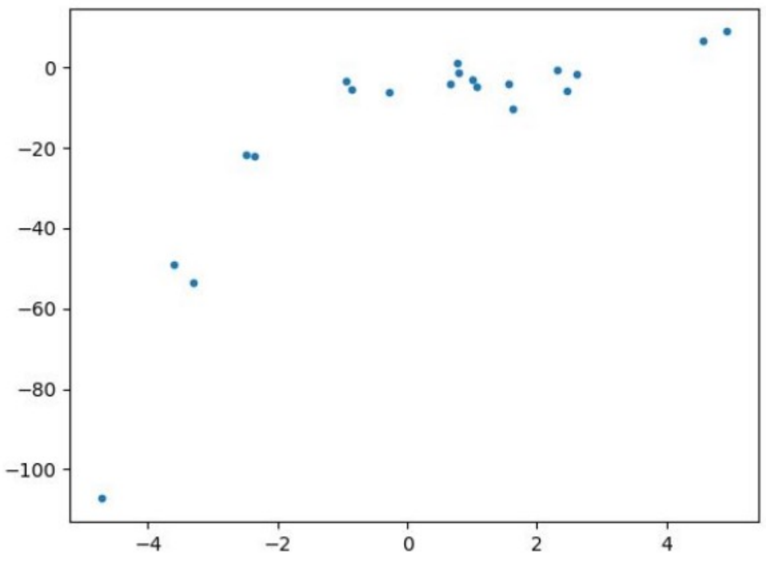
Linear regression requires the relation between the dependent variable and the independent variable to be linear.



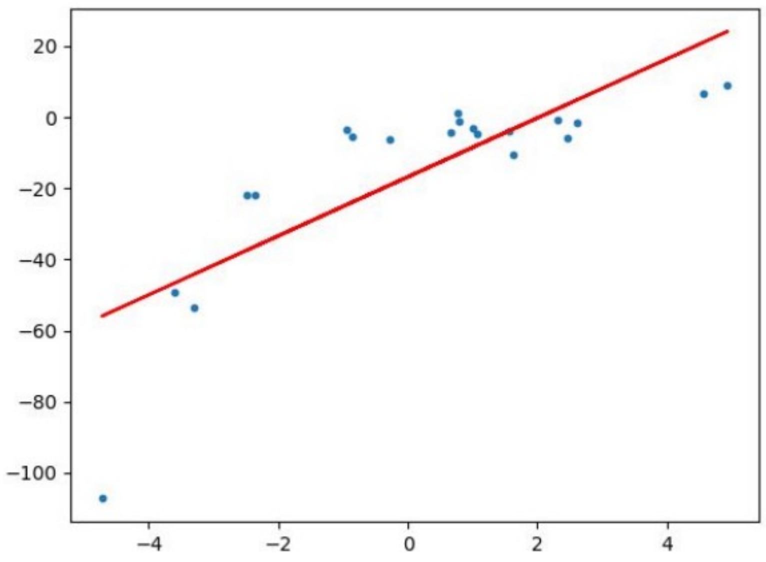
Can linear models be used to fit non-linear data?

Polynomial regression

Why Polynomial Regression?



Linear regression model

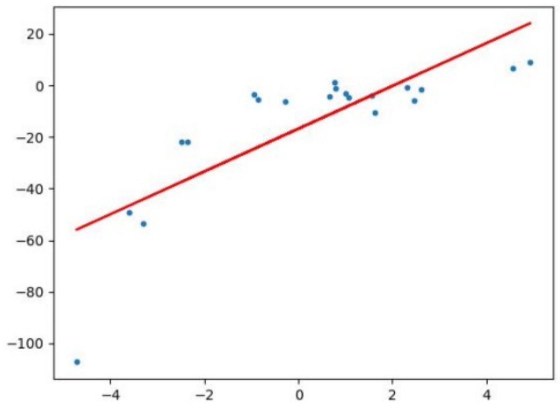


The model is unable to capture the patterns in the data. This is an example of **under-fitting**.

Polynomial regression

Why Polynomial Regression?

Linear regression model



How to overcome **under-fitting**?

