Bayesian Methods for Data Analysis in Software Engineering
ICSE 2010 – Tutorial

Mohan Sridharan* Akbar Siami Namin**

*Learning Agents and Stochastic Reasoning Laboratory
Department of Computer Science
Texas Tech University

**AdVanced Empirical Software Testing and Analysis Research Group
Department of Computer Science
Texas Tech University

May 3, 2010 - Cape Town

Linear Regressions and Classifiers Definition

- Training set
  - A collection of data observed with some attributes and classes
- Fitting a model
  - Finding a model for attributes as a function of other attributes
- Goodness of fit
  - Measure the accuracy of the model fitted
- Test set
  - Predict the behavior of unseen data using model developed
    - Assigning the unseen data to classes

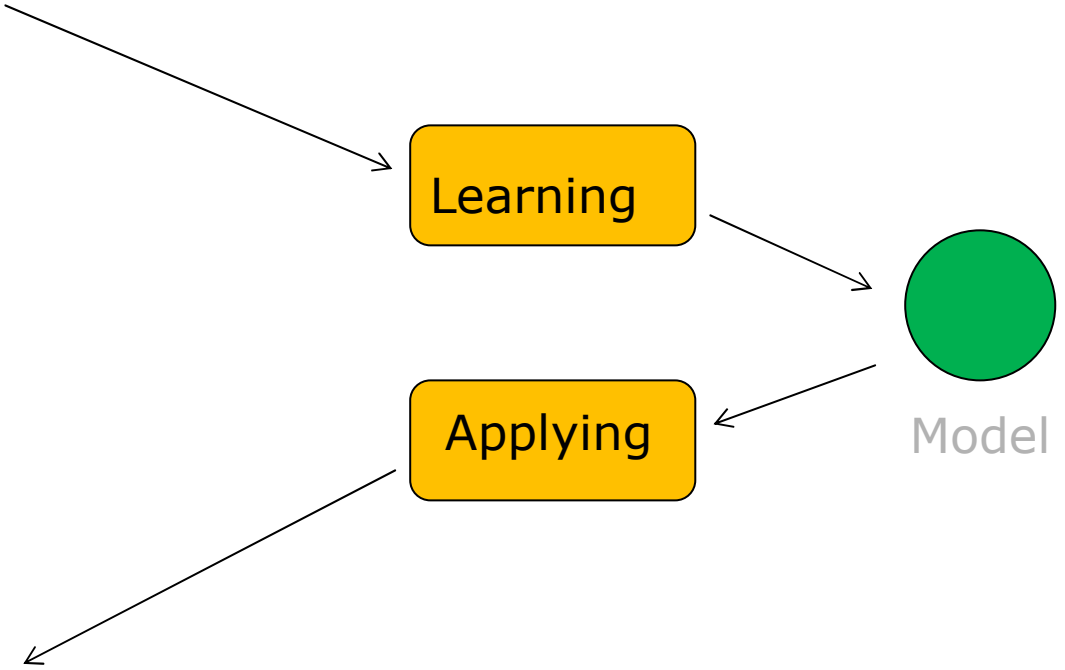
Linear Regressions and Classifiers Definition

a1	a2	a3	a4
12	20	1	Y
23	99	1	N
14	87	2	Y
22	54	3	Y

Training set

a1	a2	a3	a4
12	54	22	?
15	33	43	?
75	23	27	?

