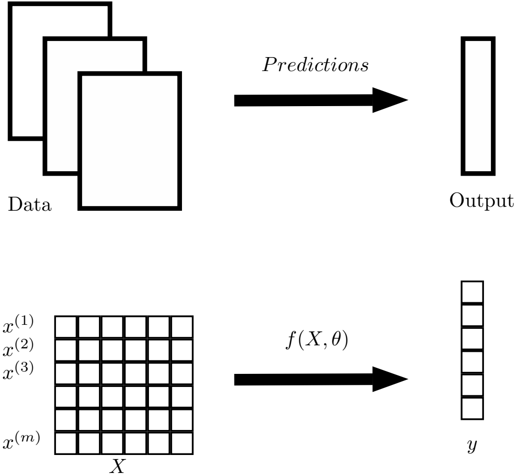
POLIMI GRADUATE SCHOOL OF MANAGEMENT

REGRESSION

Andrea Mor - andrea.mor@polimi.it

SUPERVISED LEARNING

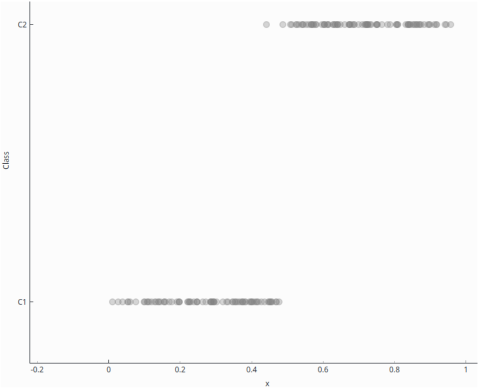


REGRESSION

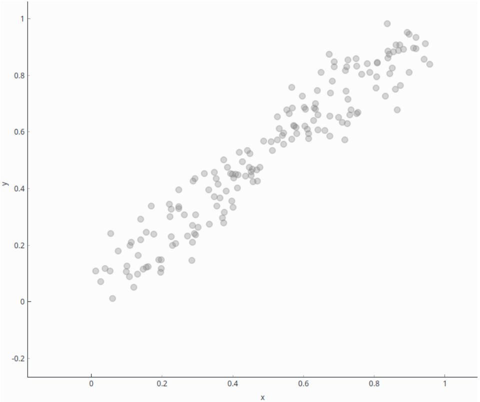
- dataset $\mathcal{D}$ contains n observations and $m+1$ attributes
- m independent/explanatory attributes/features/variables and one dependent variable/target
- observations $x_i, i \in \mathcal{N}$ are points in a n dimensional space. The target variable is denoted as y_i
- **X** is the $n \times m$ matrix of data, **y** is the target vector
- **Y** and $\mathbf{X_j}$ are random variables, $f : \mathbb{R}^m \rightarrow \mathbb{R}$

$$\mathbf{Y} = f(\mathbf{X_1}, \mathbf{X_2}, \cdots, \mathbf{X_m})$$

SUPERVISED LEARNING

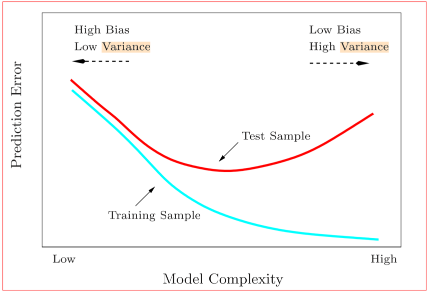


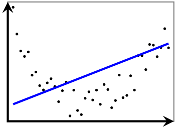
(a) Classification



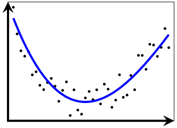
(b) regression

UNDER/OVER-FITTING

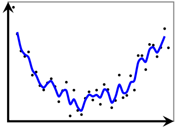




Underfitting



Balance



Overfitting

QUALITY MEASURES - REGRESSION

- Coefficient of determination

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- Mean Absolute Error:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|$$

