Data Mining with Weka

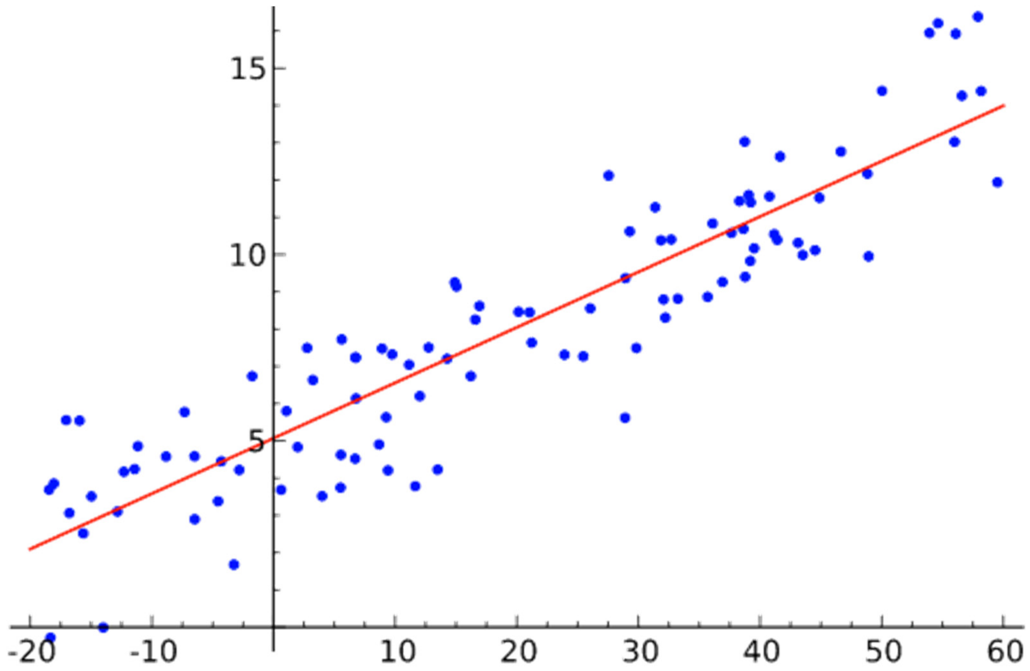
Linear regression

Ian H. Witten

Linear regression

Numeric prediction (called "regression")

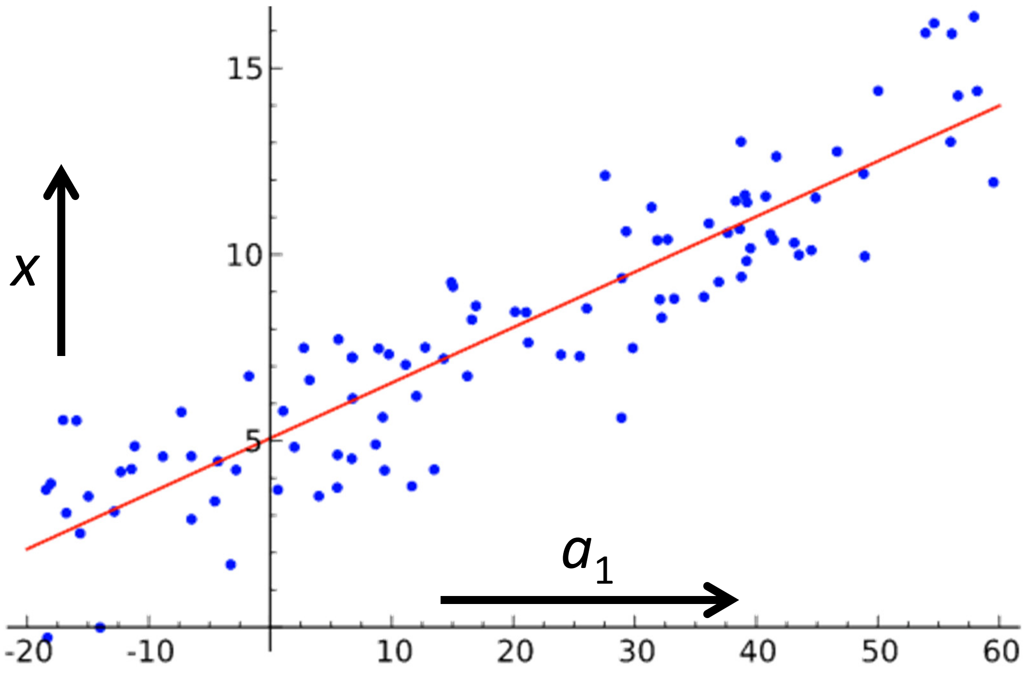
- Data sets so far: nominal and numeric attributes, but only nominal classes
- Now: numeric classes
- Classical statistical method (from 1805!)



