SVR, Trees & forest

RESTORY...

Agenda

- SVR
- Decision Trees
- Random Forest
- Statistik
  - p-value
  - Multivariable regression - variable selection
  - R_2

Support Vector Regression

- Använder sig av ett godtagbart felintervall
- Bygger på teori för support vector machines (SVM)
- Algoritmen finns i olika grad av noggrannhet
- Denna noggrannhet bestäms av en kärna (kernel)
- Den enklaste versionen är linjär SVR

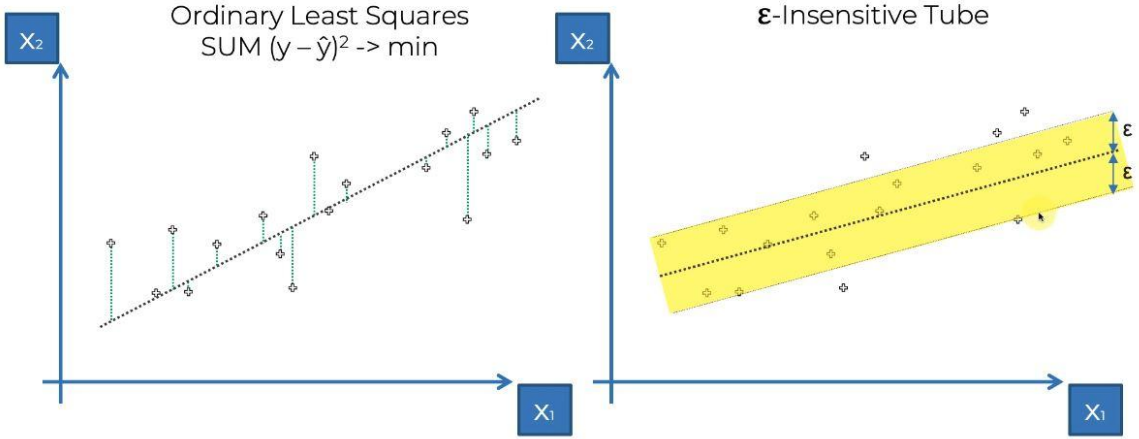


Image source: Udemy course.

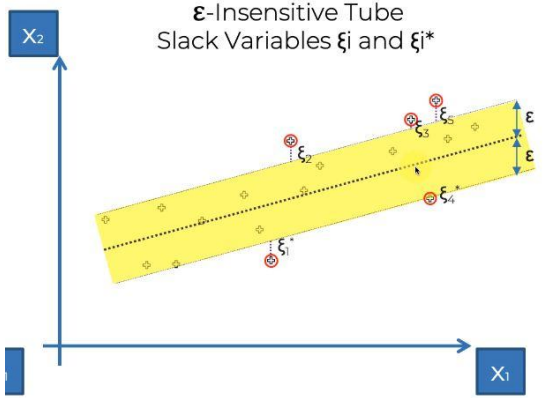
Support Vector Regression

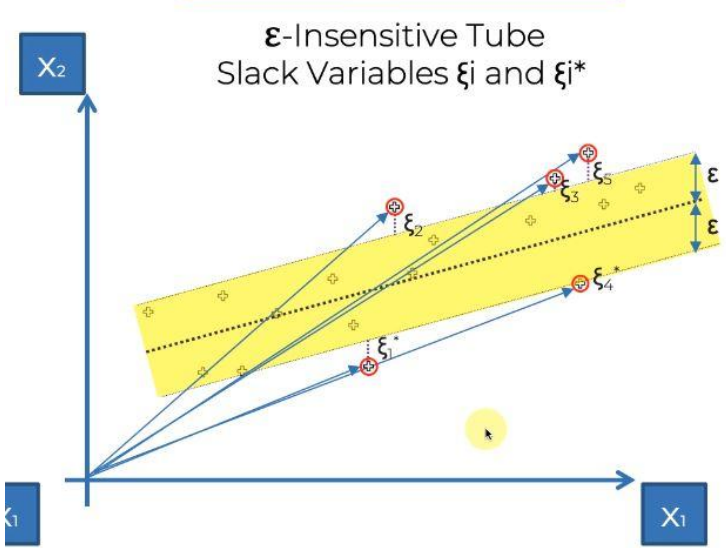
Endast punkter utanför ’röret’ bidrar till fel som ska minimeras.

