VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF COMPUTER SCIENCE

Machine Learning

Regression

Jan Platoš

November 22, 2023

Department of Computer Science
Faculty of Electrical Engineering and Computer Science
VŠB - Technical University of Ostrava

Regression

Regression

- In several cases the class label is numerical.
- The goal is to minimize the squared error of prediction.
- The predicted class label is also referred to as the response variable, dependent variable or regressand.
- The feature variables are referred to as explanatory variables, input variables, predictor variables, independent variables or regressors.
- The prediction process is referred to as regression modeling.

Regression - Linear Regression

- Let D be an $n \times d$ data matrix.
- The feature vector $\overline{X_i}$, i-th row of D, is the d-dimensional input vector.
- The corresponding response variable is y_i.
- In linear regression, the dependence of each response variable y_i on the $\overline{X_i}$ is modeled as a linear relationship:

$$y_i \approx \overline{W} \cdot \overline{X_i} \qquad \forall i \in \{1, \ldots, n\}$$

- $\overline{W} = (w_1, \ldots, w_d)$ is a d-dimensional vector of coefficients that needs to be learned from the training data to minimize the unexplained error

$$E = \sum_{i=1}^{n} \left(\overline{W} \cdot \overline{X_i} - y_i \right)^2$$

Regression - Linear Regression

- The bias b may be modeled:
  - as a part of the $\overline{W}$ and artificial dimension in training data that is set to 1.
  - removed due to mean-centered the data matrix and response variables.
- The data are normalized/standardized to ensure similar scaling and weighting for all attributes.

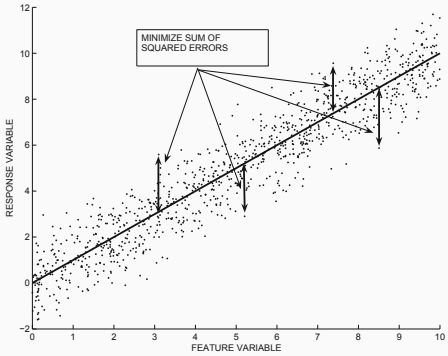


Figure 1: Linear regression

Regression - Linear Regression

- The objective function O, squared error of prediction, the have to be minimized by determination of W is defined as (where $\overline{y} = (y_1, \ldots, y_n)$):

$$O = \sum_{i=1}^{n} \left(\overline{W} \cdot \overline{X_i} - y_i\right)^2 = \left\| D\overline{W}^T - \overline{y} \right\|^2$$

- The gradient of O with respect to $\overline{W}$ is a vector $2D^T(D\overline{W}^T) = D^T\overline{y}$.
- Setting the gradient equal to o we get:

$$D^T D\overline{W}^T = D^T\overline{y}$$

- When $D^T D$ is invertible then $\overline{W}_T = (D^T D)^{-1} D^T\overline{y}$.
- otherwise we may use pseudo-inverse $D^+ = (D^T D)^{-1} D^T$ and then $\overline{W}_T = D^+\overline{y}$.

Regression - Linear Regression Regularization

- The objective function O minimizes the SSE:

$$O = \left\| D\overline{W}^T - \overline{y} \right\|^2$$

- *Ridge regression* reduce the size of the coefficient and minimizes chaotic behavior.

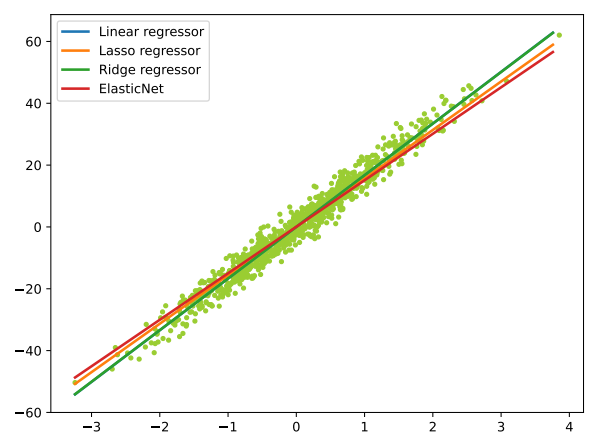
$$O = \left\| D\overline{W}^T - \overline{y} + \lambda \left\| \overline{W} \right\|^2 \right\|^2$$

- *Lasso regression* eliminates small weight (produces sparse model).