Linear regression

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k$$

(Works most naturally
with numeric attributes)



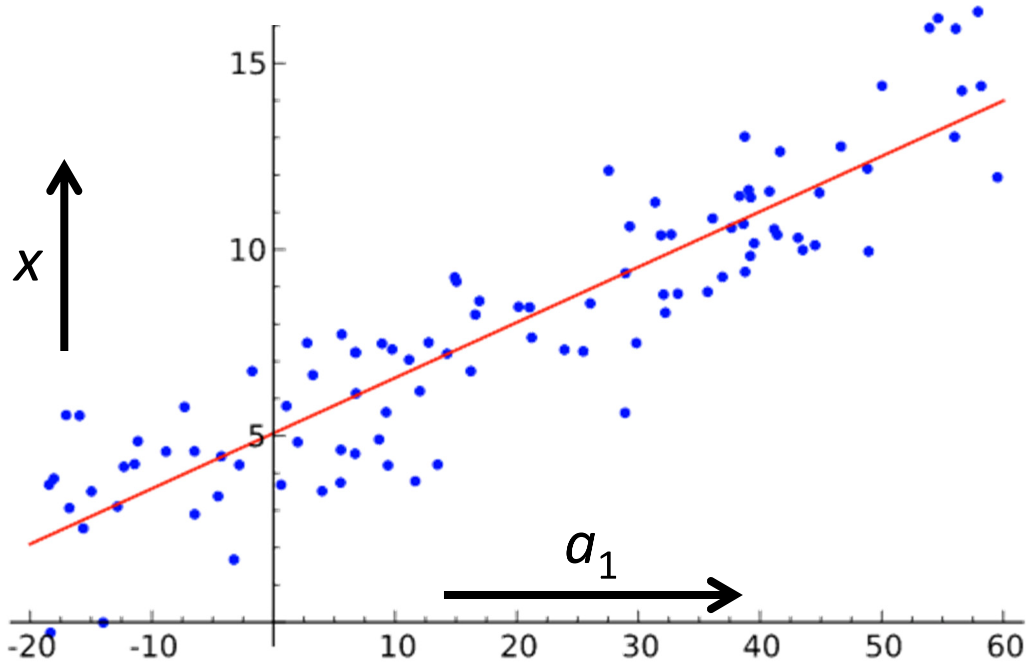
Linear regression

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k$$

- Calculate weights from training data

- Predicted value for first training instance a(1)

$$w_0 a_0^{(1)} + w_1 a_1^{(1)} + w_2 a_2^{(1)} + \ldots + w_k a_k^{(1)} = \sum_{j=0}^{k} w_j a_j^{(1)}$$



Linear regression

$$x = w_0 + w_1 a_1 + w_2 a_2 + \ldots + w_k a_k$$

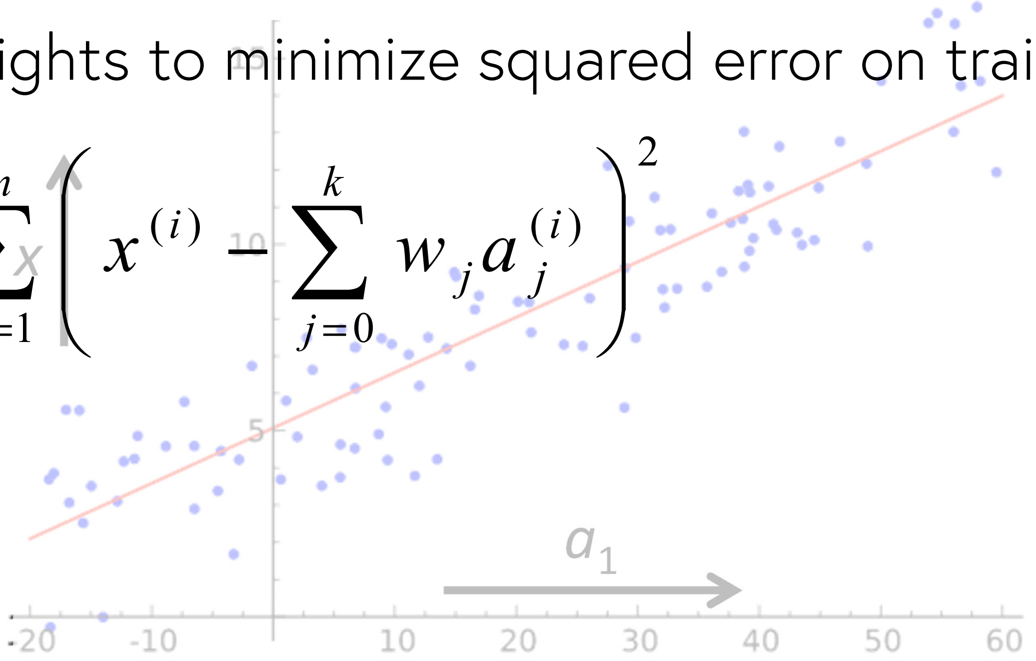
- Calculate weights from training data

- Predicted value for first training instance $a^{(1)}$

$$w_0 a_0^{(1)} + w_1 a_1^{(1)} + w_2 a_2^{(1)} + \ldots + w_k a_k^{(1)} = \sum_{j=0}^{k} w_j a_j^{(1)}$$