Increase the complexity of the model.
We need to add more features to the data

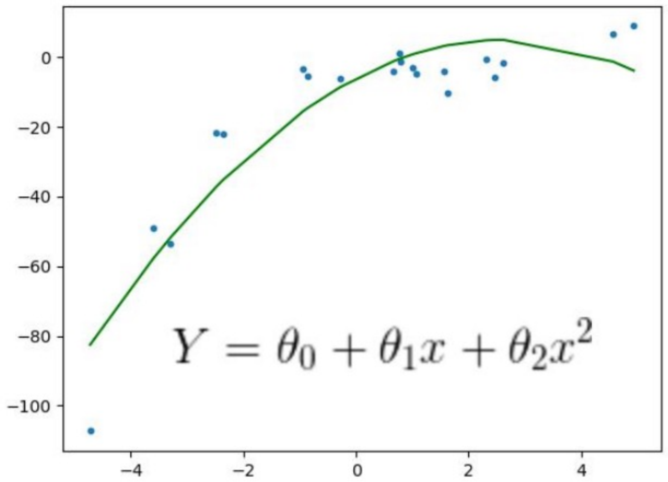
$$Y = \theta_0 + \theta_1 x \quad \Rightarrow \quad Y = \theta_0 + \theta_1 x + \theta_2 x^2$$

This is still considered to be a **linear model** as the coefficients/weights associated with the features are still linear. x^2 is only a new feature.

Polynomial regression

Why Polynomial Regression?

Linear regression model 2nd degree polynomial

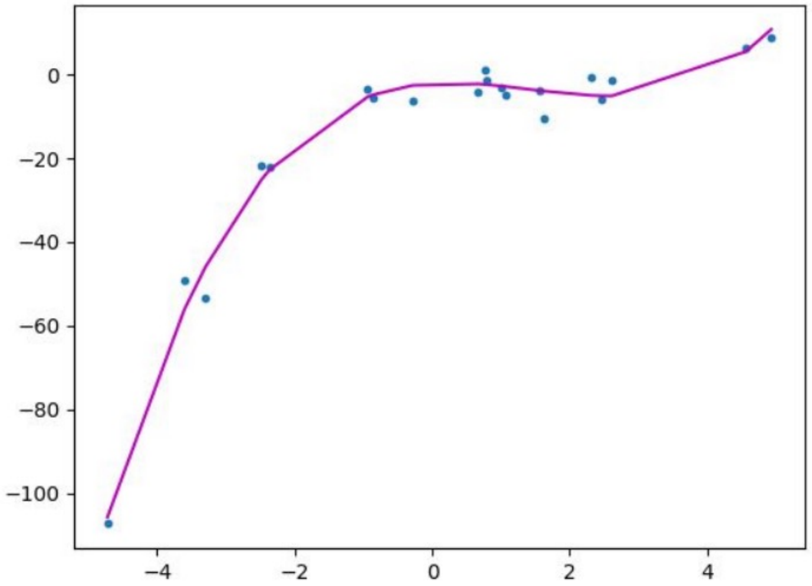


It is clear from the plot that the quadratic curve (2nd degree polynomial) is able to fit the data better than the linear line.

Polynomial regression

Why Polynomial Regression?

Linear regression model polynomial of degree 3

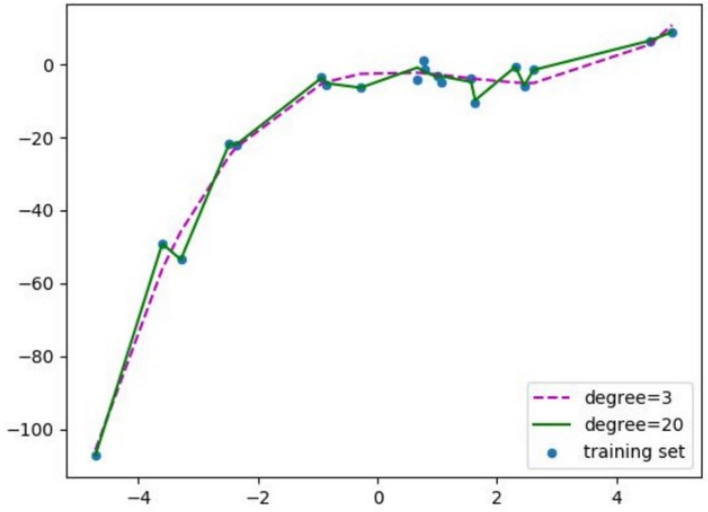


It passes through more data points than the quadratic and the linear plots.

Polynomial regression

Why Polynomial Regression?

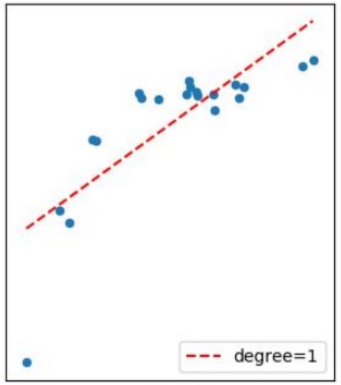
Linear regression model polynomial of degree 20



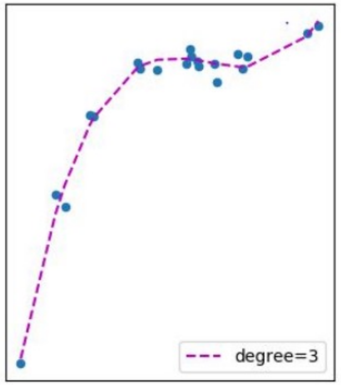
It passes through more data points but here we have a problem: **Overfitting**, i.e., the model is also capturing the noise in the data and even if it works well on the training set it will fail to generalize (the performance in the test set will be poor).

