POLITECNICO
DI TORINO

Data Science & Machine Learning for Engineering Applications

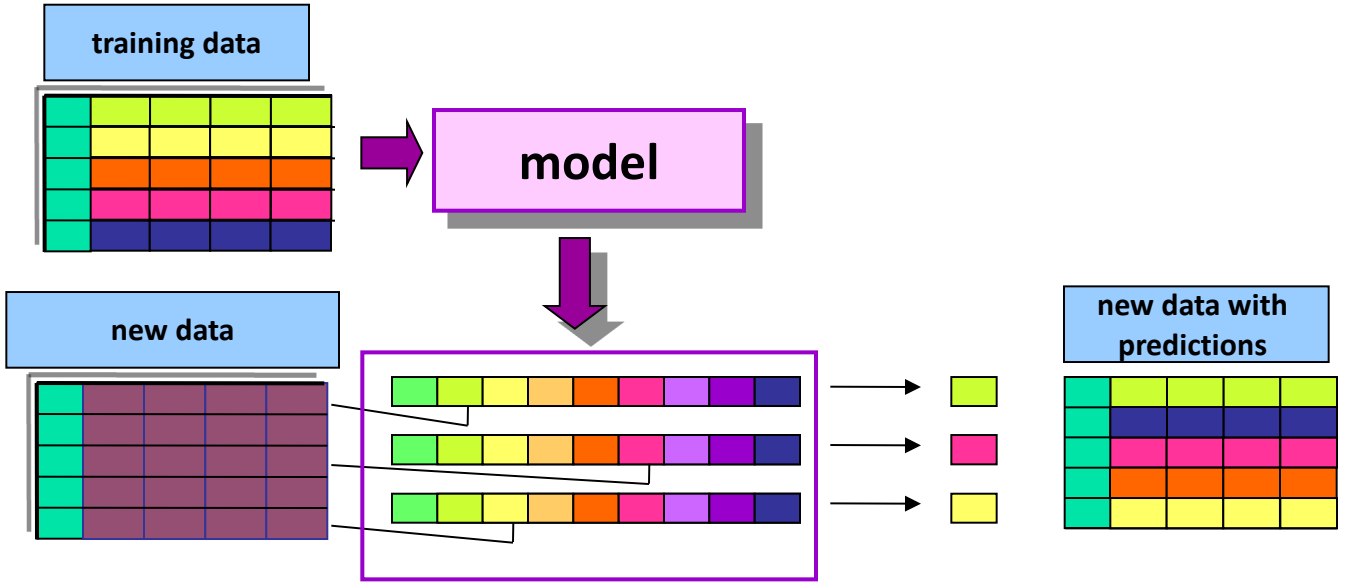
Regression Analysis: Fundamentals

DataBase and Data Mining Group

Tania Cerquitelli and Elena Baralis

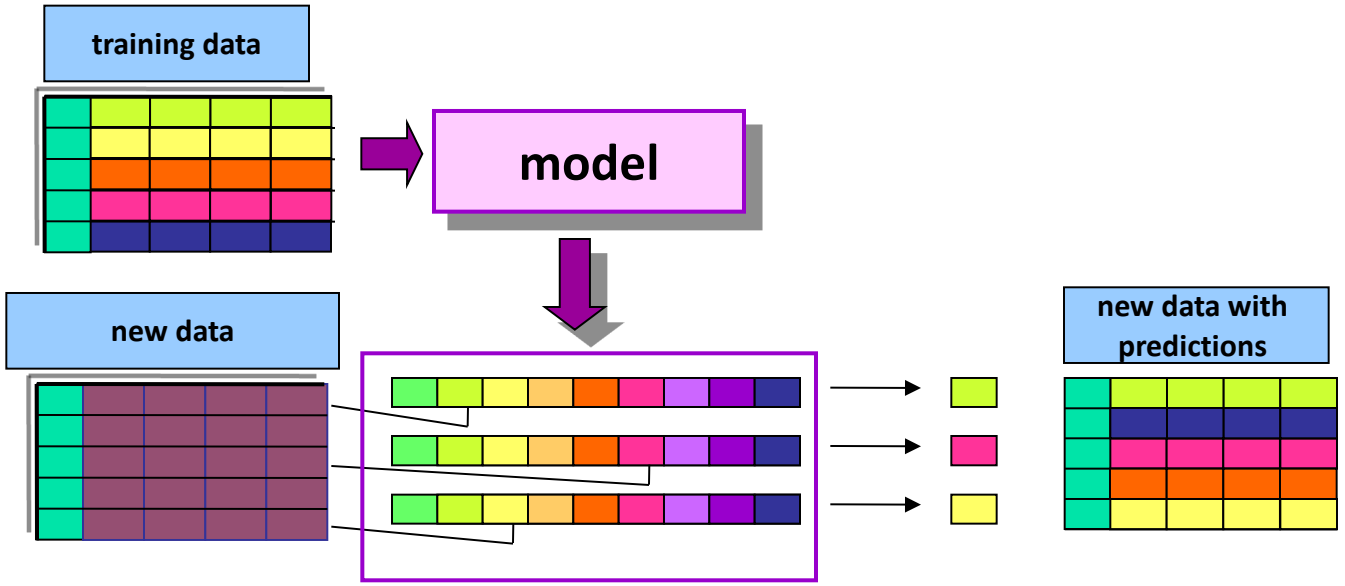
Introduction to the regression analysis

- Objectives