- Choose weights to minimize squared error on training data

$$\sum_{i=1}^{n} \left(x^{(i)} - \sum_{j=0}^{k} w_j a_j^{(i)} \right)^2$$



Linear regression

- Standard matrix problem
  - *Works if there are more instances than attributes*
    roughly speaking
- Nominal attributes
  - *two-valued: just convert to 0 and 1*
  - *multi-valued … will see in end-of-lesson quiz*

Linear regression

- Open file cpu.arff: all numeric attributes and classes
- Choose functions>LinearRegression
- Run it
- Output:
  - *Correlation coefficient*
  - *Mean absolute error*
  - *Root mean squared error*
  - *Relative absolute error*
  - *Root relative squared error*
- Examine model

$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{n}$$

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{n}}$$

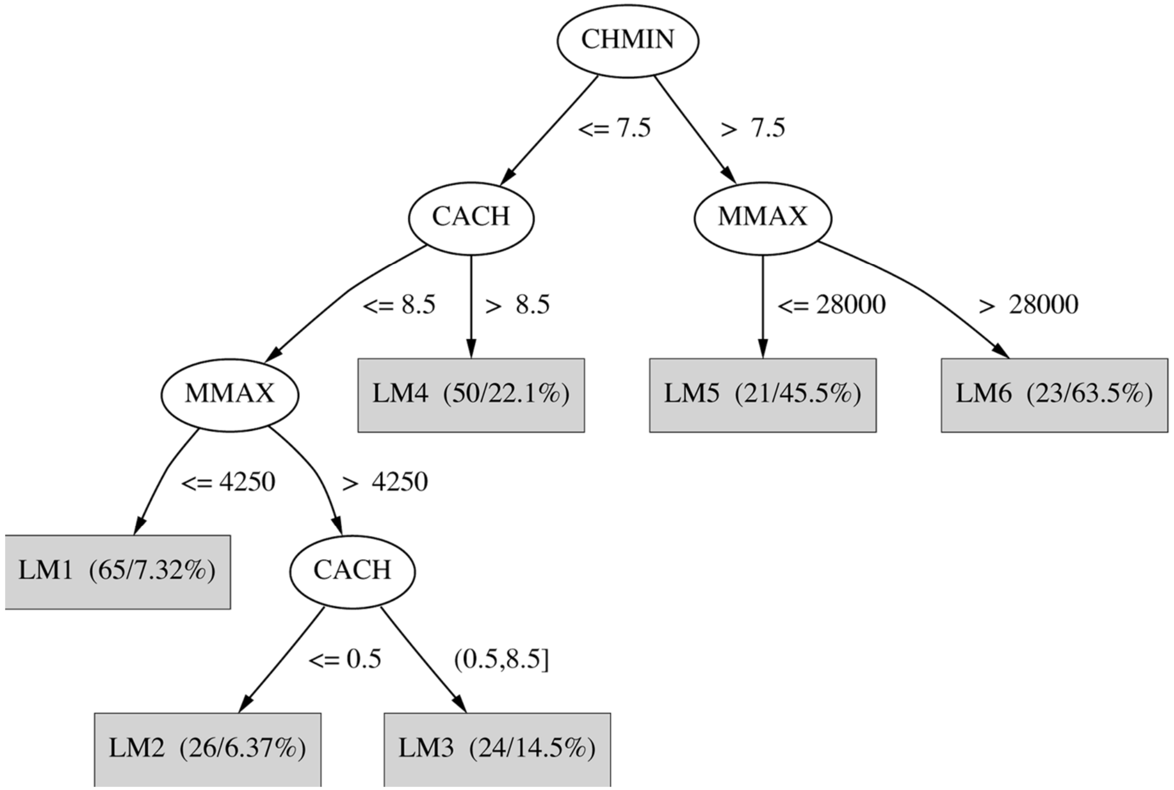
$$\frac{|p_1 - a_1| + \ldots + |p_n - a_n|}{|a_1 - \bar{a}| + \ldots + |a_n - \bar{a}|}$$

$$\sqrt{\frac{(p_1 - a_1)^2 + \ldots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \ldots + (a_n - \bar{a})^2}}$$

NON Linear regression

Model tree

- Each leaf has a linear regression model
- Linear patches approximate continuous function



NON *Linear regression*

- Choose trees>M5P
- Run it
- Output:
  - *Examine the linear models*
  - *Visualize the tree*

- Compare performance with the LinearRegression result: you do it!

Linear regression

- Well-founded, venerable mathematical technique: functions>LinearRegression
- Practical problems often require non-linear solutions
- trees>M5P builds trees of regression models