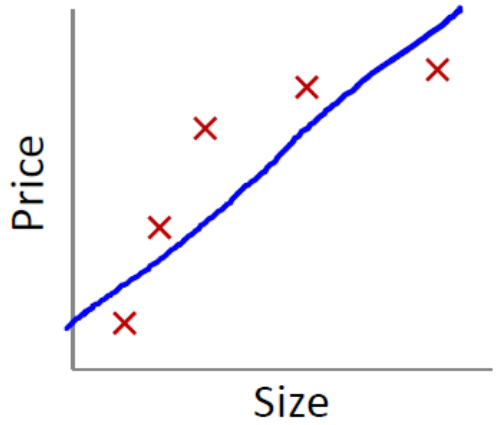
Machine Learning with R

Mohamed Rahouti

Outlines

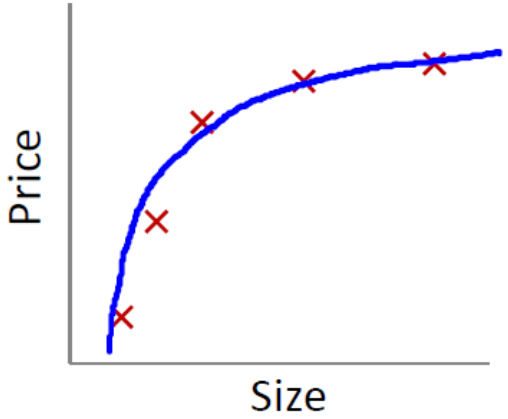
- Overfitting/underfitting problems (linear/logistic regressions)
- Decision Trees
- Unsupervised machine learning
  - K-means clustering
- H2O tool
- Conclusions

Overfitting



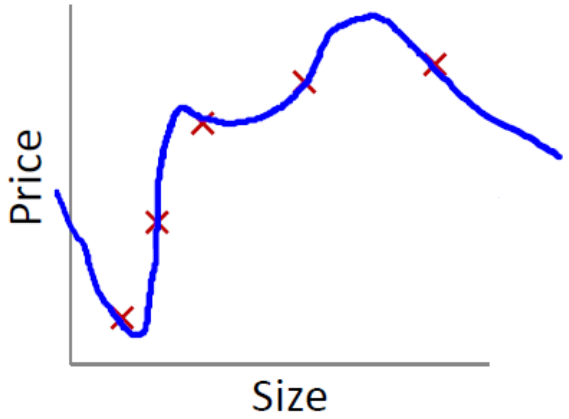
$\theta_0 + \theta_1 x$

"Underfit" "High bias"



$\theta_0 + \theta_1 x + \theta_2 x^2$

"Just right"

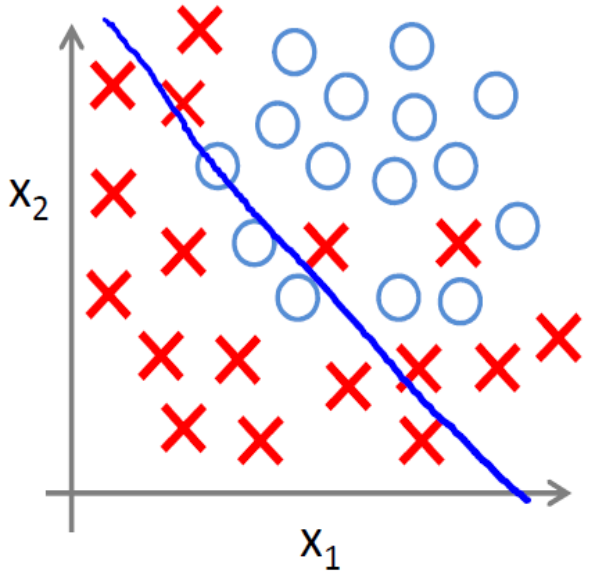


$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

"Overfit" "High variance"

Overfitting: If we have too many features, the learned hypothesis may fit the training set very well ($J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 \approx 0$), but fail to generalize to new examples (predict prices on new examples).

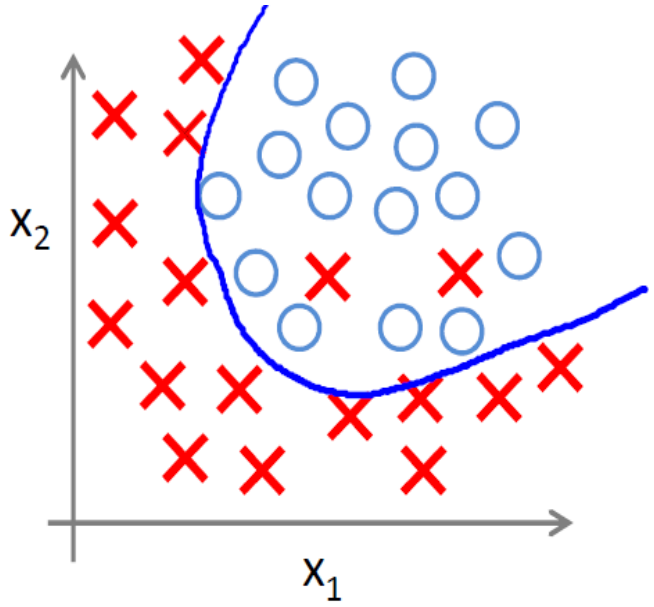
Overfitting (Cont.)