  - Prediction of a **numerical target variable**
  - Definition of an **interpretable model** of a given phenomenon



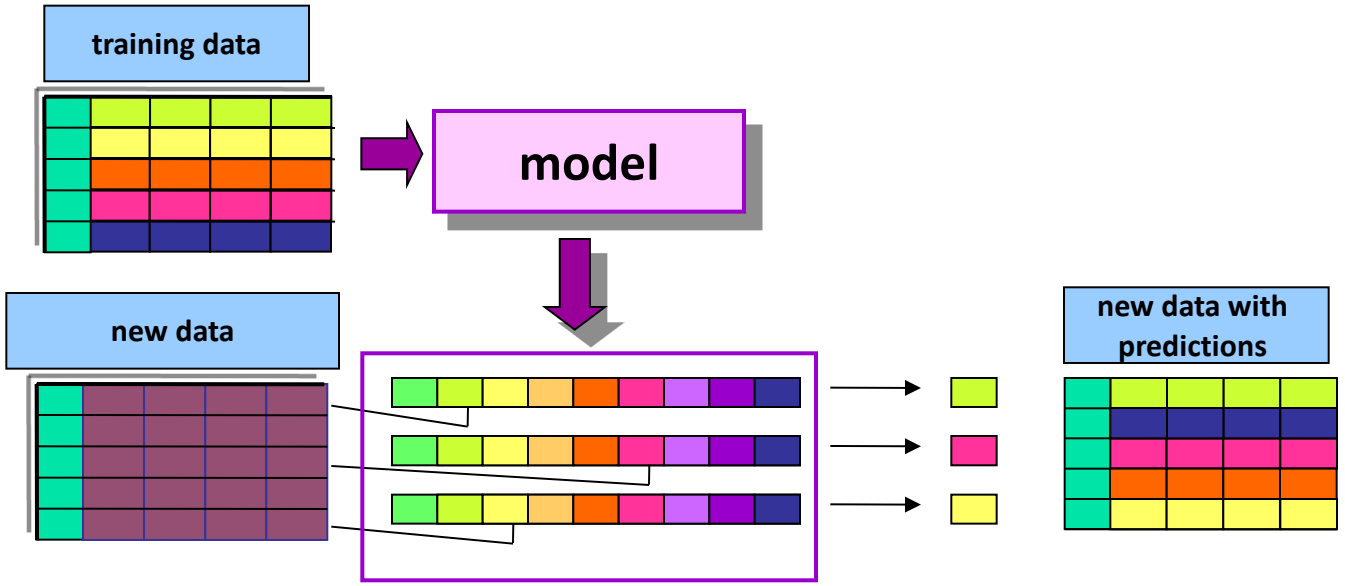
Introduction to the regression analysis

- Approach discussed in this set of slides
  - **Linear regression**
  - **SVMs (SVR)**
- Other approaches
  - k-Nearest Neighbours
  - Decision trees
  - ..