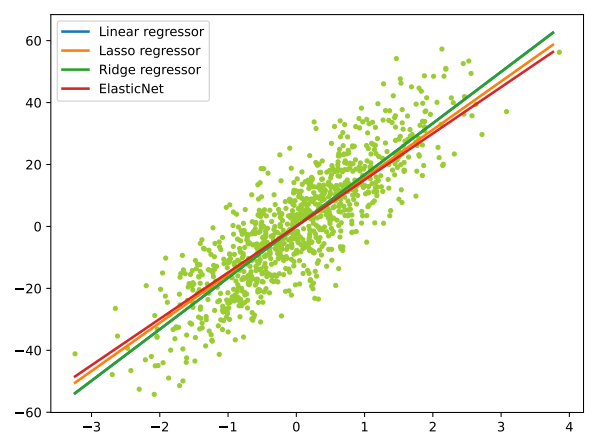
$$O = \left\| D\overline{W}^T - \overline{y} + \lambda \sum_{i=1}^{d} |w_i| \right\|^2$$

- Mixing model between *Lasso* and *Ridge* is called *ElasticNet*.

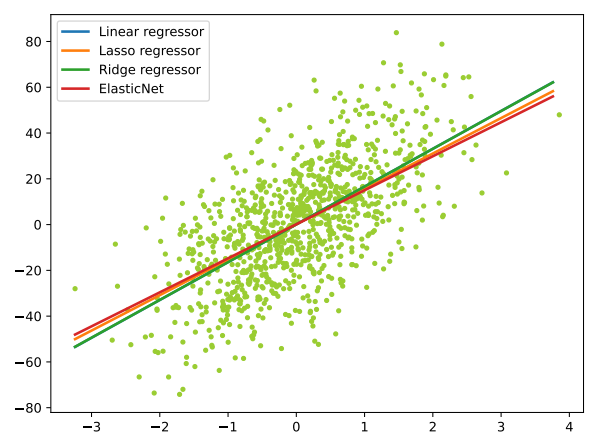
Regression - Linear Regression



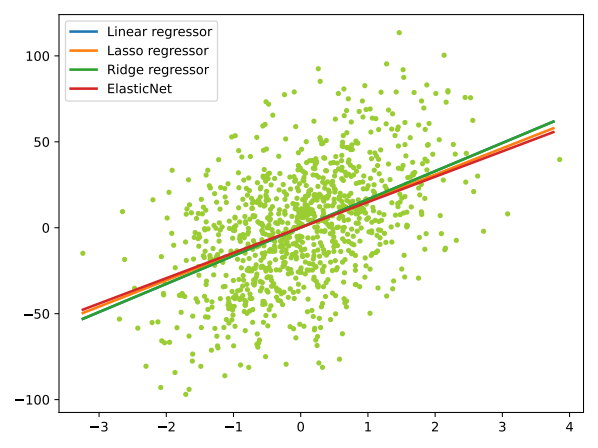
Regression - Linear Regression



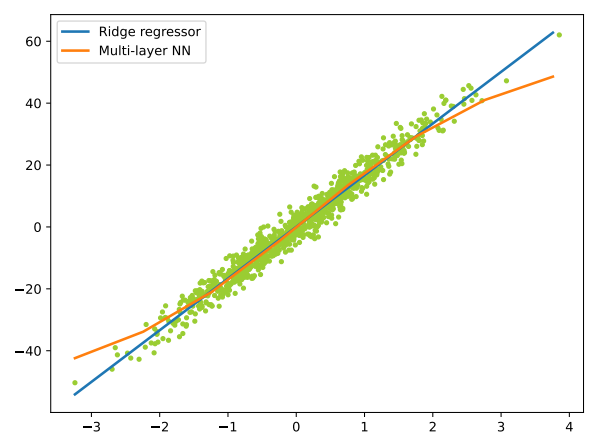
Regression - Linear Regression



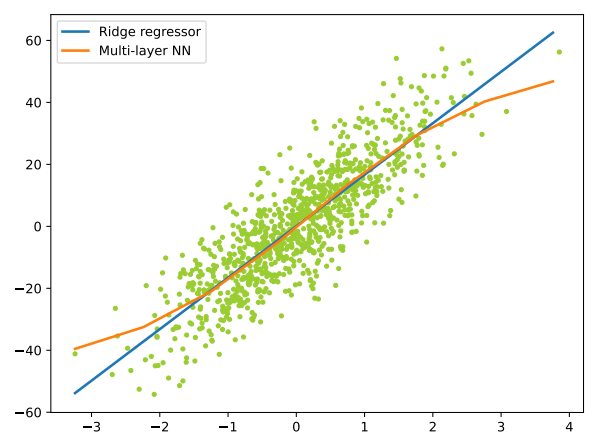
Regression - Linear Regression



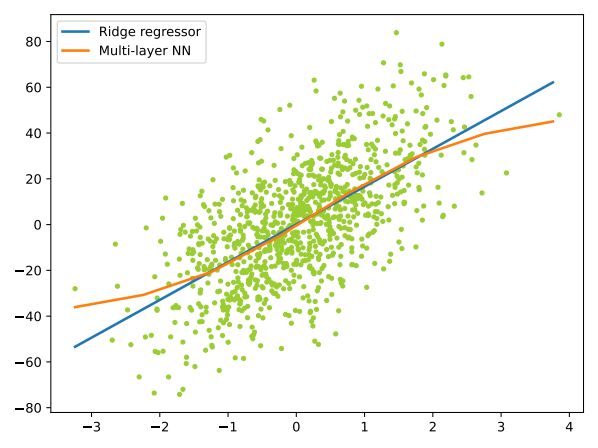
Regression - Linear Regression



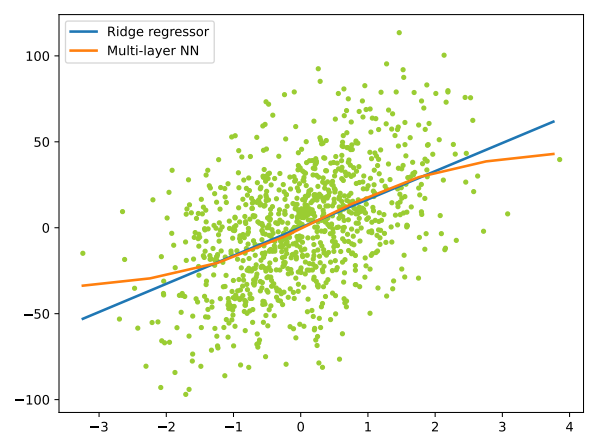
Regression - Linear Regression



Regression - Linear Regression



Regression - Linear Regression



Regression - Generalized Linear Models

- Intuitively, we expect that a constant change in a feature variable leads to the constant change in the response variable.
- This is not true in many cases, e.g. the height of the person is not linearly dependent on the age of a person.
- Moreover, such features will never be negative.
- The generalized linear models (GLM) solves this problems.
- Each responsible variable y_i is modeled as an outcome of a probability distribution with mean $f(\overline{W} \cdot \overline{X_i})$.
- The function $f(\cdot)$ is referred to as the *mean function* and its inverse as *link function*.
- The selection of the mean/link function and corresponding probability distribution should maximize effectiveness and interpretability of the model.

Regression - Generalized Linear Models

- Intuitively, we expect that a constant change in a feature variable leads to the constant change in the response variable.
- This is not true in many cases, e.g. the height of the person is not linearly dependent on the age of a person.
- Moreover, such features will never be negative.
- The generalized linear models (GLM) solves this problems.
- Each responsible variable y_i is modeled as an outcome of a probability distribution with mean $f(\overline{W} \cdot \overline{X_i})$.
- The function $f(\cdot)$ is referred to as the *mean function* and its inverse as *link function*.
- The selection of the mean/link function and corresponding probability distribution should maximize effectiveness and interpretability of the model.
- the response variable is modeled using probability, the $\overline{W}$ is determined using maximum likelihood approach.

Regression - Nonlinear and polynomial regression

- Linear regression cannot capture nonlinear relationships.
- Linear approach may be applied on the derived features.
- The derivation means an application of non-linear functions on the each input points.
- The new set of points may have different number of dimensions.