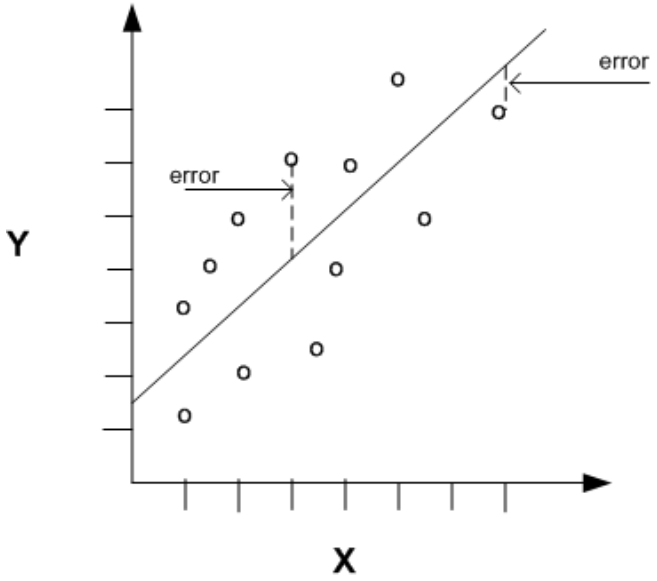
Test set



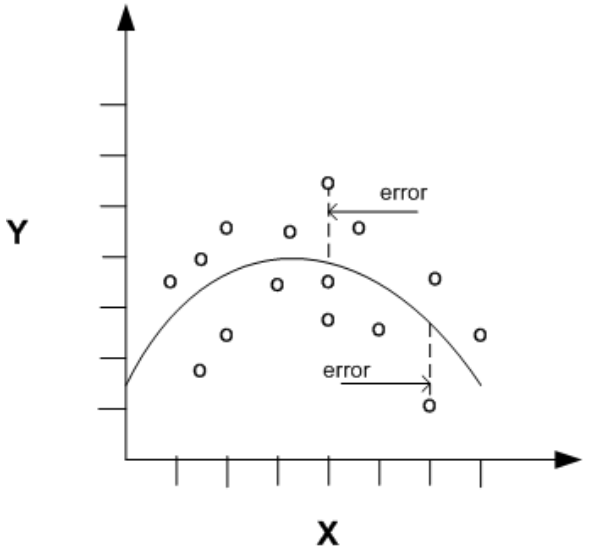
Linear Regressions

- Fitting a line to the data
- Analysis of regression
- Goodness of fit

Linear Regressions
Different Types of Regression Lines



Linear Regression



Non-linear Regression

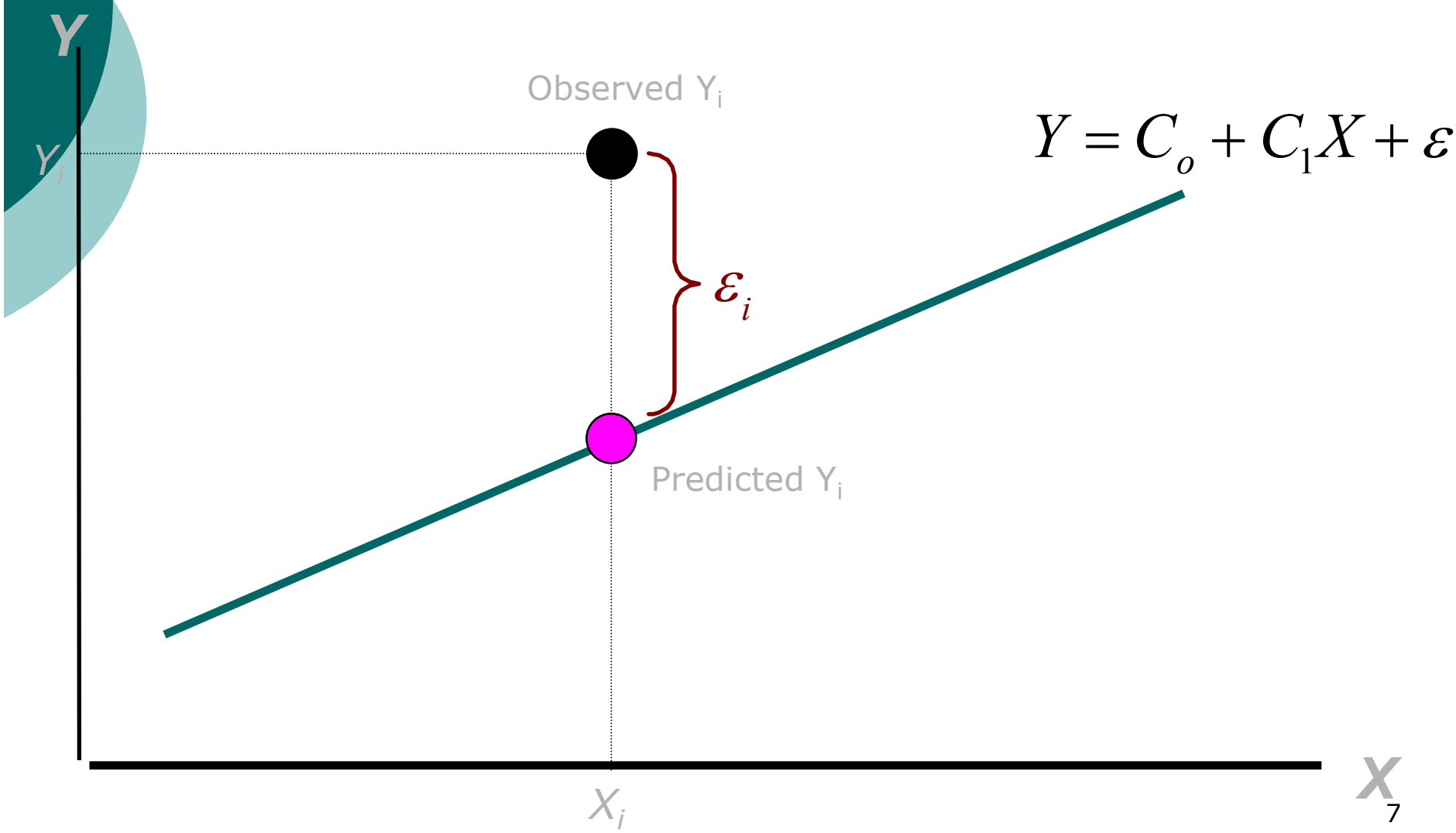
Linear Regressions
Simple Linear Regressions