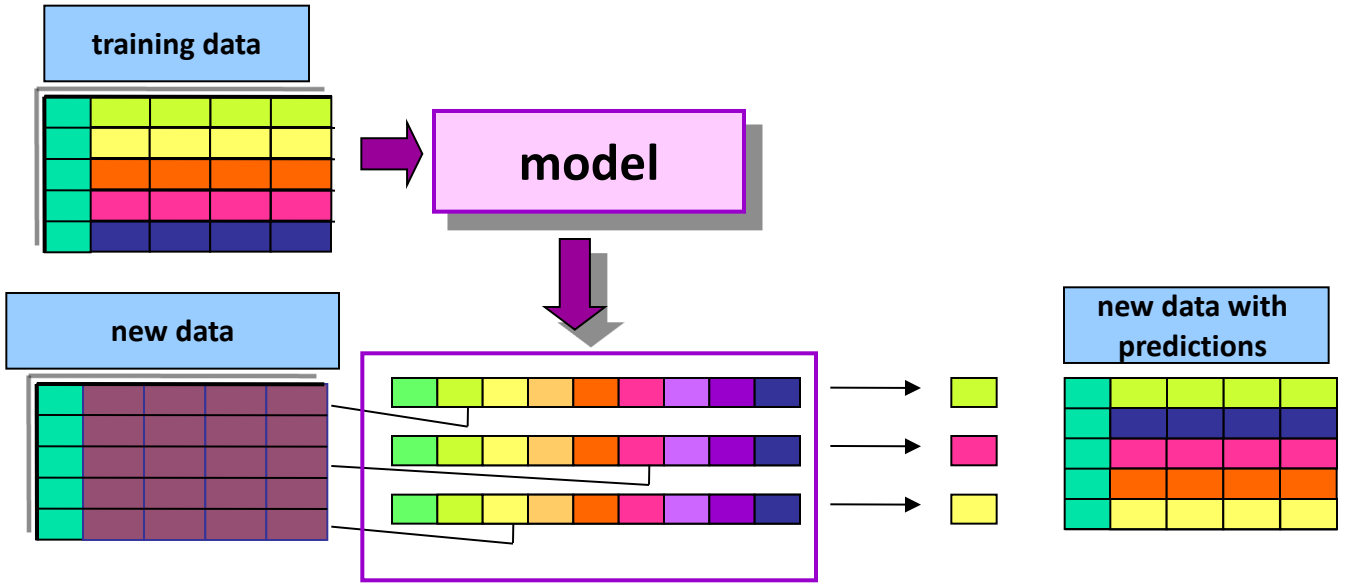
Introduction to the regression analysis

- Requirements
  - **accuracy**
  - interpretability
  - scalability
  - **noise and outlier management**



Introduction to the regression analysis

- Applications
  - Estimating the cost of a house
  - Estimating the remaining useful life (RUL) of an industrial equipment
  - Industrial Vehicle Usage Predictions
  - Predicting the Number of Free Floating Car Sharing Vehicles within Urban Areas
  - …



Introduction to regression

- The term "regression" was coined by **Francis Galton** in 1877 to describe a biological phenomenon
  - the heights of descendants of tall ancestors tend to regress down towards a normal average (i.e, regression toward the mean)
- Father of regression **Carl F. Gauss** (1777-1855)

Definition

- Given
  - A numerical target attribute
  - A collection of data objects also characterized by the target attribute

- The regression task finds a model that allows predicting the target variable value of new objects through
  - $y=f(x_1, x_2, \ldots x_n)$

Regression analysis

- Regression analysis can be classified based on
  - **Number of explanatory variables**
    - Simple regression: single explanatory variable
    - Multiple regression: includes any number of explanatory variables
  - **Types of relationship**
    - Linear regression: straight-line relationship
    - Non-linear: implies curved relationships (e.g., logarithmic relationships)
  - **Temporal dimension**
    - Cross Sectional: data gathered from the same time period
    - Time Series: involves data observed over equally spaced points in time

Simple linear regression

$$y = \beta_0 + \beta_1 x$$

- The regression line provides an **interpretable model** of the phenomenon under analysis
  - y: **estimated** (or predicted) **value**
  - β_0: estimation of the **regression intercept**
    - The intercept represents the estimated value of y when x assumes 0
  - β_1: estimation of the **regression slope**
  - x: **independent variable**

Simple linear regression

$$y = \beta_0 + \beta_1 x$$

- *Least squares method*
  - β_0 and β_1 can be obtained by **minimizing the Residual sum of squares (RSS)** that is the sum of the squared residuals
    - differences between actual values (y) and estimated ones $(\hat{y})$

$$min\, RSS = min \sum_i (y_i - \hat{y}_i)^2 =$$

$$min \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2$$

Estimation of the parameters by least squares

$$y = \beta_0 + \beta_1 x$$

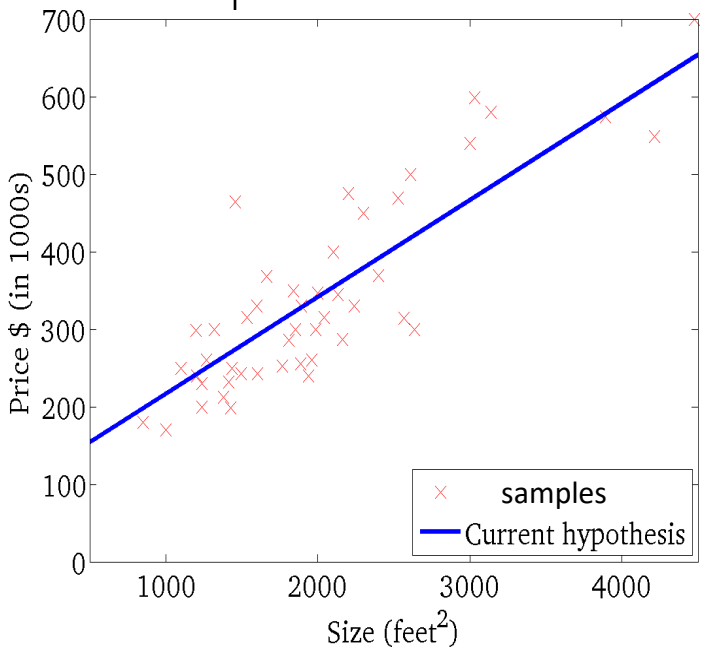
$$\beta_1 = \frac{\Sigma_i (x_i - \bar{x})(y_i - \bar{y})}{\Sigma_i (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

