LING 506 - TOPICS IN COMPUTATIONAL LINGUISTICS

# Introductory Machine Learning

*Yan Tang*

Department of Linguistics, UIUC

Week 13

# Last week…

- Polynomial regression
  - Adds new features by including powers of each raw feature
- Learning curve
  - Underfitting and overfitting
  - Trade-off between bias (simple model) and variance (complex model)

# Last week…

- Ridge Regression

$$C = MSE(c) + \frac{\alpha}{2}\sum_{i=1}^{n} c_i^2$$

- Lasso Regression

$$C = MSE(c) + \alpha\sum_{i=1}^{n} |c_i|$$

- Elastic Net

$$C = MSE(c) + r\alpha\sum_{i=1}^{n} |c_i| + (1-r)\frac{\alpha}{2}\sum_{i=1}^{n} c_i^2$$

# Parametric vs Non-parametric regression

- Parametric regression models
  - Assumes a relation that can be specified using a formula, e.g. $y = c_0 \cdot x_0 + c_1 \cdot x_1 + \ldots + c_N \cdot x_N$
  - Fixed number of model parameters (Not hyperparameters!)
  - Easy to interpret
  - e.g. linear regression and logistic regression

- Creating a model with minimum prediction error can be difficult if there are several predictors

# Parametric vs Non-parametric regression

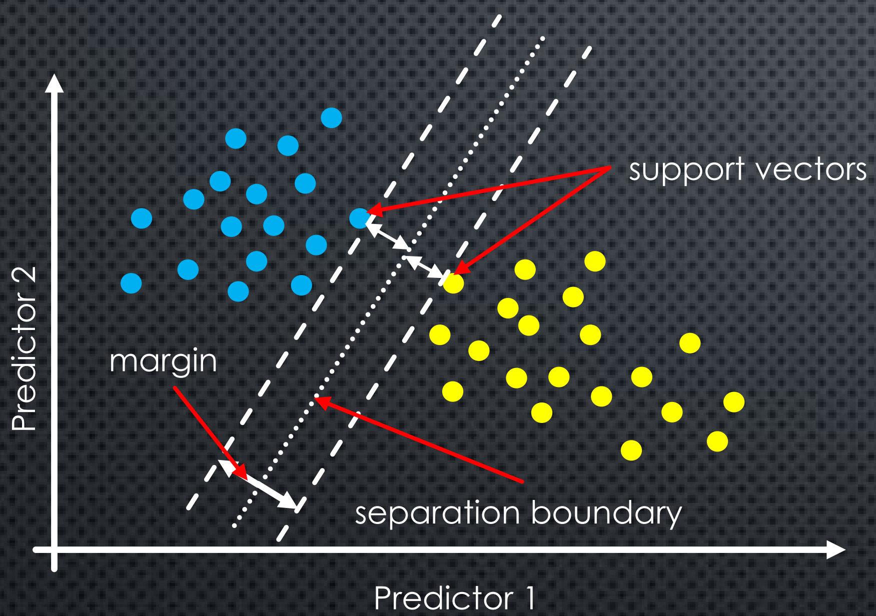
- Non-parametric regression models
  - Do not fit the regression model based on a given formula
  - Can provide more accurate prediction but are difficult to interpret
  - Tend to be overfitting
  - e.g. SVMs, Decision Trees and Gaussian Process Regression

- The primary purpose of the model: predicting the response for unknown observations