- A statistical technique using a variable X to estimate variable Y
  - X: independent variable
  - Y: Dependent variable, response variable→

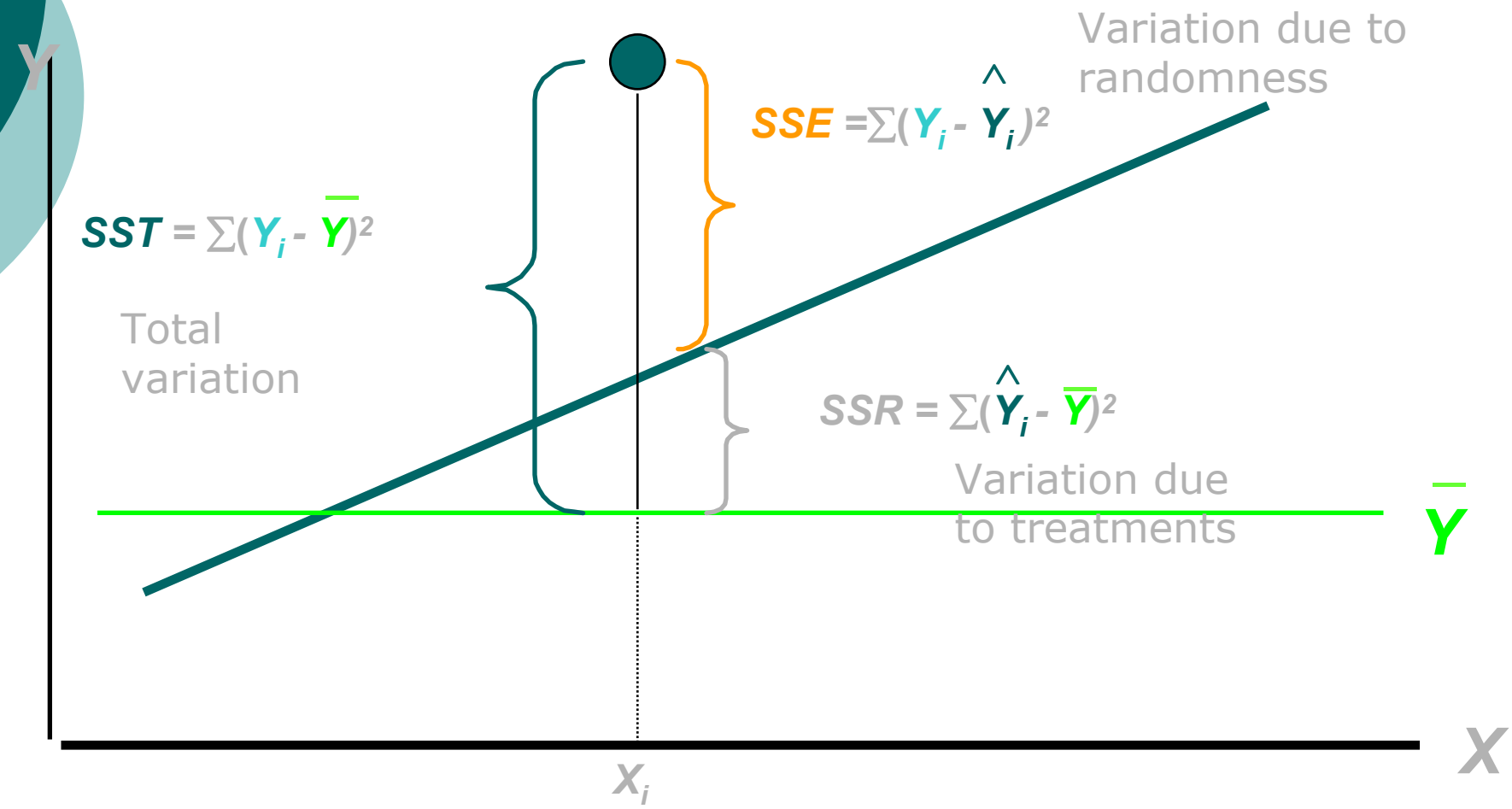
$$Y = C_o + C_1 X + \varepsilon$$

- Regression coefficients: C_o, C_1
- Residuals: ε

Linear Regressions
Simple Linear Regressions



Linear Regressions
Visualization of Variation: SST = SSE + SSR



Linear Regressions
Variation to Randomness, SSE

- Choosing the prediction that minimizes the sum of squares of the deviations of the observed from the predicted values
  - Denoted by SSE

$$SSE = \sum (Y - \hat{Y})^2$$

  - Y: true values
  - $\hat{y}$: estimated values

Linear Regressions
Variation to Regression, SSR

- The mean values
  - An unbiased estimator
  - Minimize the deviations of the estimated from those computed by an unbiased estimator

$$SSR = \sum (\hat{Y} - \overline{Y})^2$$

  - $\overline{Y}$: mean (unbiased estimated) values
  - $\hat{Y}$: estimated values

Linear Regressions
SST = SSE + SSR

- SST =SSE + SSR

$$SSE = \sum (Y - \hat{Y})^2$$