- where means $\bar{y} = \frac{1}{n}\sum_i y_i$ and $\bar{x} = \frac{1}{n}\sum_i x_i$ are the sample

Simple linear regression: example

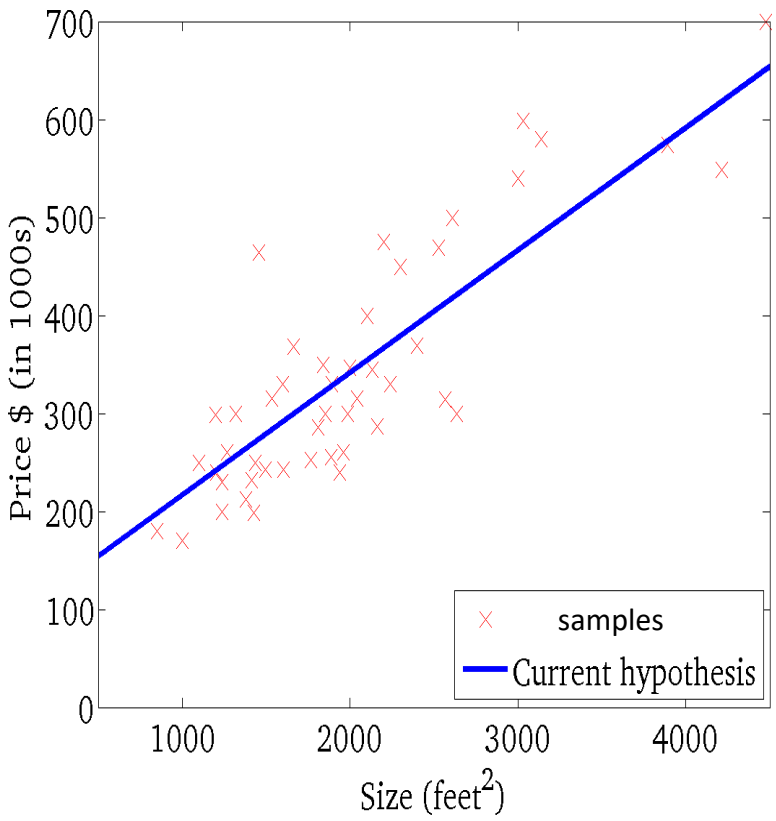
Size in feet^2	Price ($) in 1000's
2104	460
1416	232
1534	315
852	178
…	…



- Goal of a **real estate agency**
  - Estimate the selling price of a home based on the value of size in square feet
- Simple linear regression finds **a linear model** of the problem
  - x = Size in feet^2
  - y = Price ($) in 1000's

$$y = \beta_0 + \beta_1 x$$

Simple linear regression: example



- β_0: The **intercept** represents the estimated value of y when x assumes 0
  - No house had 0 square feet, but β_0 is the portion of house price not explained by square feet
- β_1: the **slope** measures the estimated change in the y value as for every one-unit change in x
  - The average value of a square foot of size

Multiple linear regression

$$y = f(\boldsymbol{x}) = \beta_0 + \beta_1 \boldsymbol{x_1} + \beta_2 \boldsymbol{x_2} + \beta_3 \boldsymbol{x_3} + \ldots + \beta_n \boldsymbol{x_n} + \boldsymbol{\xi}$$

- **Dependant variable** (y): the single variable being explained/predicted by the regression model
- **Independent or explanatory variables** (x_i): The variables used to predict/explain the dependant variable
- **Coefficients** (β_i): values, computed by the regression task, reflecting explanatory to dependent variable relationships
- **Residuals** $(\boldsymbol{\xi})$: the portion of the dependent variable that is not explained by the model
  - The model performs under or over predictions

Interpretating regression coefficients

- Uncorrelated predictors
  - Each coefficient can be estimated and tested separately
  - Interpretation: a unit change in x_i is associated with a β_i change in y, white all the other variables stay fixed
    - β_i represents the average effect on y of a one unit increase in x_i, holding all other predictors fixed
- Correlation among predictors cause problems
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become complex: when x_j changes, everything else changes
- The claim of causality should be avoided for the observational data