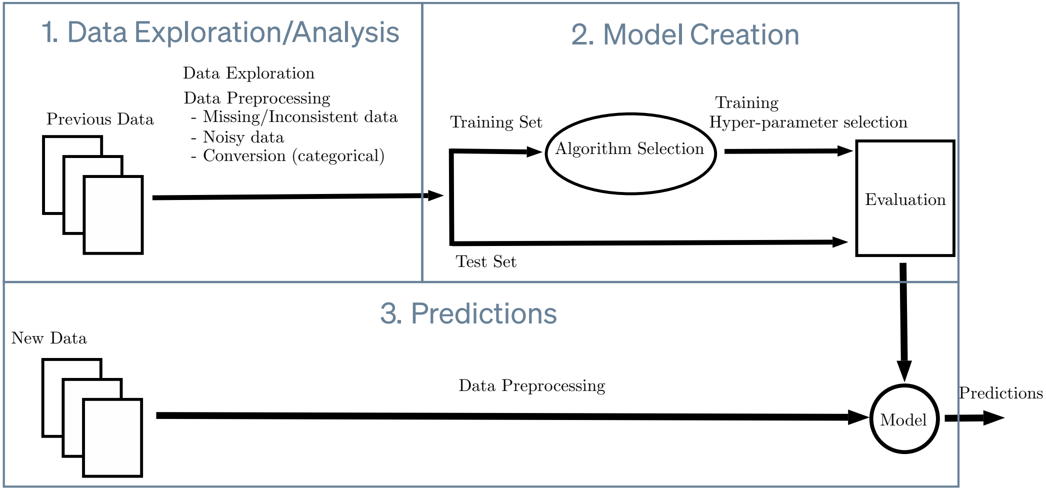
- Mean Squared Error:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Root Mean Squared Error: $RMSE = \sqrt{MSE}$
- Mean Absolute Percentage Error:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

SUPERVISED LEARNING WORKFLOW



1. Data Exploration/Analysis

Previous Data

Data Exploration

Data Preprocessing
- Missing/Inconsistent data
- Noisy data
- Conversion (categorical)

2. Model Creation

Training Set

Algorithm Selection

Training
Hyper-parameter selection

Evaluation

Test Set

3. Predictions

New Data

Data Preprocessing

Model

Predictions

THE ALGORITHMS

REGRESSION MODELS

- Heuristics Methods
  - Nearest Neighbours
  - Regression Trees
- Optimization based Methods
  - Linear models
  - Support vector machine
  - Neural Networks

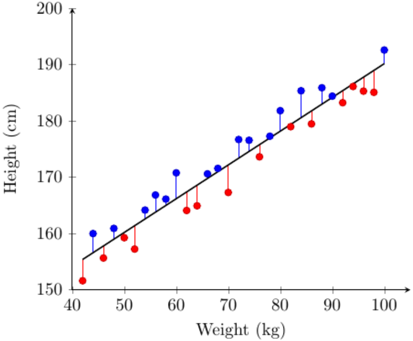
SIMPLE LINEAR REGRESSION

- Deterministic model

$$Y = w\,X + b$$

- Probabilistic model

$$Y = w\,X + b + \varepsilon$$



REGRESSION MODELS (N=1)

- Linear

$$Y = b + \sum_{j=1}^{n} w_j X_j \; = \; b + w_1 X_1 + w_2 X_2 + \cdots + w_n X_n \; = \; b + Xw$$

- Quadratic

$$\begin{aligned} Y &= b + Xw + X^2 d \qquad\quad Z = X^2 \\ &= b + Xw + Zd \end{aligned}$$

- Exponential

$$\begin{aligned} Y &= e^{b+Xw} \qquad\qquad\quad Z = logY \\ &= b + Xw \end{aligned}$$

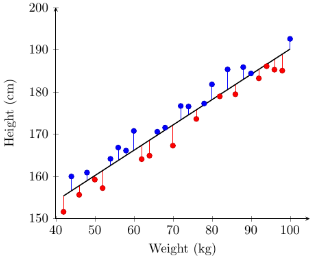
SIMPLE LINEAR REGRESSION

- Residuals

$$e_i = y_i - f(x_i) = y_i - wx_i - b \qquad i \in \mathcal{M}$$

- Least square regression

$$SSE = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} [y_i - wx_i - b]^2$$