Bias vs Variance Trade off

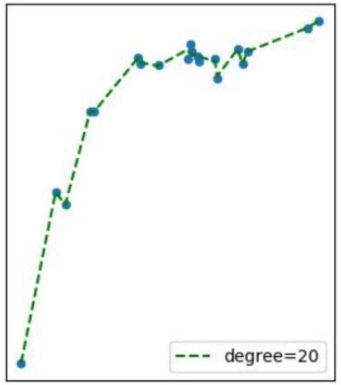
- **Bias**: error due to the model's simplistic assumptions in fitting the data. A **high bias** means that the model is unable to capture the patterns in the data and this results in **underfitting**.
- **Variance:** error due to the complex model trying to fit the data. **High variance** results in **over-fitting** the data. Variance refers to how much the model is dependent on the training data.



Underfit
High Bias
Low Variance

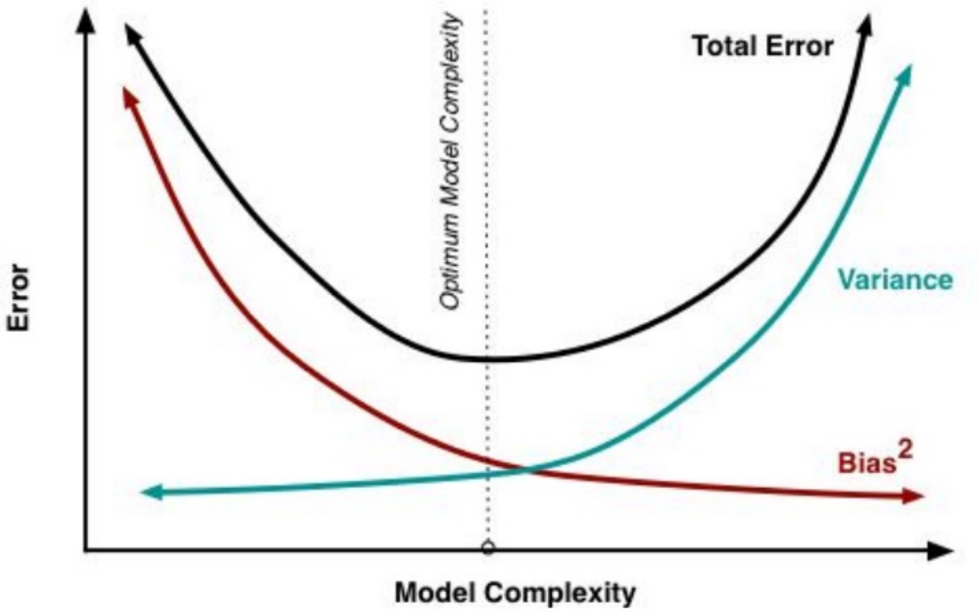


Correct Fit
Low Bias
Low Variance



Overfit
Low Bias
High Variance

Bias vs Variance Trade off



Bias is related with a model failing to fit the training set and variance is related with a model failing to fit the testing set.