Feature selection

- In case of a high dimensional data set, in terms of number of dependent variables, **some of the variables** might provide **redundant information**.
- Feature selection and removal (correlation-based approach)
  - simplifying the model computation
  - improving the model performance
  - Enhancing the model interpretation (i.e., better explainability of the dependent variables)
- Variable/feature selection
  - Driven by the business understanding and domain knowledge
  - Feature selection based on correlation test
    - Features highly-correlated with other attributes could be discarded from the analysis
    - having dependence or association in any statistical relationship, whether causal or not

Polynomial regression

- The polynomial models can be used in those situations where the **relationship** between dependent and explanatory variables is **curvilinear**.
- Polynomial regression consists of**:**
  - Computing new **features** that are power functions of the input features
  - Applying **linear** regression on these new features

$$y = \beta + \beta_1 x + \beta_2 x^2 + \varepsilon$$

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

  - The above models are also linear (i.e., A model is linear when it is linear in parameters)
  - They are the second order polynomials in one and two variables respectively.
- Sometimes a nonlinear relationship in a small range of explanatory variables can also be modeled by polynomials.

Polynomial model in one variable

- The k[th] order polynomial model in one variable is given by

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k + \varepsilon$$

- It is included in the linear regression model below

$$y = X\beta + \varepsilon$$

- Techniques for fitting linear model can be used for fitting the polynomial regression model
- For example, $y = \beta_0 + \beta_1 x + \beta_2 x^2$
  - Is a polynomial regression model in one variable and is called as **second order model** or **quadratic model**, where the coefficients
    - β_1 is the linear effect parameter
    - β_2 is the quadratic effect parameter
- The polynomial models can be used to approximate a complex nonlinear relationship

Polynomial regression: considerations in case of one variable

- Order of the model
  - Keep the order of the polynomial model as low as possible
    - Up to the **second order** polynomial
    - If necessary, you should apply some **data transformations**
  - Arbitrary fitting of higher order polynomials can be a serious abuse of regression analysis.
    - Data overfitting issue
- Different model building strategies do not necessarily lead to the same model
  - **Forward selection procedure**: to successively fit the models in increasing order and test the significance of regression coefficients at each step of model fitting.
    - Keep the order increasing until t-test for the highest order term is nonsignificant
    - The significance of highest order term is tested through the null hypothesis
  - **Backward elimination:** to fit the appropriate highest order model and then delete terms one at a time starting with highest order. This is continued until the highest order remaining term has a significant t-test
- The first and second order polynomials are mostly used in practice.

Polynomial models in two or more variables

- The techniques of fitting of polynomial model in one variable can be extended to fitting of polynomial models in two or more variables.
- A second order polynomial is more used in practice and its model is specified by

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

- This is also called **response surface**.

Strong and weak points of Polynomial Regression

- Advantages of using Polynomial Regression:
  - **Broad range of function** can be fit under it.
  - Polynomial basically fits **wide range of curvature**.
  - Polynomial usually provides the **best approximation** of the relationship between dependent and independent variable.
- Disadvantages of using Polynomial Regression
  - They are **too sensitive to the outliers**.
    - The presence of a few outliers in the data can seriously affect the results of a nonlinear analysis.
  - Higher polynomial degree means **higher flexibility** of your model, but also **data overfitting**
    - Overfitting occurs in those cases when you have a few samples and a model that has high flexibility
    - It is always possible for a polynomial of order $(n-1)$ to pass through n points so that a polynomial of sufficiently high degree can always be found that provides a "good" fit to the data.
    - Those models **never enhance the understanding** of the unknown function and they are **never good predictors**.

To avoid data overfitting

- Use more training data (if possible)
- Use lower model complexity
- Use regularization techniques
  - e.g., Ridge and Lasso

RIDGE and LASSO

- Regression analysis methods that perform both **variable selection** and **regularization** in order to enhance the prediction **accuracy** and **interpretability** of the statistical model it produces.
- Useful **to reduce model complexity** and **prevent overfitting** when
  - The number of variables describing each observation exceeds the number of observations
  - The number of variables does not exceed the number of observations, but the learned model suffers from poor generalization.
- Techniques of training a linear regression (or a linear regression with polynomial features)
  - They try to assign values **closer to zero (RIDGE) or zero (LASSO)** to the coefficients assigned to features that are not useful for the regression
  - The effect is the **decreasing of the complexity of the model**