ξ - över

ξ^* - under

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





Varför heter det support vectors?

- Punkterna utanför “tuben” bestämmer dess

Image source: Udemy course.

SVR - application

In general, you can use SVR to solve the same problems you would use linear regression for.

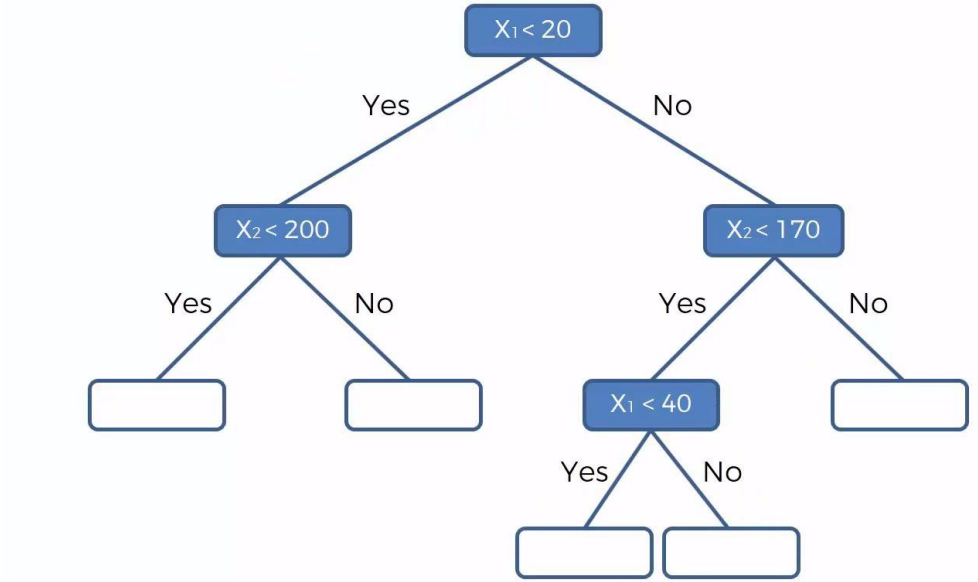
Unlike linear regression, though, SVR also allows you to model non-linear relationships between variables and

provides the flexibility to adjust the model's robustness by tuning hyperparameters.

