BABD

INTERNATIONAL MASTER IN BUSINESS ANALYTICS AND BIG DATA

Supervised Learning - Regression

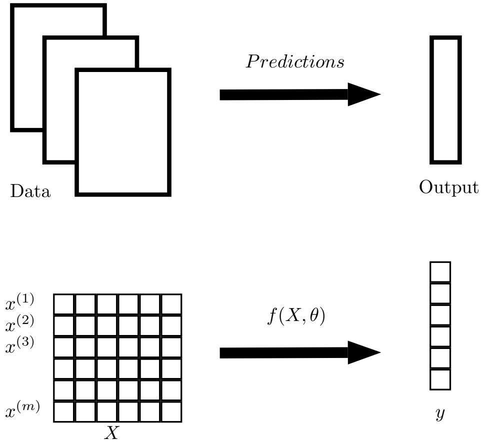
MIP POLITECNICO DI MILANO GRADUATE SCHOOL OF BUSINESS

Cefriel POLITECNICO DI MILANO

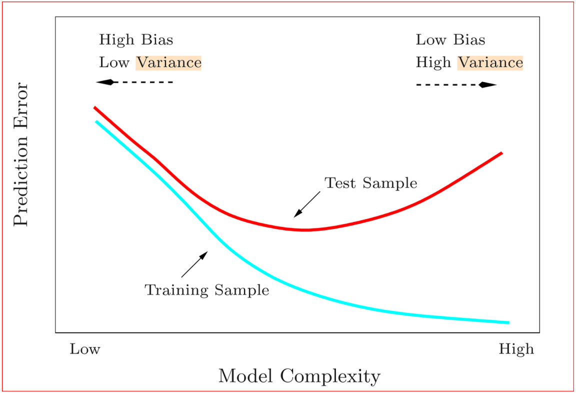
IBM

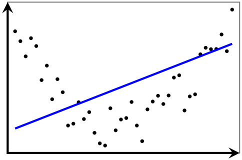
FT Executive Education Ranking 2019 | FT European Business Schools Ranking 2018 | 2018 BEST B-SCHOOLS | EFMD EQUIS ACCREDITED | AMBA ACCREDITED | EFMD EOCCS CERTIFIED

Supervised Learning

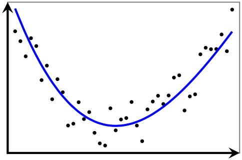


Under/Over-fitting

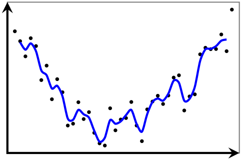




Underfitting



Balance



Overfitting

Quality measures - Regression

- Coefficient of determination

$$R^2 = 1 - \sum_{i=1}^{m}(y_i - \hat{y}_i)^2 / \sum_{i=1}^{m}(y_i - \bar{y})^2$$

- Mean Absolute Error :

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

- Mean Squared Error :

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

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

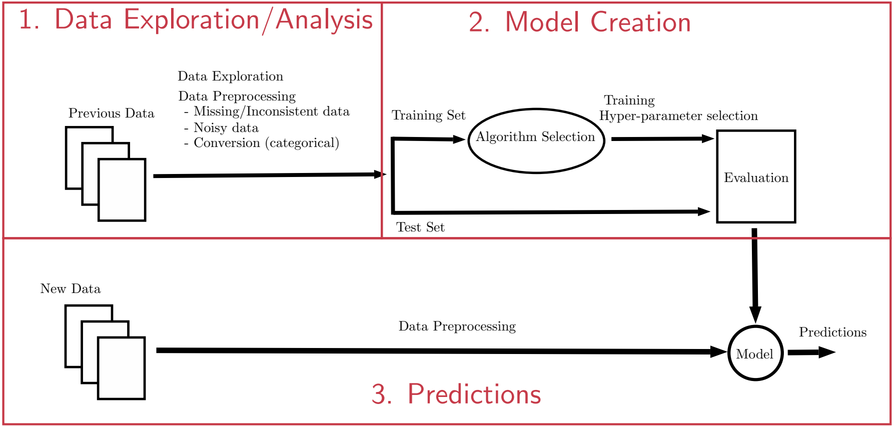
Regression model

- Dataset $\mathcal{D}$ contain m observation/records and $n+1$ attributes.
- n independent features and a single continuous dependent attribute: target
- We can represent our dataset as a numeric matrix X of dimension $m \times n$
- Our aim is to find a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ such that the *associated error* to our prediction