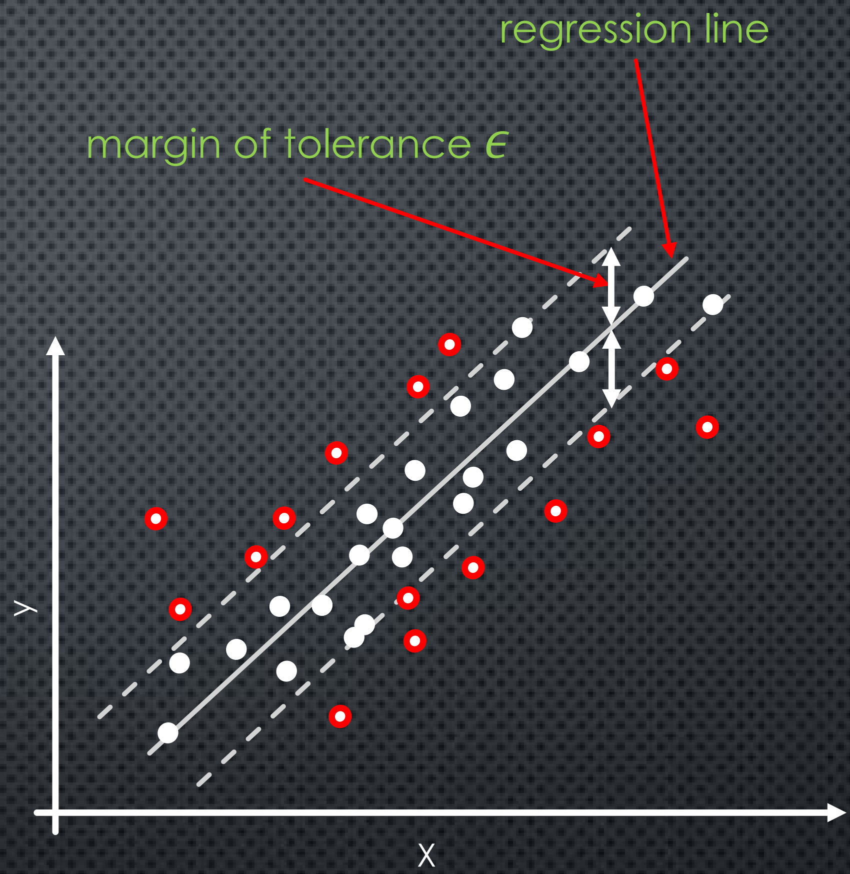
# SVM Regression

- Support Vector Classification can be extended to solve regression problems
- Fitting a model only depends on a subset of the training data
- The objective resembles the opposition of that for classification problems

# SVM Regression




**SVM classification**




**SVM regression**

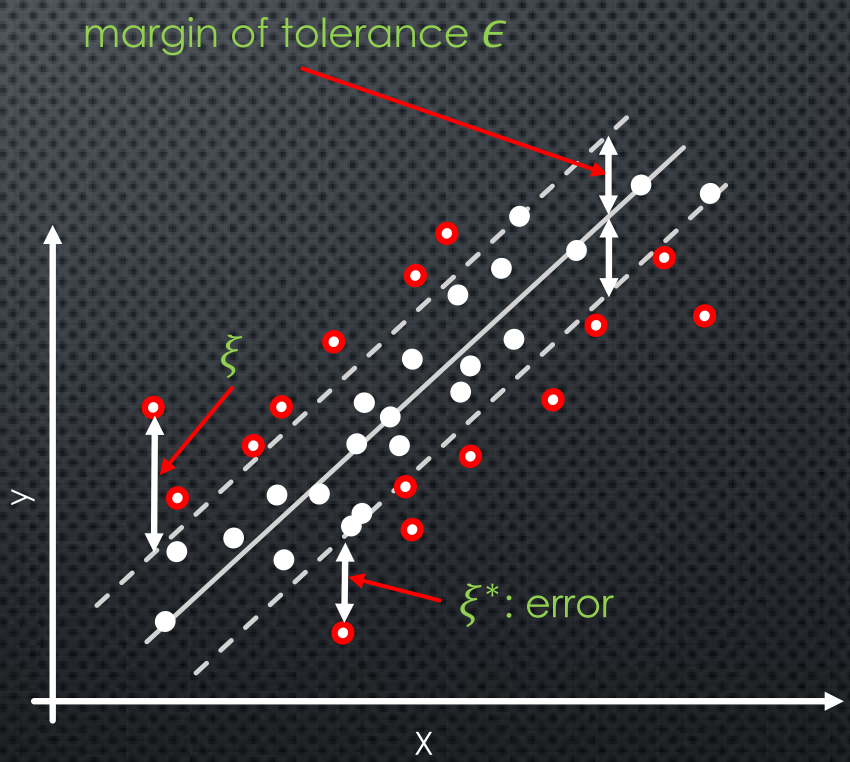
- Hyperparameter $\epsilon$: the greater the value, the larger the margin
  - Opposite to the $C$ for classification

# SVM Regression

- Linear SVM Regression model is $\epsilon$-insensitive
  - Adding more training samples within the margin does not father change the model's behaviour
- Given training data $(X_i, y_i)$, the objective function:

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

$$\left.\begin{aligned}&\forall n:\ y_i - (cx_i + b) \le \epsilon + \xi_i\\ &\forall n:\ (cx_i + b) - y_i \le \epsilon + \xi_i^*\\ &\forall n: \xi_i \ge 0, \xi_i^* \ge 0\end{aligned}\right\}\ \text{Constraints}$$



# Decision-Tree Regression

Determination of variable $X[i]$ and criterion $t_x^i$

CART cost function for regression:

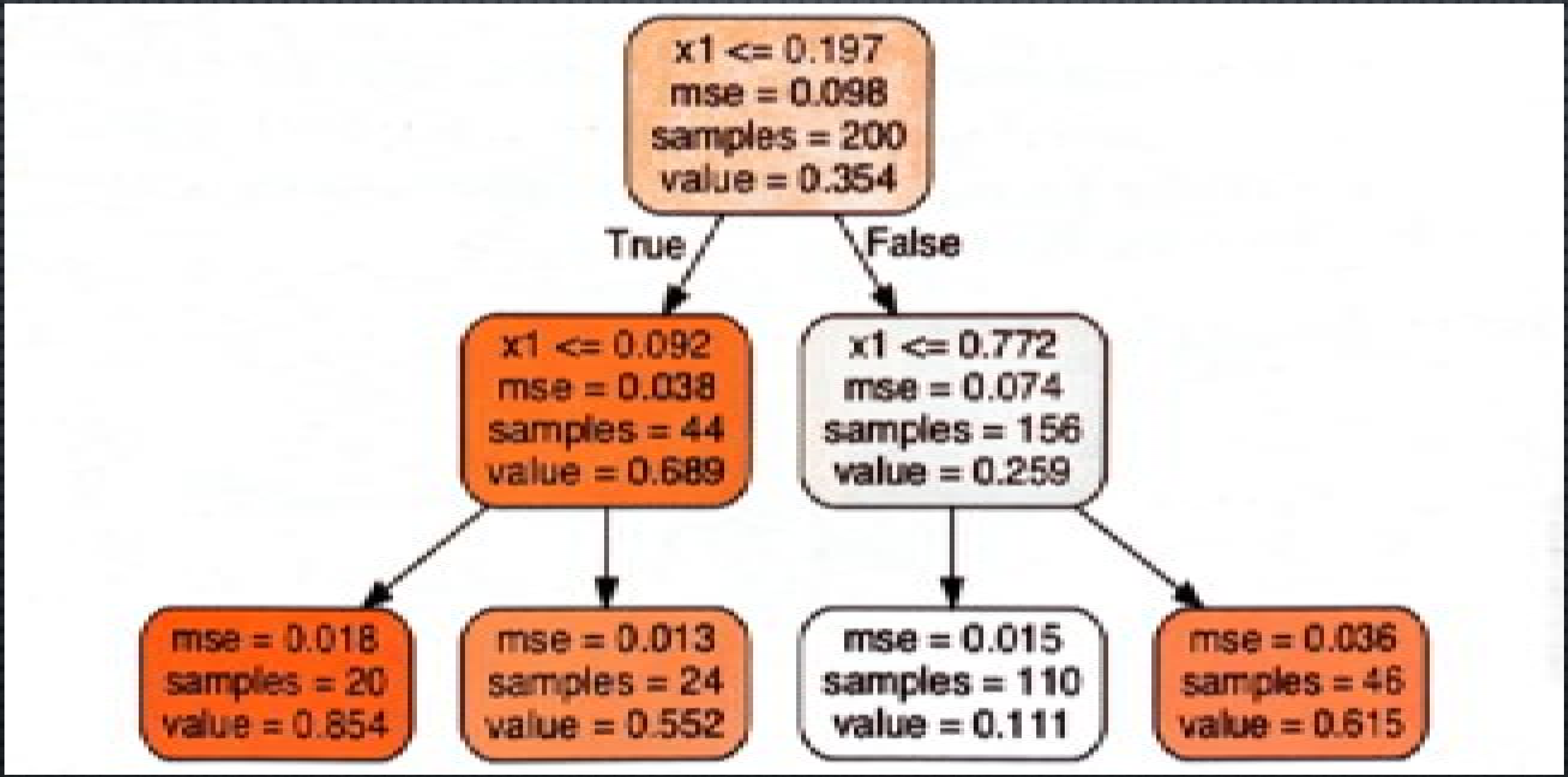
$$C\left(X[i], t_x^i\right) = \frac{m_{left}}{M} MSE_{left} + \frac{m_{right}}{M} MSE_{right}$$