Source: TDS

Decision Trees

Decision Trees - Structure

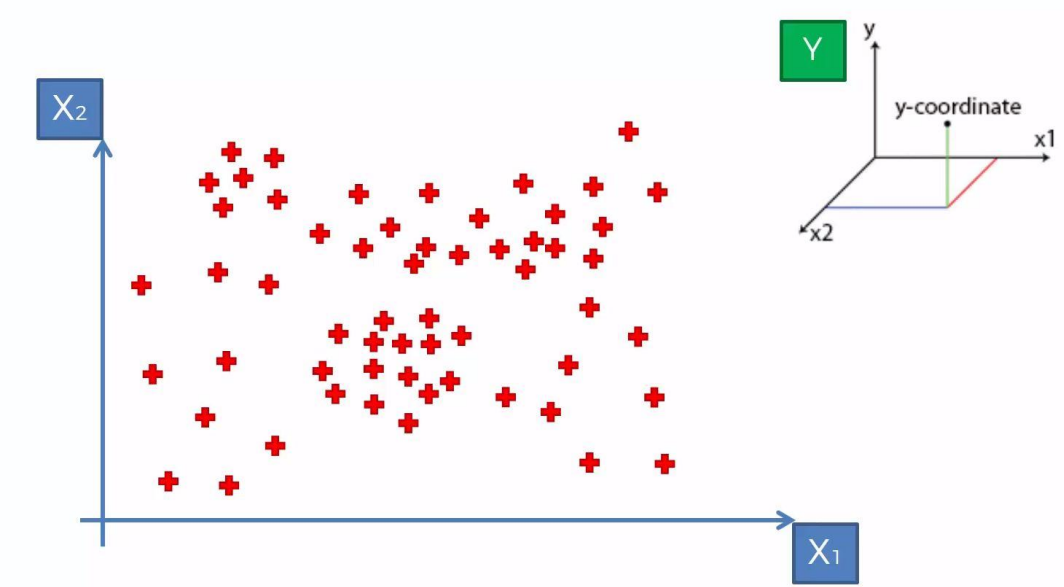


Decision Trees

CART (Classification and regression trees)

x_1 och x_2 är oberoende

y är in en tredje dimension

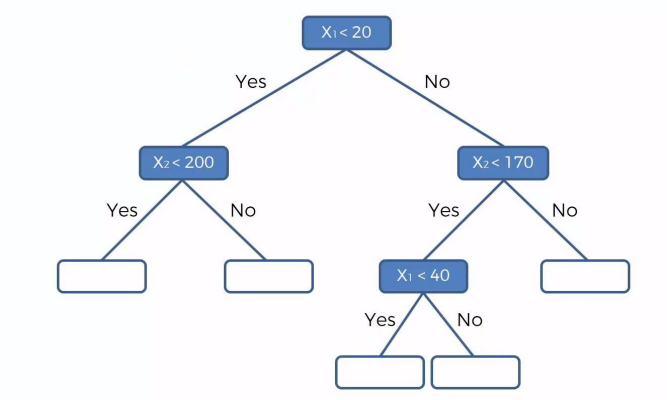


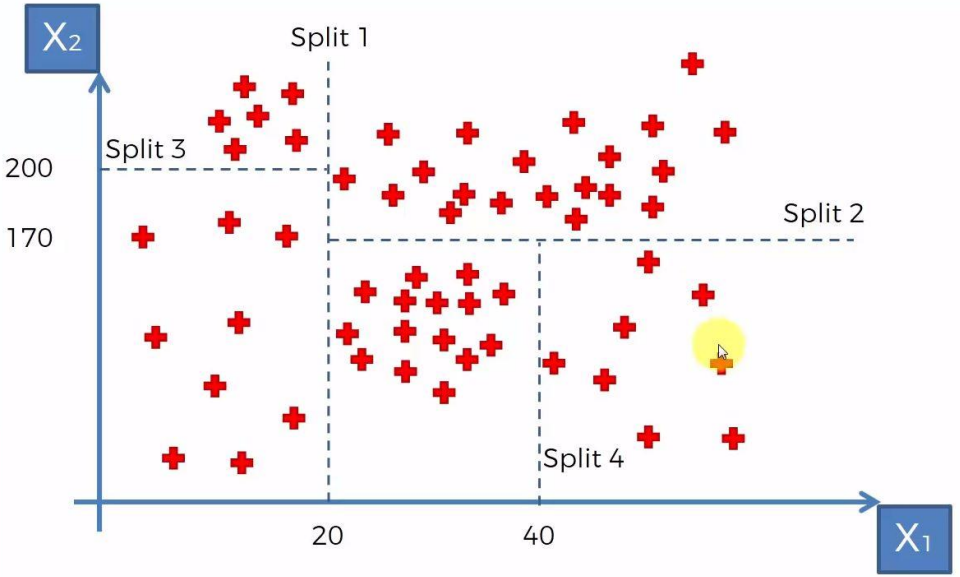
Decision Trees

Steg för att dela upp intervallet

“Information entropy” - lägger en ny uppdelning mer information om punkterna?

Annan stoppunkt - mindre än 5% av alla datapunkter i en grupp.