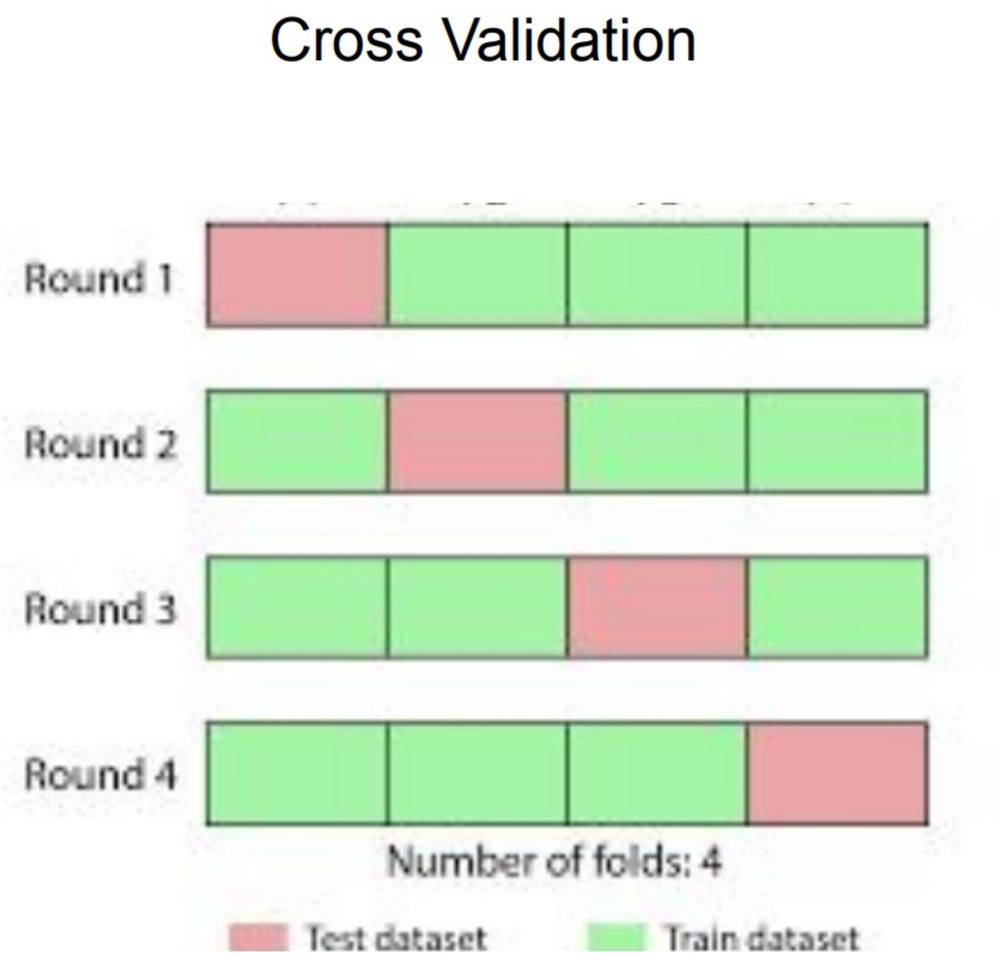
Regularization

How to address overfitting

- In classical linear regression if the model feels like one particular feature is particularly important, the model may place a large weight to the feature. This might lead to **overfitting** in small datasets.

- Potential solutions:
  - Reduce Number of feature (Manually or Model selection algorithm)
  - **Regularization**
  - Collect More data
  - **Cross Validation**
  - Ensembling (Combine prediction from multiple separate models)

Regularization



Regularization

How to address overfitting

- In classical linear regression if the model feels like one particular feature is particularly important, the model may place a large weight to the feature. This might lead to **overfitting** in small datasets.

- **Regularization techniques consist of modifying the cost function, adding a penalty term, to restrict the values of our coefficients.**

- Regularization techniques:
  - **Lasso**, also called L1-Norm
  - **Ridge**, also called L2-Norm
  - **Elastic Net**: is a combination of Lasso and Ridge

Regularization

LASSO

- Adds an additional term to the cost function, adding the sum of the coefficient values (the L-1 norm) multiplied by a constant lambda.

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

- Lambda set the coefficients of the bad predictors mentioned above 0 (**feature selection**).
- If lambda=0, we effectively have no regularization.
- Large lambda leads coefficients to 0.

Regularization

RIDGE

- Sums the **squares of coefficient** values (the L-2 norm) and multiplies it by some constant lambda.

$$\hat{\beta}^{ridge} = \underset{\beta \in \mathbb{R}}{argmin} \|y - XB\|_2^2 + \lambda \|B\|_2^2$$

- will decrease the values of coefficients, but is unable to force a coefficient to exactly 0 (**No feature selection**).
- If lambda=0, we effectively have no regularization
- It will also select groups of colinear features (grouping effect)

Analysis of both Lasso and Ridge regression has shown that neither technique is consistently better than the other; try both methods to determine which to use!

Regularization

Elastic Net

- Includes both L-1 and L-2 norm regularization terms.

$$\hat{\beta} \equiv \underset{\beta}{\operatorname{argmin}}(\|y - X\beta\|^2 + \lambda_2 \|\beta\|^2 + \lambda_1 \|\beta\|_1).$$

- It seems to have the predictive power better than Lasso, while still performing feature selection. We therefore get the best of both worlds, performing **feature selection** of Lasso with the **feature-group selection** of Ridge.

Regularization

- So what do we use?
  - Ridge is faster to converge, Lasso tends to bounce around during Gradient Descent
  - It is always preferable to have at least a little bit of regularization, so try to avoid regression without regularization
  - Ridge is good default to start with, and to check convergence potential (whether we train or not)
  - If you suspect that only few features are useful (and you don't know which ones) you should go for Lasso or ElasticNet because they tend to reduce the useless feature's weights down to zero

Thank you!

BARCELONA TECHNOLOGY SCHOOL