$$\hat{y} = f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$$

is small

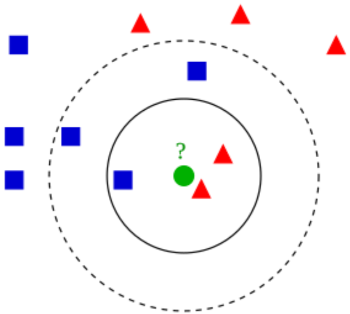
Supervised Learning Workflow



Regression Models

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

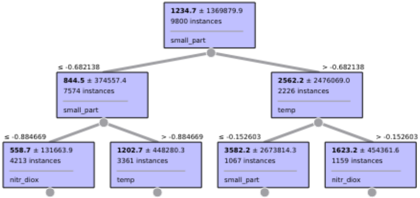
KNN K-nearest Neighbours



Main Parameters

- k : number of neighbours
- neighbour weights
- distances

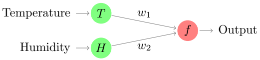
Regression tree

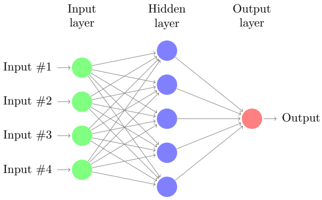


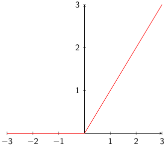
Main Parameters

- variability measure: mse (variance from mean), mae (error from median)
- 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

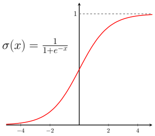
Multi-Layer Perceptron







ReLU

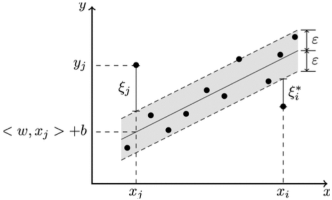


Sigmoid

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



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

$$\begin{aligned} \text{subject to } & y_i - w^T\phi(x_i) - b \leq \varepsilon + \zeta_i, \\ & w^T\phi(x_i) + b - y_i \leq \varepsilon + \zeta_i^*, \\ & \zeta_i, \zeta_i^* \geq 0, i = 1, ..., n \end{aligned}$$

Main Parameters

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

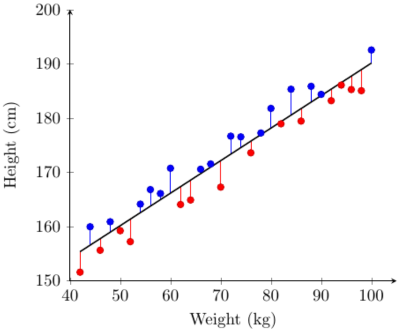
Simple linear regression

- Deterministic model

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

- Probabilistic model

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



Regression models

- **linear**

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

- **quadratic**

$$Y = b + wX + dX^2 \qquad Z = X^2$$

$$Y = b + wX + dZ.$$

- **exponential**

$$Y = e^{b+wX} \qquad Z = \log Y \qquad Z = b + wX.$$

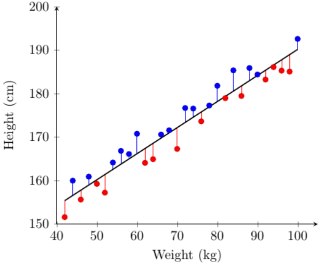
Simple linear regression

- Residuals

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

- Least square regression

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



Least square linear regression

- $$\hat{w} = \frac{\sigma_{xy}}{\sigma_{xx}},$$
  $$\hat{b} = \bar{\mu}_y - \hat{w}\bar{\mu}_x,$$

- $$\bar{\mu}_x = \frac{\sum_{i=1}^{m} x_i}{m}, \qquad \bar{\mu}_y = \frac{\sum_{i=1}^{m} y_i}{m}.$$

$$\sigma_{xy} = \sum_{i=1}^{m}(x_i - \bar{\mu}_x)(y_i - \bar{\mu}_y),$$
$$\sigma_{xx} = \sum_{i=1}^{m}(x_i - \bar{\mu}_x)^2,$$
$$\sigma_{yy} = \sum_{i=1}^{m}(y_i - \bar{\mu}_y)^2,$$

Least square multiple linear regression

- **probabilistic model**

$$Y = w_1 X_1 + w_2 X_2 + \cdots + w_n X_n + b + \varepsilon.$$

$$\mathbf{e} = (e_1, e_2, \ldots, e_m)$$

$$\mathbf{w} = (b, \mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n)$$

- **extend matrix X by a vector with all components = 1**

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}.$$

Least square multiple linear regression

- **sum of squared residuals**

$$\mathrm{SSE} = \sum_{i=1}^{m} e_i^2 = \|\mathbf{e}\|^2 = \sum_{i=1}^{m} (y_i - \mathbf{w}'\mathbf{x}_i)^2$$
$$= (\mathbf{y} - \mathbf{X}\mathbf{w})'(\mathbf{y} - \mathbf{X}\mathbf{w}).$$

- **null partial derivatives**

$$\frac{\partial\,\mathrm{SSE}}{\partial \mathbf{w}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{w} = \mathbf{0}.$$

- **normal equation**

$$\mathbf{X}'\mathbf{X}\mathbf{w} = \mathbf{X}'\mathbf{y}.$$

- **minimum point**

$$\hat{\mathbf{w}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Least square multiple linear regression