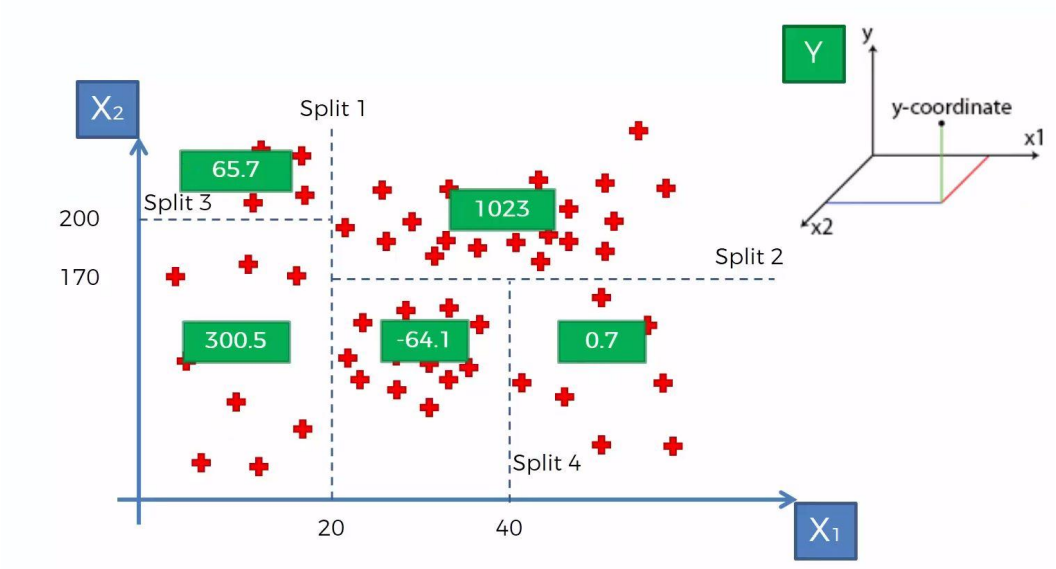




RESTORY...