LEAST SQUARE LINEAR REGRESSION

$$\frac{\partial SSE}{\partial b} = -2\sum_{i=1}^{m}[y_i - wx_i - b] = 0 \Rightarrow \qquad w\sum_{i=1}^{m} x_i + bm = \sum_{i=1}^{m} y_i$$

$$\frac{\partial SSE}{\partial w} = -2\sum_{i=1}^{m}[y_i - wx_i - b]x_i = 0 \Rightarrow \qquad w\sum_{i=1}^{m} x_i^2 + b\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} x_i y_i$$

LEAST SQUARE LINEAR REGRESSION

$$\frac{\partial SSE}{\partial b} = -2\sum_{i=1}^{m}[y_i - wx_i - b] = 0 \Rightarrow \qquad w\sum_{i=1}^{m} x_i + bm = \sum_{i=1}^{m} y_i$$

$$\frac{\partial SSE}{\partial w} = -2\sum_{i=1}^{m}[y_i - wx_i - b]x_i = 0 \Rightarrow \qquad w\sum_{i=1}^{m} x_i^2 + b\sum_{i=1}^{m} x_i = \sum_{i=1}^{m} x_i y_i$$

$$w^* = \frac{\sigma_{xy}}{\sigma_{xx}}, \quad b^* = \overline{\mu}_y - w^*\overline{\mu}_x$$

$$\sigma_{xx} = \sum_{i=1}^{m}(x_i - \overline{\mu}_x)^2$$

$$\sigma_{xy} = \sum_{i=1}^{m}(x_i - \overline{\mu}_x)(y_i - \overline{\mu}_y)$$

LEAST SQUARE MULTIPLE LINEAR REGRESSION

- If we extend the matrix X with a vector of “ones” then the linear model can be expressed as

$$y = Xw + e$$

-

$$SSE = \sum_{i=1}^{m} e_i^2 = ||e||^2 = (y - Xw)^\top (y - Xw)$$

-

$$\nabla SSE = -2X^\top y + 2X^\top X w = 0$$

-

$$X^\top X w = X^\top y$$

$$\boxed{w^* = (X^\top X)^{-1} X^\top y}$$

-

LEAST SQUARE MULTIPLE LINEAR REGRESSION

- Solution:

$$w^* = (X^\top X)^{-1} X^\top y$$

- Predicted values

$$\hat{y} = Xw^* = (X(X^\top X)^{-1} X^\top)y = Hy$$

- Hat matrix

$$H = X(X^\top X)^{-1} X^\top$$

- Residuals

$$e = y - \hat{y} = (I - H)y$$

GENERAL LINEAR MODELS

- We consider a set of bases functions: polynomials, kernels, etc.

$$Y = \sum_h w_h \, g_h(X_1, X_2, \ldots, X_n) + b + \varepsilon$$

- For example, for $n = 2$

$$Y = X_1 w_1 + X_2 w_2 + X_1^2 w_3 + X_2^2 w_4 + [X_1 X_2] w_5 + b + \varepsilon$$

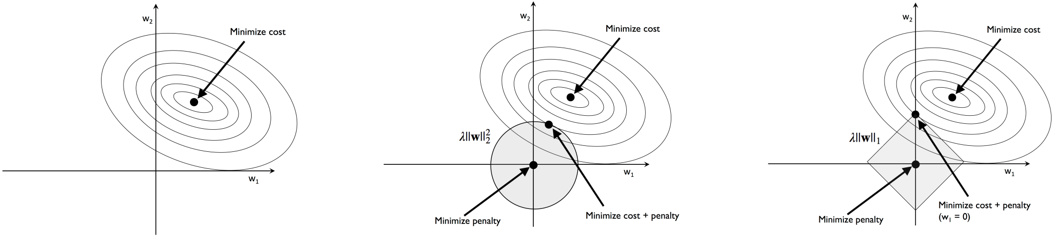
LINEAR MODELS REGULARIZATION

- Ridge:

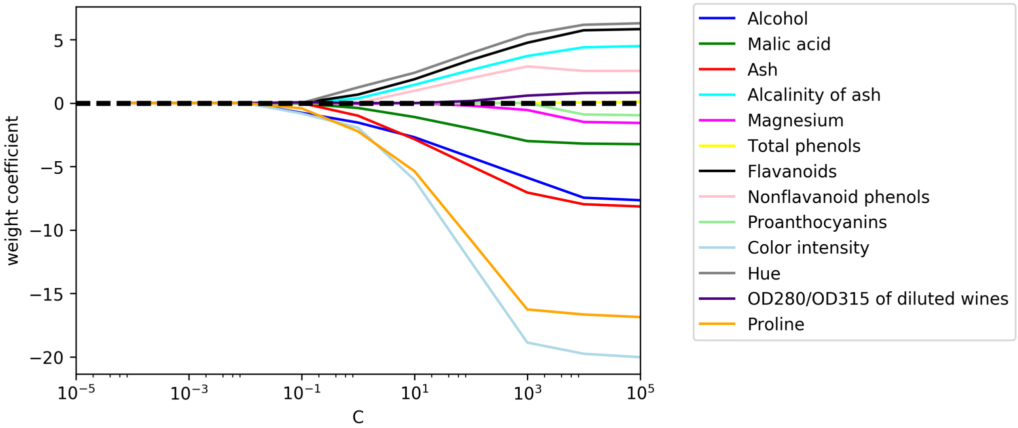
$$\min_{w} \lambda||w||^2 + ||e||^2 = \min_{w} \lambda||w||^2 + (y - Xw)^\top (y - Xw)$$

- Lasso:

$$\min_{w} \lambda|w| + ||e||^2 = \min_{w} \lambda|w| + (y - Xw)^\top (y - Xw)$$



REGULARIZATION EFFECT

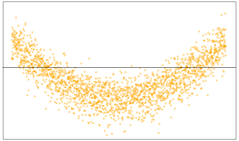


$$C = 1/\lambda$$

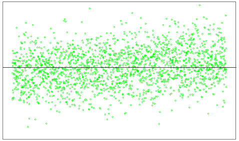
RESIDUAL ASSUMPTIONS

Independence, $E(\varepsilon_i|\mathbf{x_i}) = 0,$ $Var(\varepsilon_i|\mathbf{x_i}) = \sigma^2$

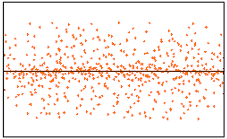
Pattern in Relationship



No Pattern in Relationship

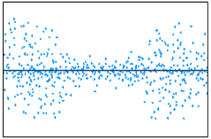


Homoscedasticity



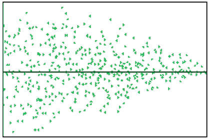
Random Cloud (No Discernible Pattern)

Heteroscedasticity



Bow Tie Shape (Pattern)

Heteroscedasticity



Fan Shape (Pattern)

LINEAR MODELS - SIGNIFICANCE OF COEFFICIENTS

- By assuming residuals independent and normal distribution
- Variance of coefficients

$$Var(\hat{w}) = (X'X)^{-1}\sigma^2 \quad \hat{w} \sim \mathcal{N}(w, (X'X)^{-1}\sigma^2)$$

- Empirical Variance

$$\hat{\sigma} = \frac{SSE}{m-n-1} = \frac{\sum_{i=1}^{m}(y_i - \mathbf{w}'\mathbf{x_i})^2}{m-n-1}$$