- values predicted by model

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\mathbf{w}} = (\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{y} = \mathbf{H}\mathbf{y}$$

- hat matrix

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

- residuals

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

Linear Models generalizations

- Regularization

$$\min_{\mathbf{w}} RR(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^{m} (y_i - \mathbf{w}'\mathbf{x}_i)^2$$
$$= \min_{\mathbf{w}} \lambda \|\mathbf{w}\|^2 + (\mathbf{y} - \mathbf{X}\mathbf{w})'(\mathbf{y} - \mathbf{X}\mathbf{w}).$$

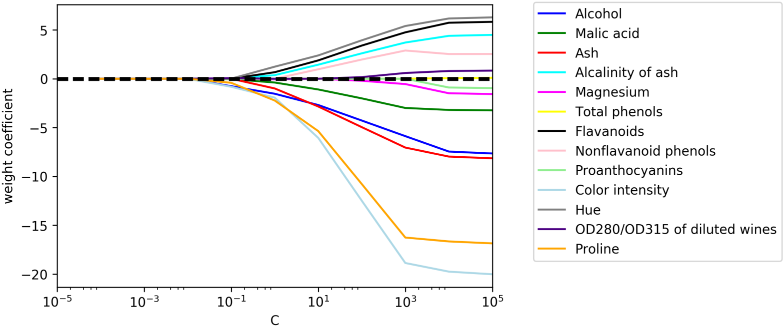
$$\min_{\mathbf{w}} LR(\mathbf{w}, \mathcal{D}) = \min_{\mathbf{w}} \lambda |\mathbf{w}| + \sum_{i=1}^{m} (y_i - \mathbf{w}'\mathbf{x}_i)^2$$
$$= \min_{\mathbf{w}} \lambda |\mathbf{w}| + (\mathbf{y} - \mathbf{X}\mathbf{w})'(\mathbf{y} - \mathbf{X}\mathbf{w}).$$

- General Linear Models
  - **Functions g_h represent any set of bases, such as polynomials, kernels and other groups of nonlinear functions**

$$Y = \sum_{h} w_h g_h(X_1, X_2, \ldots, X_n) + b + \varepsilon$$

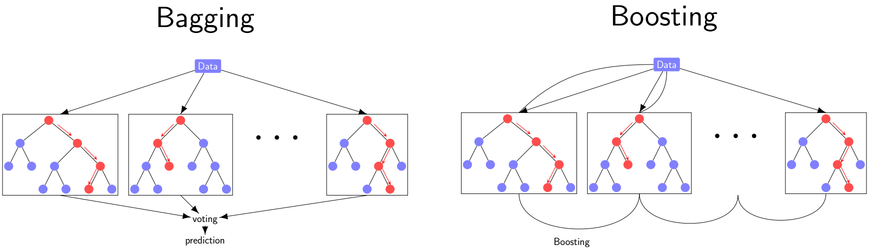
  - **Coefficients w_h and b can be determined through the minimization of the sum of squared errors. Function SSE in this formulation is more complex than for linear regression, solution of the minimization problem more difficult**

Regularization effect



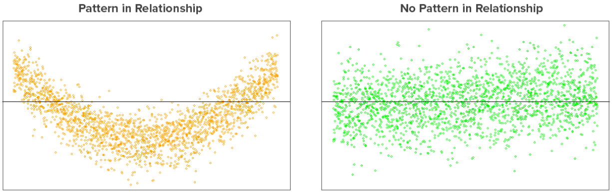
$$C = 1/\lambda$$

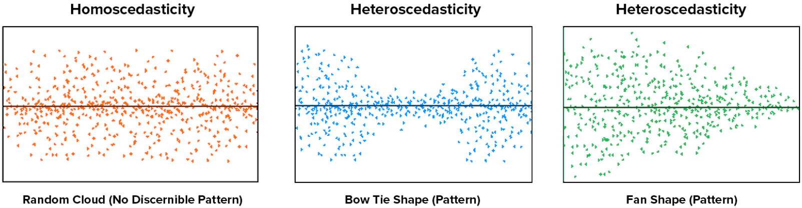
Ensemble Methods



Residual assumptions

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





Linear models - Significance of coefficients

- By assuming residuals independent and normal distributed
- 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

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

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

- 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

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

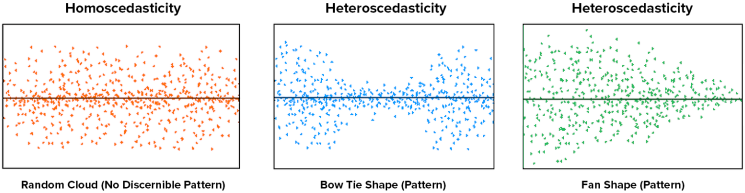
Variance inflation factor

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

if bigger than five indicates the existence of multicollinearity.

Normal residual assumption

- Graphical distribution



- 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.)