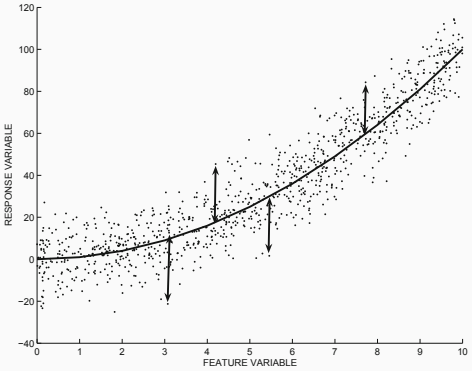


Figure 2: Linear regression

Regression - Nonlinear and polynomial regression

- The new set of m features denoted as $h_1(\overline{X_i}) \dots h_m(\overline{X_i})$ for the i-th data point.
- The $h(\cdot)$ represents nonlinear transformation from the d-dimensional input feature space into 1-dimensional space.
- The size of the new dataset D_h is $n \times m$.
- The linear relationship is then defined as :

$$y = \sum_{i=1}^{m} w_i h_i(\overline{X})$$

- The *polynomial regression* expands the number of features by factor r

$$\overline{X} = (x_1, \dots, x_d) \Rightarrow \overline{X^h} = (x_1, x_1^2, x_1^3, \dots, x_1^r, x_2, \dots, x_d^r)$$

- The Kernel trick is allows by the reformulation of the regression problem with dot-products.

Regression - Regression Trees

- In reality, local linear regression may be quite effective even when the relationships is nonlinear.
- This is used in Regression Trees.
- Each test instance is classified with its locally optimized linear regression by determining its appropriate partition.
- The partition is determined using split criteria in the internal nodes, i.e. the same as the Decision trees.
- The general strategy of tree construction is the same as for Decision Trees.
- The splits are univariate (single variable/axis parallel).
- The changes are done in splitting criterion determination and in the pruning.
- The number of points used for training need to be high to avoid over-fitting

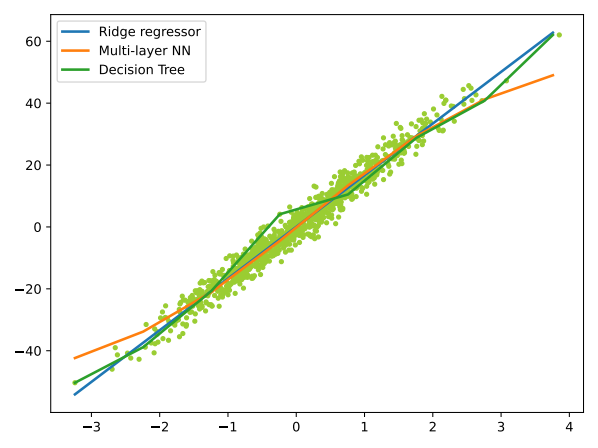
Regression - Regression Trees - Splitting criterion

- Due to numeric nature of the class variable, error-based measure have to be used instead of entropy or Gini index.
- The regression modeling is applied on each child resulting from potential split.
- The aggregated squared error of prediction of all training points is computed.
- The split point with the minimum aggregated error is selected.
- The complete regression modeling is computationally very expensive.
- An average variance of the numeric class variable may be used instead.
- The linear regression models are constructed at the leaf nodes after the tree is created.
- This results in larger trees but it its computational expensiveness is much lower.

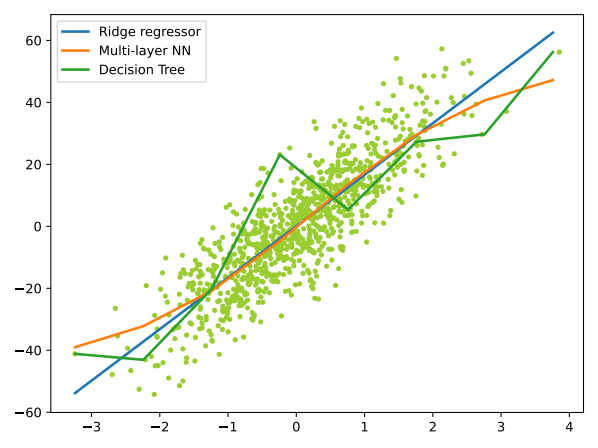
Pruning criterion

- A portion of the training data is not used during construction phase.
- This set is used for evaluation of the squared error of the prediction.
- Leaf nodes are iteratively removed if the accuracy not decreases.