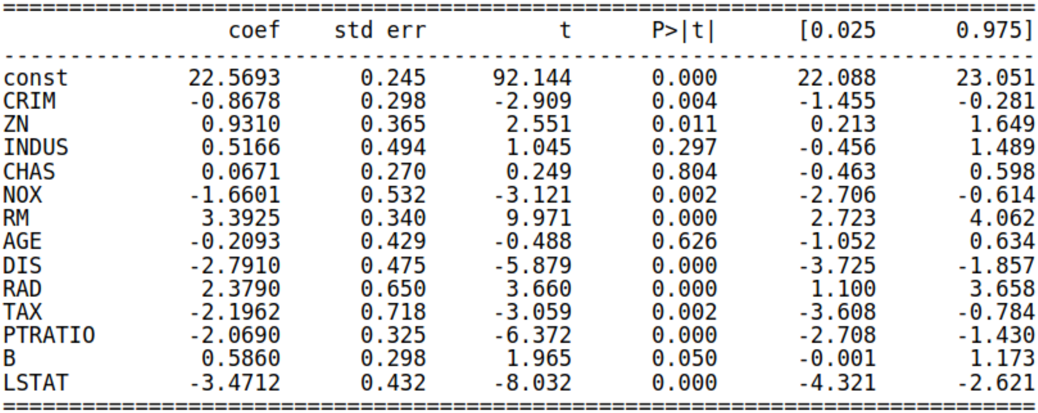
-

$$(m-n-1)\,\hat{\sigma}^2 \sim \sigma^2\chi^2_{m-n-1}$$

- Under the null hypothesis $w_i = 0$ then

$$\frac{\hat{w}_i}{\hat{\sigma}\sqrt{(X'X)_{ii}}} \sim t_{m-n-1}$$

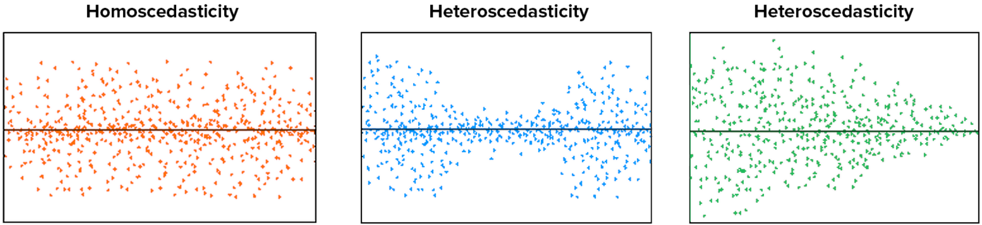
LINEAR MODELS - SIGNIFICANCE OF COEFFICIENTS



```
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         22.5693      0.245     92.144      0.000      22.088      23.051
CRIM          -0.8678      0.298     -2.909      0.004      -1.455      -0.281
ZN             0.9310      0.365      2.551      0.011       0.213       1.649
INDUS          0.5166      0.494      1.045      0.297      -0.456       1.489
CHAS           0.0671      0.270      0.249      0.804      -0.463       0.598
NOX           -1.6601      0.532     -3.121      0.002      -2.706      -0.614
RM             3.3925      0.340      9.971      0.000       2.723       4.062
AGE           -0.2093      0.429     -0.488      0.626      -1.052       0.634
DIS           -2.7910      0.475     -5.879      0.000      -3.725      -1.857
RAD            2.3790      0.650      3.660      0.000       1.100       3.658
TAX           -2.1962      0.718     -3.059      0.002      -3.608      -0.784
PTRATIO       -2.0690      0.325     -6.372      0.000      -2.708      -1.430
B              0.5860      0.298      1.965      0.050      -0.001       1.173
LSTAT         -3.4712      0.432     -8.032      0.000      -4.321      -2.621
==============================================================================
```

NORMAL RESIDUAL ASSUMPTION

- Graphical distribution



Random Cloud (No Discernible Pattern)

Bow Tie Shape (Pattern)

Fan Shape (Pattern)

- Graphically compare error distribution against a normal distribution with QQ-plots
- Apply an hypothesis test to check the normality of the errors (Kolmogorov–Smirnov, D’Agostino, etc.)

MULTI-COLLINEARITY OF FEATURES

$$Var(\hat{w}_j) = \frac{\sigma^2}{(m-1)Var(X_j)} \times \frac{1}{1-R_j^2}$$

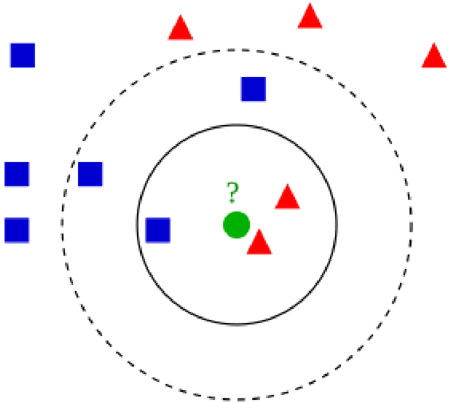
where R_j is the coefficient of determination for the linear regression explaining X_j with the remaining explanatory variables.