$$MSE = \frac{1}{m} \sum_{n=1}^{m} (\bar{y} - y_n)^2$$

$m_{left}$, $m_{right}$: number of observations in the left/right branch

$\bar{y}$: the mean value of responses in a branch

# Decision-Tree Regression



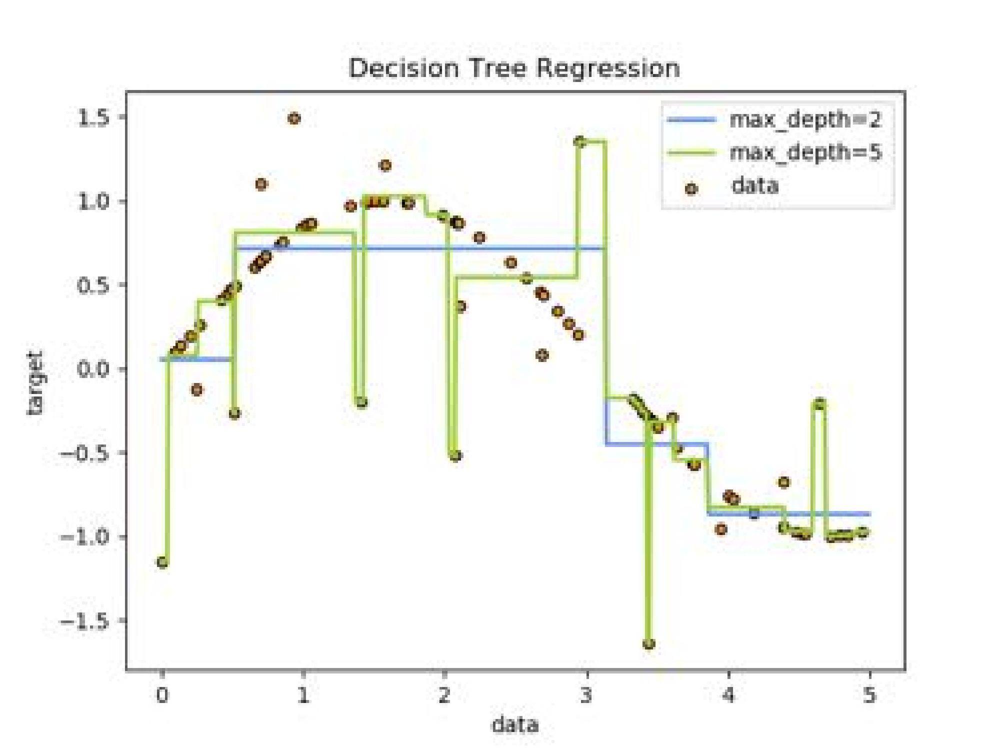



Chapter 6, *Aurélien Géron, 2019*

# Decision-Tree Regression

- For regression problems Decision Trees are prone to overfitting

- Hyperparameters need to be regularised reduce computational resource consumption

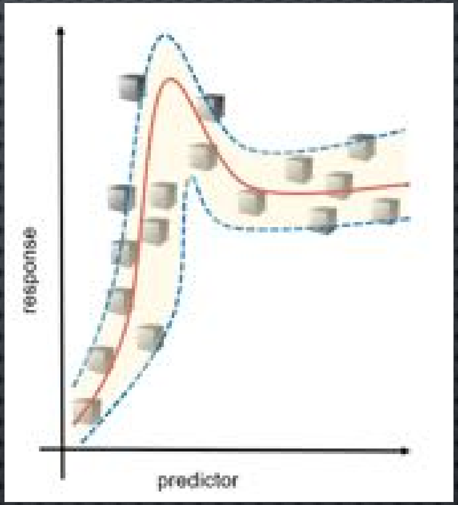





https://scikit-learn.org/stable/auto_examples/tree/plot_tree_regression.html

# Gaussian Process Regression

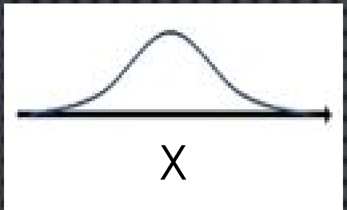
- Gaussian Process Regression (GPR) is another non-parametric regression technique

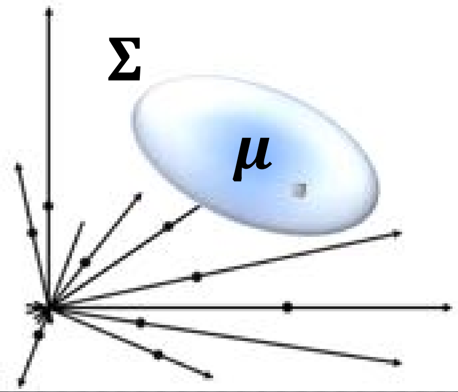


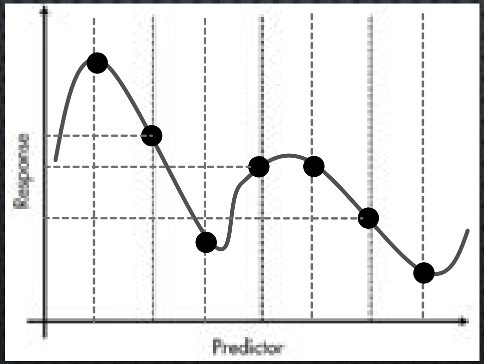

- GPR models optionally return the standard deviation and prediction confidence intervals