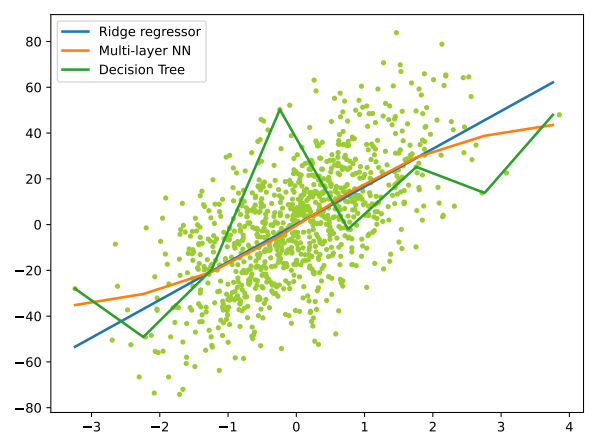
Regression - Regression Trees



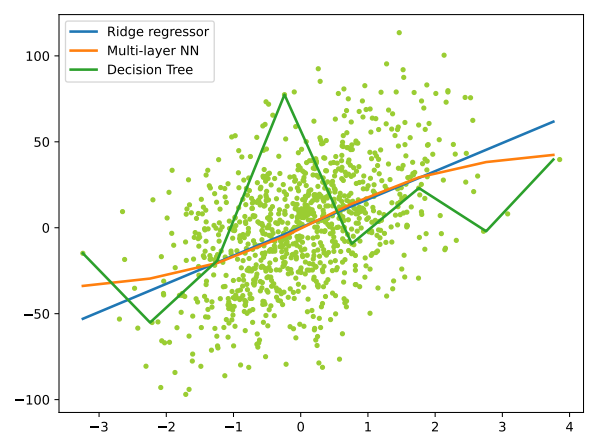
Regression - Regression Trees



Regression - Regression Trees



Regression - Regression Trees



Regression- Assessing Model Effectiveness

- **Mean Absolute Error (MAE)** - is the average of the absolute difference between the predicted and actual value. It is highly affected by outliers.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - g(\overline{X_i})|$$

Regression- Assessing Model Effectiveness

- **Mean Squared Error (MSE)** - is the average of the squared difference between the predicted and actual value. It is differentiable and may be used for optimization.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - g(\overline{X_i})\right)^2$$

- **Root Mean Squared Error (RMSE)** - is the square root of the average of the squared difference of the predicted and actual value. The root mean is able penalize large errors.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - g\left(\overline{X_i}\right)\right)^2}$$

Regression- Assessing Model Effectiveness

- The effectiveness of the linear regression models can be evaluated with a measure known as **R^2-*statistics*** or *coefficient of determination*.
- The standard Sum of Squared Error is defined for a model $g(\overline{X})$ as:

$$SSE = \sum_{i=1}^{n} \left(y_i - g(\overline{X_i})\right)^2$$

- The Squared Error of the response variable about its mean is defined as:

$$SST = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \frac{y_j}{n} \right)^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Regression- Assessing Model Effectiveness

- The R^2-*statistics* is then defined as:

$$R^2 = 1 - \frac{SSE}{SST}$$

- The value is always between 0 and 1 and higher are more desirable.
- For high dimension data, **adjusted** version is more accurate:

$$R^2 = 1 - \frac{(n-d)SSE}{(n-1)SST}$$

- The R^2-*statistics* is not applicable on the nonlinear models.
- The nonlinear regression may be evaluated using pure SSE.

Regression- Assessing Model Effectiveness

- **Mean Average Percentage Error (MAPE)** - is the average percentage error between the predicted and actual value.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} \left| \frac{y_i - g(\overline{X_i})}{y_i} \right|$$

- **Symmetric Mean Average Percentage Error (SMAPE)** - is the symmetric average percentage error between the predicted and actual value.

$$SMAPE = \frac{100}{n} \sum_{i=1}^{n} \frac{\left| y_i - g(\overline{X_i}) \right|}{\frac{|y_i| + \left| g(\overline{X_i}) \right|}{2}}$$

Questions?