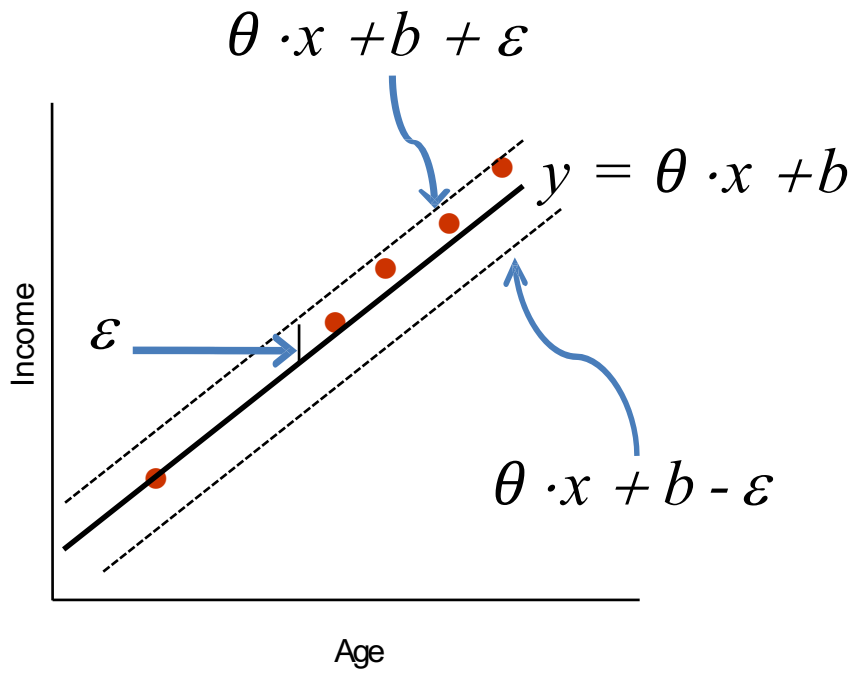
LASSO

- LASSO means **Least Absolute Shrinkage and Selection Operator**
- Term coined by Robert Tibshirani in 1996, but it was originally introduced in geophysics literature 10 years before
- Lasso **regularization** was originally defined for **least squares**, but it is easily extended to a wide variety of statistical models in a straightforward fashion
  - E.g., generalized linear models
- The Lasso's **variable selection** relies on the form of the **constraint**
  - It forces the sum of the absolute value of the regression coefficients to be less than a fixed constraint, which forces some coefficients to be set to zero
  - The selected model is simpler since it does not include coefficients set to zero.
- It is similar to RIDGE regression but usually identifier a simpler model
  - **RIDGE** simplifies the model **by shrinking the size of some coefficients**, while **LASSO sets some coefficients to zero**.

Support Vector Machine – Regression

- Find a function, $f(x)$, that performs a prediction of the target attribute y with a maximum error equal to ε

We do not care about errors as long as they are less than ε



Support Vector Regression: linear model

- The (training) problem can be formulated as a convex optimization problem

$$\min \frac{1}{2} \| \theta \|^2$$

$$s.t. \quad y^i - \theta \cdot x^i - b \leq \varepsilon;$$
$$\theta \cdot x^i + b - y^i \leq \varepsilon$$

y^i = value of the target attribute of the i^{th} training object

x^i = value of the predictive attributes of the i^{th} training object

θ and b = parameter of the regression model

Support Vector Regression: Soft margin

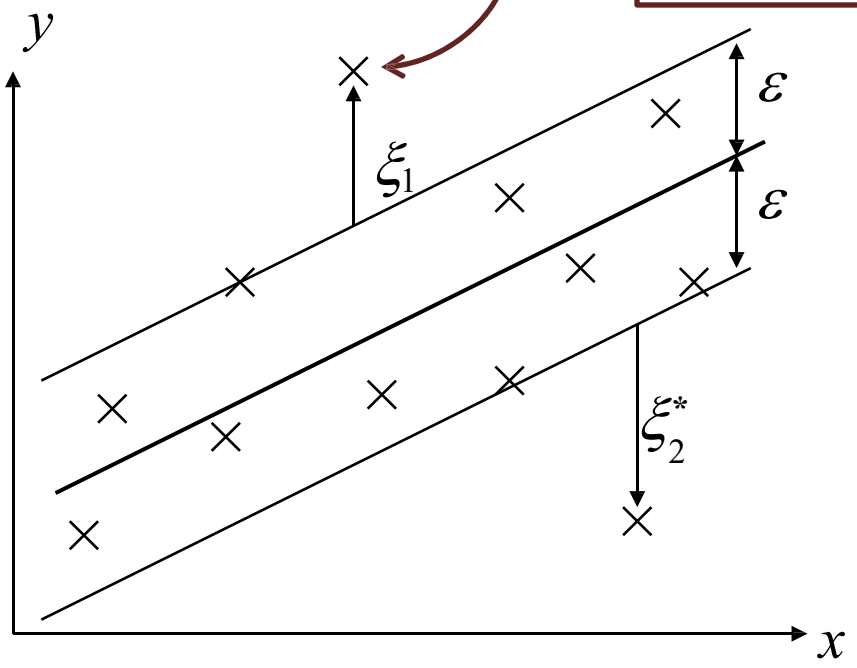
- Given a specific value of ε, the problem is not always feasible
- **Soft margin**
  - Reformulate the problem by considering the errors related to the predictions that do not satisfy the ε **maximum distance**