$$SSR = \sum (\hat{Y} - \overline{Y})^2$$

$$\Rightarrow SST = \sum (Y - \overline{Y})^2$$

Linear Regressions
Computing the Regression Coefficients

$$SST = \sum (Y - \overline{Y})^2$$

$$SSx = \sum (X - \overline{X})^2 \qquad SSy = \sum (Y - \overline{Y})^2$$

$$SSxy = \sum (X - \overline{X})(X - \overline{Y})$$

$$C_1 = SSxy / SSx \qquad C_0 = \overline{Y} - C_1 \overline{X}$$

Linear Regressions
Goodness of Fit

- Standard error
- Mean square error
- Coefficient of determination: R-squared

Linear Regressions
Standard Errors

- Standard errors of coefficients - SSE
  - Measures of the variation of the intercept and slops from the expected values
- Standard error of regression - SSR
  - Measure of the variation of the regression line

Linear Regressions
Mean Squared Error

- Mean squared error of coefficients with df degree of freedom
  - MSE = SSE / df
- Mean squared error of regression line with df degree of freedom
  - MSR = SSR / df

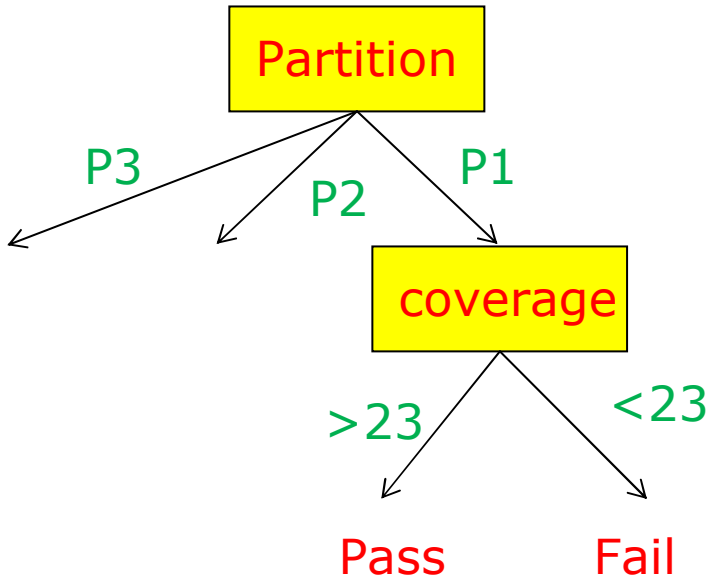
Linear Regressions
Coefficient of Determination: R-squared