Variance inflation factor

$$VIF_j = \frac{1}{1-R_j^2}$$

empirically if bigger than 5/10 indicates the existence of multicollinearity.

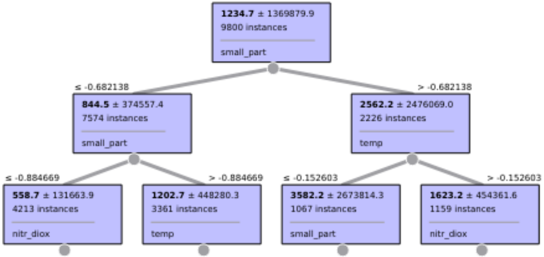
KNN K-NEAREST NEIGHBOURS



Main Parameters

- k : number of neighbours
- neighbour weights
- distances

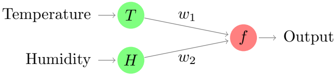
REGRESSION TREE

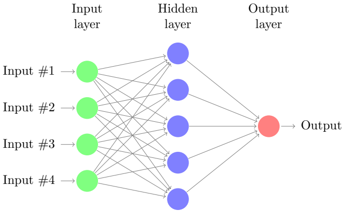


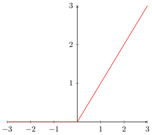
Main Parameters

- variability measure: mse (i.e., reduction in variance), mae, ...
- max_depth
- min_samples_split: minimum number of samples to split an internal node
- min_sample_leaf: minimum number of samples required to be at a leaf node

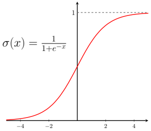
MULTI-LAYER PERCEPTRON







ReLU

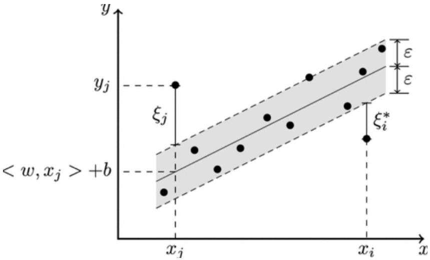


Sigmoid

Main Parameters

- hidden_layer_sizes: $(n_1, n_2, \ldots, n_L)$
- activation: identity, logistic, tanh, relu
- alpha regularization term parameter
- Resolution algorithm parameters: solver, tol, batch_size, learning_rate, max_iter.

SVR



$$\min_{w,b,\zeta,\zeta^*} \frac{1}{2}\|w\| + C\sum_{i=1}^{n}(\zeta_i + \zeta_i^*)$$

subject to $y_i - w^T\phi(x_i) - b \leq \varepsilon + \zeta_i,$

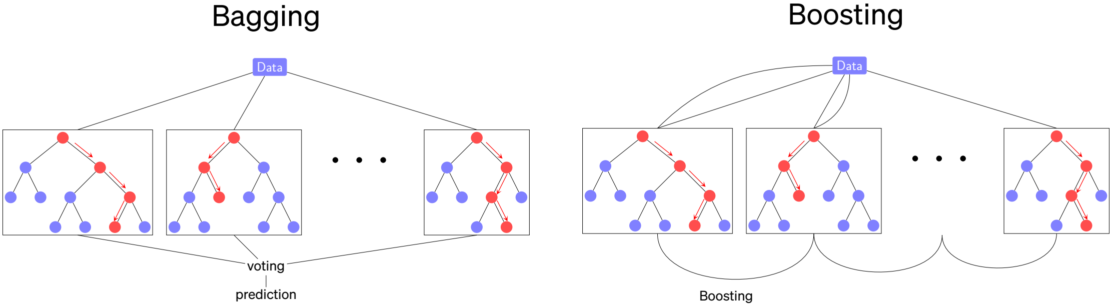
$$w^T\phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*,$$

$$\zeta_i, \zeta_i^* \geq 0, i = 1, ..., n$$

Main Parameters

- C: inverse of regularization strength
- ε: tolerance
- kernel
- Resolution algorithm parameters

ENSEMBLE METHODS



THANK YOU

POLIMI GRADUATE SCHOOL OF MANAGEMENT