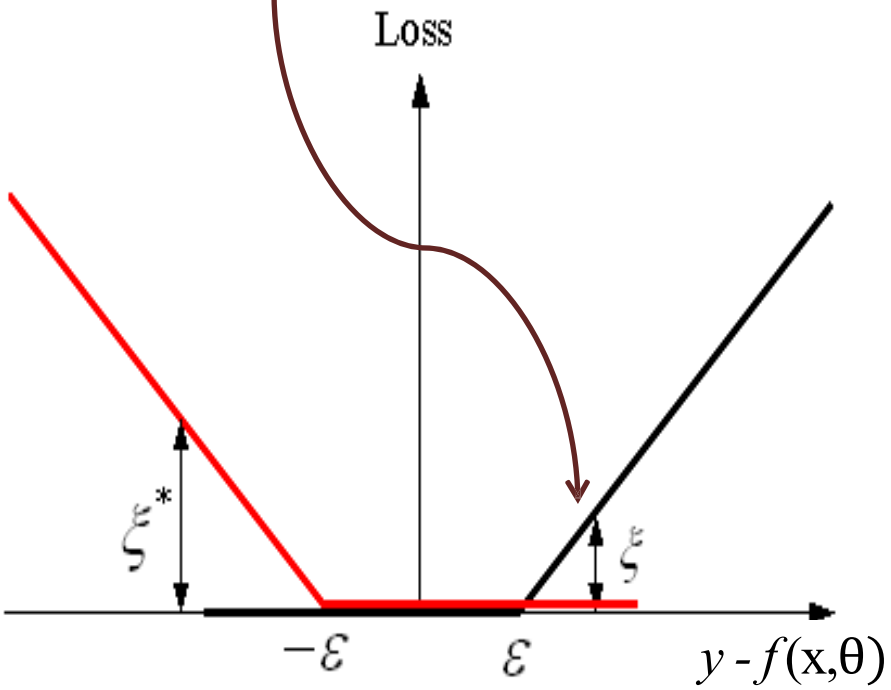
Support Vector Regression: Soft margin

Assume linear model

$$f(\mathrm{x},\theta) = \theta \cdot \mathrm{x} + b$$

Only the point outside the ε-region contribute to the final cost





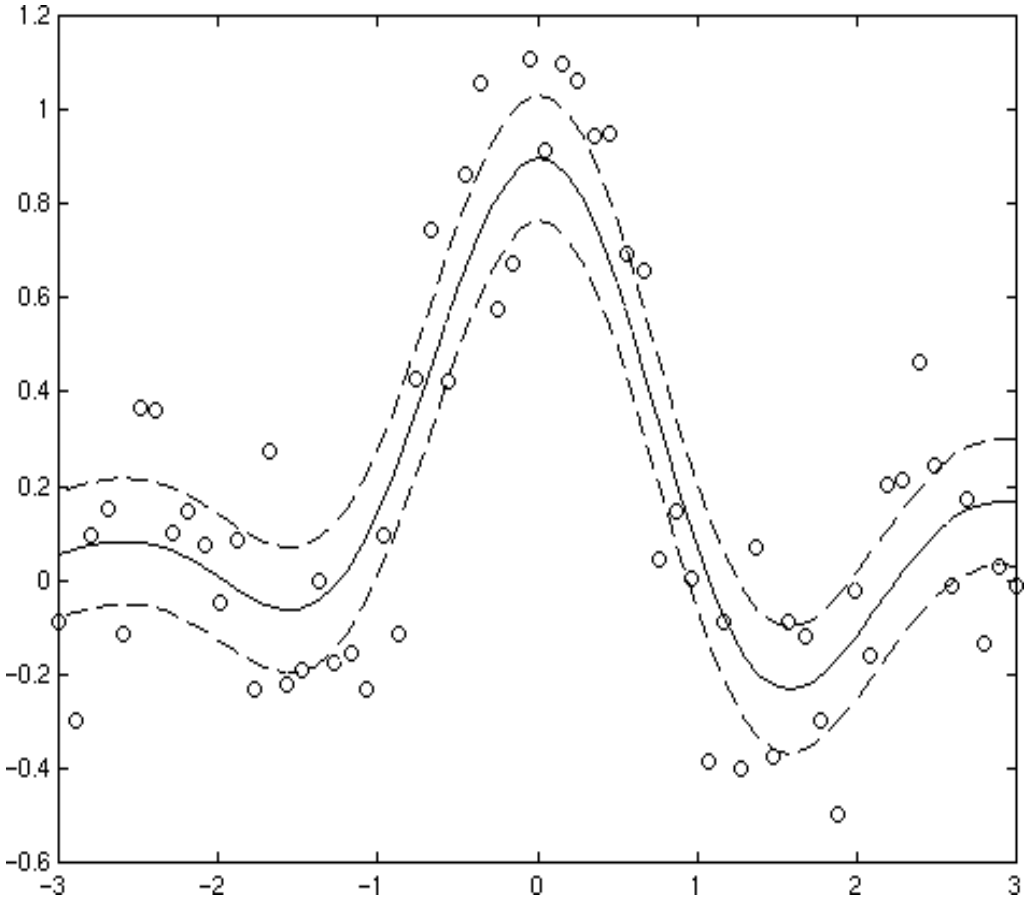
Support Vector Regression: Soft margin

- The (training) problem can be formulated as a convex optimization problem

$$\min \frac{1}{2} \| \theta \|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*)$$

$$\begin{aligned} s.t. \quad & y^i - \theta \cdot x^i - b \leq \varepsilon + \xi_i; \\ & \theta \cdot x^i + b - y^i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, i = 1, \ldots, m \end{aligned}$$

How about a non-linear case?



Linear versus Non-linear SVR

- Map the original features into a higher order dimensional space
- Apply a kernel transformation
  - Polynomial
  - Gaussian radial
  - …
- Transform the input data by means of the kernel function φ and then solve the previous problem

Linear versus Non-linear SVR

- φ maps the input data into a new dimensional space

$$\min \frac{1}{2} \| \theta \|^2 + C \sum_{i=1}^{m} (\xi_i + \xi_i^*)$$

$$\begin{aligned} s.t. \quad & y^i - \theta \cdot \phi(x^i) - b \leq \varepsilon + \xi_i; \\ & \theta \cdot \phi(x^i) + b - y^i \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, i = 1, \ldots, m \end{aligned}$$

Evaluating regression

- Evaluation metrics for regression:
  - MAE (Mean Absolute Error)
  - MSE (Mean Squared Error)
  - RSE: Residual Standard Error
  - R^2
  - Adjusted R^2

- The evaluation is performed by comparing
  - y: the actual value (**ground truth**)
  - $\hat{y}$: the predicted value through the regression model

Evaluating regression

- MAE (Mean Absolute Error)
  - the average vertical distance between each real value and the predicted one

$$MAE = \frac{1}{n}\sum_i |y_i - \hat{y}_i|$$

- MSE (Mean Squared Error)
  - the average of the squares of the errors
  - the average squared difference between the estimated values and the actual value.
  - MSE tends to penalize less errors close to 0

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

- MAE and MSE always > 0
  - The lower the values of MAE and MSE the better the model
  - It is mainly affected by the domains of data sample

Evaluating regression

- Overall accuracy of the model