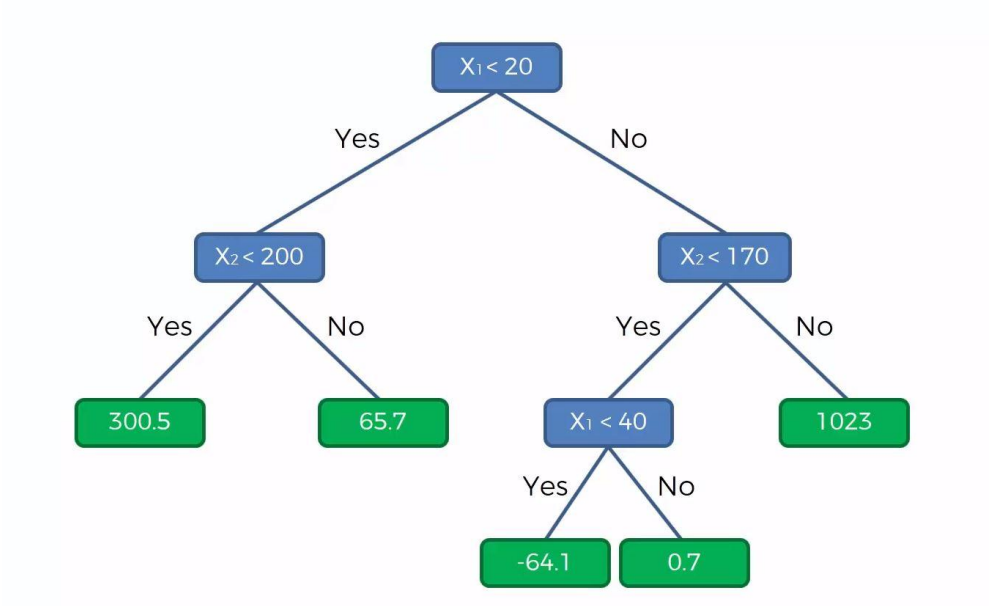
Decision Trees - Leaves

- Modellens värde i ett ’löv’(leaf)
- anges av medelvärdet av y



Decision Trees - Leaves

Motsvarande beslutsträd



Random forest - Ensemble learning

- Välj ut ett slumpmässigt antal datapunkter K från träningsdata
- Bygg ett beslutsträd på dessa K datapunkter
- Välj ett antal träd Ntree du vill bygga och upprepa ovanstående steg - flera träd
- För en ny datapunkt som du vill hitta ett modellvärde till, ta medelvärdet av de Y värden alla Ntree träden producerat.
- N = 500 minst

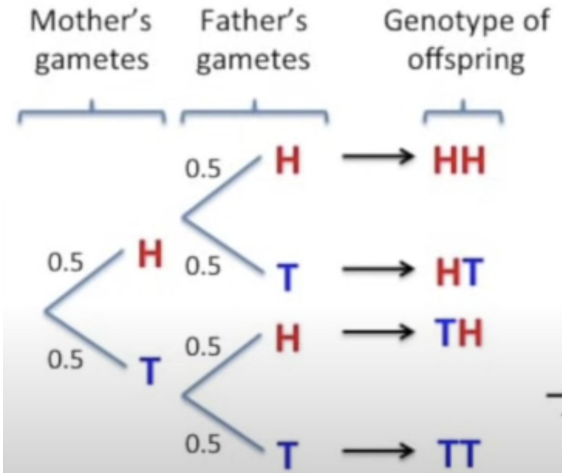
- Ensemble learning - en algoritm används flera gånger
- Random forest är en typ av ensemble learning
- Stabilare och mer kraftfull

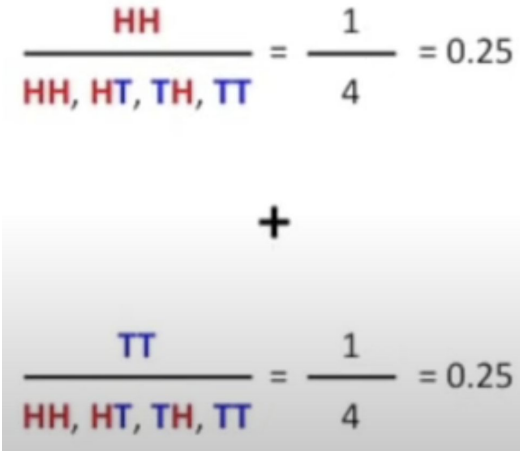
Statistik-ish

p-value

- Statistical significance
- Null hypothesis H_0 - assume true
- Is it correct? - experiment
  - 0.50
  - halveras ...
  - $\alpha = 0.05$
  - domain dependent

" A p-value is the probability that random chance generated the data, or something else that is equal or rarer"





p-value ≠ probability

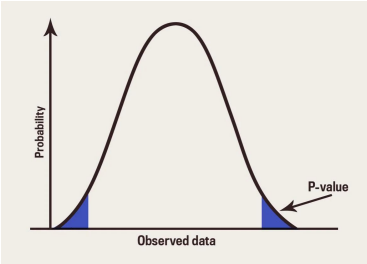


Bild ref.

StatQuest Youtube vid

RESTORY...

Multivariable regression - variable selection

Why select?

- GIGO
- Noise
- Explain

1. **All** Variables
   a. Domain knowledge - predictors
   b. Forced - by company
   c. Preparing for no. 2

2. **Backward** Elimination
   a. Set significance level (e.g. sl = 0.05)
   b. Fit full model
   c. Remove predictor with highest **p**-val if **p>sl**.
   d. Refit model
   e. Repeat c & d until no p > sl.