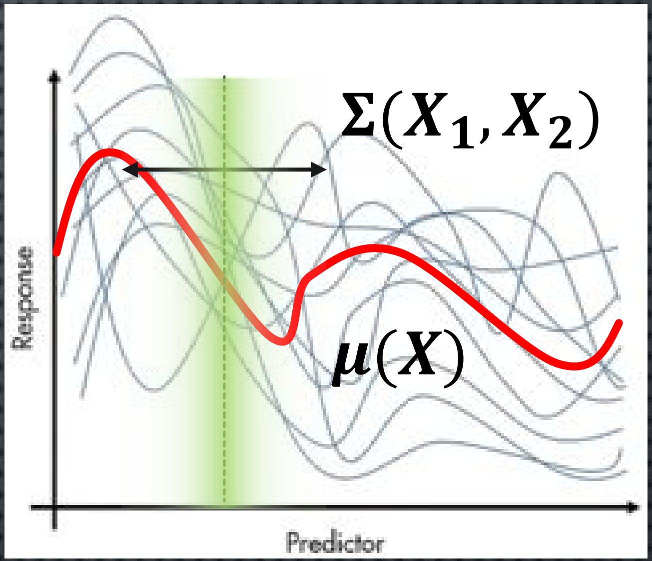
- GPR allows different kernels to be specified

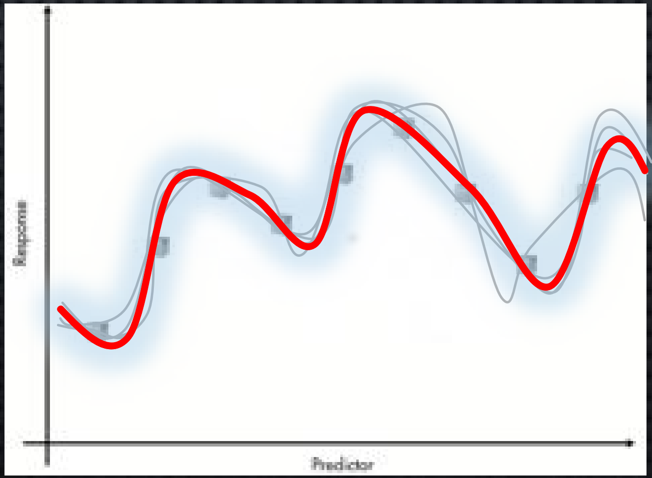
# Gaussian Process Regression



# Gaussian Process Regression

- The disadvantages of Gaussian processes
  - They are not sparse, i.e., they use the whole samples information to perform predictions
  - They lose efficiency in high dimensional spaces when the number of features goes large - low scalability!
- Extensions:
  - Deep Gaussian Process
  - Spare Gaussian Process

# Stepwise Linear Regression

- Stepwise linear regression: a method of regressing multiple variables while simultaneously removing those that are less important
  - Resembles sequential feature reduction and feature elimination
  - Using t- or F-statistics of the estimated coefficients as criteria for removing or adding variables/features
  - Assumptions: normally-distributed data, uncorrelated variables

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_3 x_3$$

$$y = c_0 + c_1 (x_1)^2 + c_2 (x_2)^2 + c_3 (x_3)^2$$

$$y = c_0 + c_1 x_1 + c_2 x_2 + c_3 (x_3)^3$$

$$y = c_0 + c_1 (x_1)^2 + c_2 x_1 x_2 + c_3 (x_2)^2 + c_4 (x_4)^2 + c_5 x_1 x_4$$

# Stepwise Linear Regression

- Stepwise linear regression is highly controversial !
  - + More control over variables
  - + Efficient for fining-tuning a linear model by adding or reducing variables
  - - Models may be oversimplified; underfitting
  - - Tests are biased on the same data: overfitting
  - - Difficult to interpret its reason of feature selection

- Alternatives:
  - LASSO
  - Least Angle Regression (LAR)

# Data preparation for linear regressors

- Encode categorical data
  - ML algorithms like numbers
- Linear assumption
  - The basic assumption on the data
- Remove noise / outliers
  - Linear regression is sensitive to outliers

# Data preparation for linear regressors

- Remove collinearity
  - Highly-correlated features could lead to overfitting
- The effect of a normal distribution
  - Better predictions on normally-distributed data
- Rescale features
  - Most algorithms except features in comparable spaces for efficiency, consistence and accuracy