  - RSE: Residual Standard Error

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

    - n is the number of samples
    - RSS is the residual sum of squares
  - RSE is always greater than 0
    - The lower the RSE value the better the regression model

Evaluating regression

- R^2: R-squared measures the goodness of fit of a model
  - how well the regression predictions approximate the real data points.
  - It estimates a normalized error

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

  - RSS is the residual sum of squares $RSS = \sum_i (y_i - \hat{y}_i)^2$
  - TSS is the total sum of squares $TSS = \sum_i (y_i - \bar{y}_i)^2$ with $\bar{y} = \frac{1}{n}\sum_i y_i$

Evaluating regression: R^2

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - FVU$$

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

- R^2 represents the proportion of variance of y explained by variation in x
  - FVU means the fraction of variance unexplained
    - Ratio between the unexplained variance (variance of the model's errors) and the total variance

Evaluating regression: R^2

- R^2 value
  - $R^2 = 1$
    - A perfect linear relationship between x and y
    - 100% of the Y variation is explained by variation in x
  - R^2 close to 1
    - A very good linear relationship between x and y
    - Good predictions
  - $0 < R^2 << 1$
    - Weaker linear relationship between x and y
    - A portion of the variation in y is not explained by variation in x
  - $R^2 = 0$
    - No linear relationship between x and y
    - The value of y does not depend on the value of x

Evaluating regression: R^2 adjusted

- Drawback of R^2
  - In the context of multiple linear regression, if new predictors (X_i) are added to the model, R^2 only increases or remains constant but it never decreases.
  - However, it is not always true that by increasing the complexity of regression model, the latter will be more accurate
- The Adjusted R-Squared is the modified form of R-Squared that has been adjusted to incorporate model's degree of freedom.
- It should be used to evaluate the quality of a multiple linear regression model

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

  - p = number of explanatory variables
  - n = number of samples

- The adjusted R-Squared only increases if the new term improves the model accuracy.