3. **Forward** Selection
   a. Select sl
   b. Fit all simple reg. models for y and X for all x separately.
   c. Select the one with the lowest p-val ($\mathbf{x_n}$).
   d. Fit $\mathbf{x_n}$ with all other x variables (X-$\mathbf{x_n}$) separately for every model. (y and [x_a, x_b])
   e. If the model with the lowest p-val has a p<sl, add it to the predictor, and repeat d.

Udemy kurs, Lek. 54.

Multivariable regression - variable selection

Stepwise regression: 2,3,**4**

4. ***Bidirectional*** *Elimination*
   a. Select an entry and stay sl (can be different)
   b. Add on new variable using Forward Selection (p < sl_enter).
   c. Do all steps of Backward elimination. Keep only variables where p < sl_stay is fulfilled.
   d. Repeat b and c until no new variables can enter or exit.
5. Score Comparison
   a. Set a fit score minimum
   b. Create a model for all possible variable combinations (2^n-1)
   c. Select the one with the best score

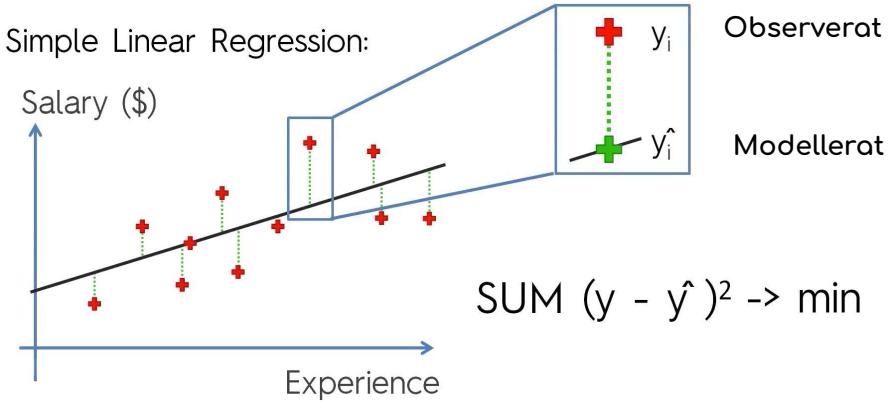
Scikit-learn does this automatically!

Udemy kurs, Lek. 54.

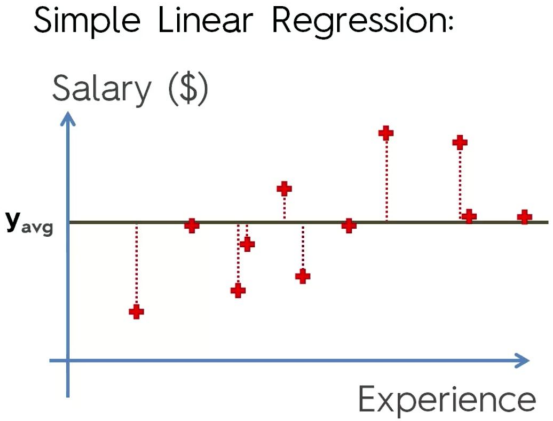
R^2

R^2

How much better than average?



$$SUM\ (y - \hat{y})^2 \rightarrow min$$



$$SS_{res} = SUM\ (y_i - \hat{y}_i)^2$$

$$SS_{tot} = SUM\ (y_i - y_{avg})^2$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Udemy kurs, Lek.99.

R^2 och flera variabler

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

R^2 – Goodness of fit
(greater is better)

$y = b_0 + b_1 * x_1$

Problem:

$y = b_0 + b_1 * x_1 + b_2 * x_2$ ← $+ b_3 * x_3$

SS_{res} -> Min → R^2 will never decrease

Udemy kurs, Lek. 99.

Justerad R^2

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

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

p - number of regressors

n - sample size

Udemy kurs, Lek. 100.

Länkar

- SVR and SVM
- SVR - TDS
- SVR kapitel från bok
- SVM applications
- Decision Trees applications irl
- Random forest vs DT
- Random forest applications irl
- Komplexitet och modeller