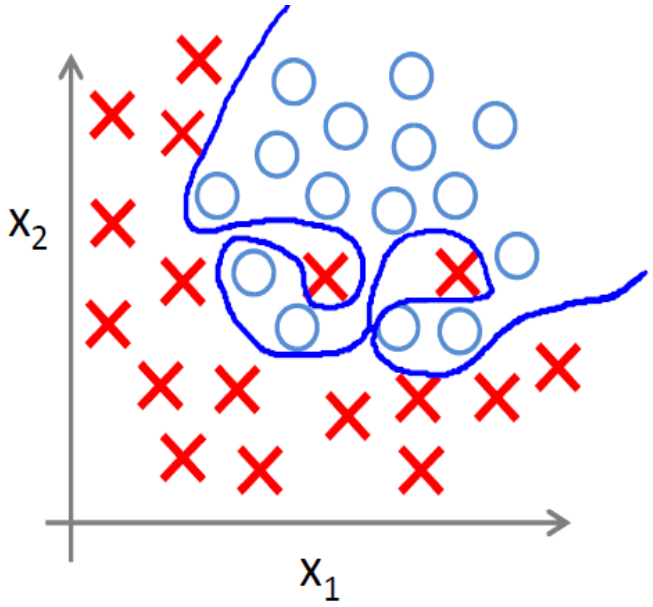
$$h_\theta(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

(g = sigmoid function)

"Underfit"



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$

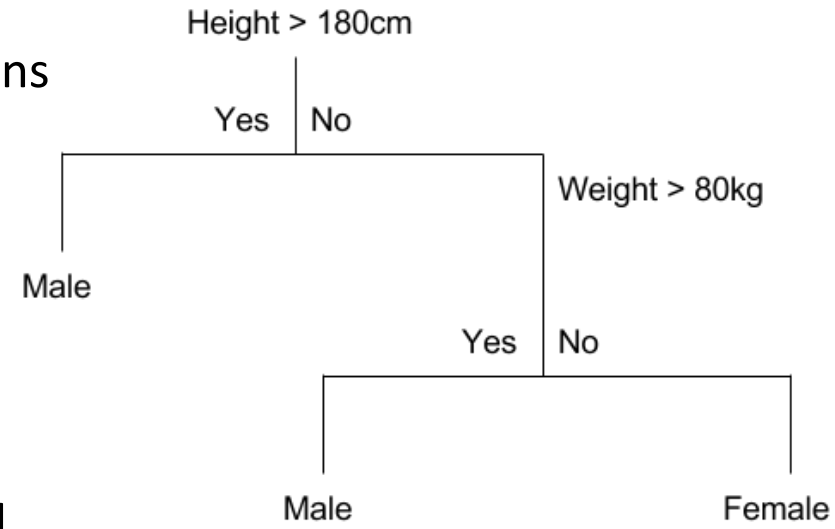
"Overfit"

ML Algorithms

- Machine learning algorithms:
  - Supervised learning
  - Unsupervised learning
- Others: Reinforcement learning, recommender systems.
- Also talk about: Practical advice for applying learning algorithms.

Decision Tree

- Graph G to represent choices & their results in form of a tree.
  - Nodes in G represent an event or choice
  - Edges in graph represent the decision rules or conditions
- Example:
  - Predicting an email as spam or not spam
  - Predicting of a tumor is cancerous
  - Predicting a loan as a good or bad credit risk based on the factors in each of these
- A model is created with observed data also called training data
- A set of validation data is used to verify and improve the model



Decision Tree in R

- R has packages which are used to create and visualize decision trees.
- For new set of predictor variable, we use this model to arrive at a decision on the category (yes/No, spam/not spam) of the data.
- The R package **"party"** is used to create decision trees.
- The basic syntax for creating a decision tree in R
  - *> ctree(formula, data)*
  - **formula** is a formula describing the predictor and response variables.
  - **data** is the name of the data set used.

Application in R

- Lab 3

Thanks for your Attention

- Questions??

Credits

- Some material was borrowed from:
  - Machine Learning with R and H2O
  - Introduction to ML with Applications in R
  - ML in R, by Alexandros Karatzoglou
  - ML course, by A. NG