- Measure of how much variation of the expected values of the response variable is explained by the estimated values
  - R^2 = SSR/SST
    - A value between [0, 1], higher better

Tree-based Classifiers
Example of A Decision Tree

Test Case	Partition	Line Coverage	Pass| fail
T1	P1	23	Pass
T2	P2	15	Pass
T3	P2	45	Fail
T4	P3	57	Fail
T5	P1	32	Pass

Training Set



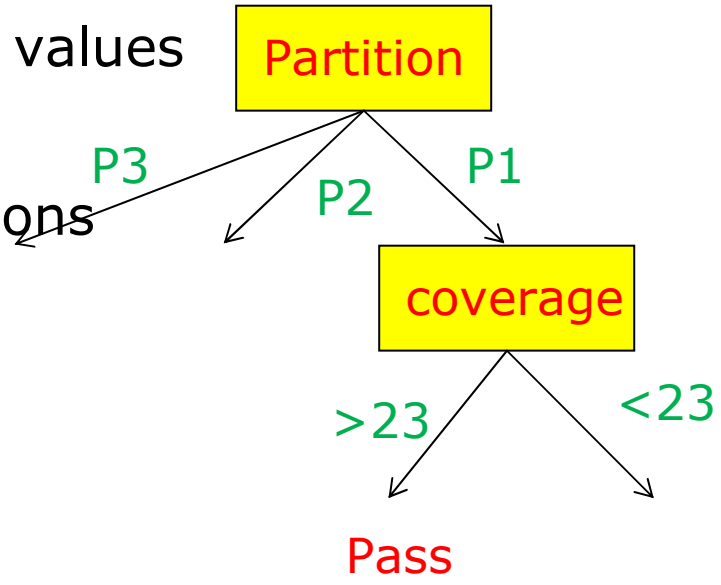
Decision Table: Model

Tree-based Classifiers
Tree Induction – Splitting Strategy

- Split based on an attribute test that optimize certain criteria
- Determine how to split the records
  - How to specify the attribute test condition?
  - How to determine the best split?
- Determine when to stop splitting

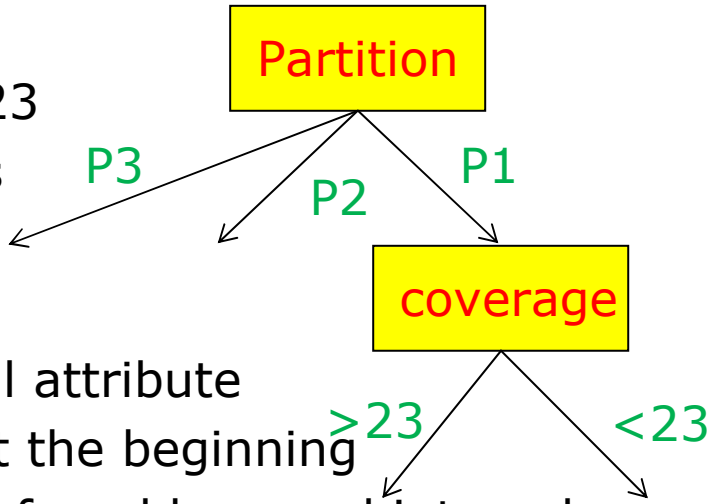
Tree-based Classifiers
Splitting Test Conditions

- Two-way split vs. Multi-way split
- Multi-way split
  - Many partitions as distinct values
- Two-way split
  - Divide into only two partitions



Tree-based Classifiers
Splitting Test Conditions

- Splitting based on continuous attributes
  - Binary Decision
    - E.g. coverage <23 and >23
    - Consider all possible splits
      - Identify the best one
  - Discretization
    - form an ordinal categorical attribute
    - Static – discretize once at the beginning
    - Dynamic – ranges can be found by equal interval
      - E.g. bucketing, equal frequency bucketing (percentiles), or clustering.



Tree-based Classifiers
Identifying The Best Split

- For each node, the notions of:
  - Homogenous class distribution
  - Impurity

C0: 5 C1: 5	C0: 9 C1: 1
Non-homogeneous, **High degree of impurity**	**Homogeneous,** **Low degree of impurity**

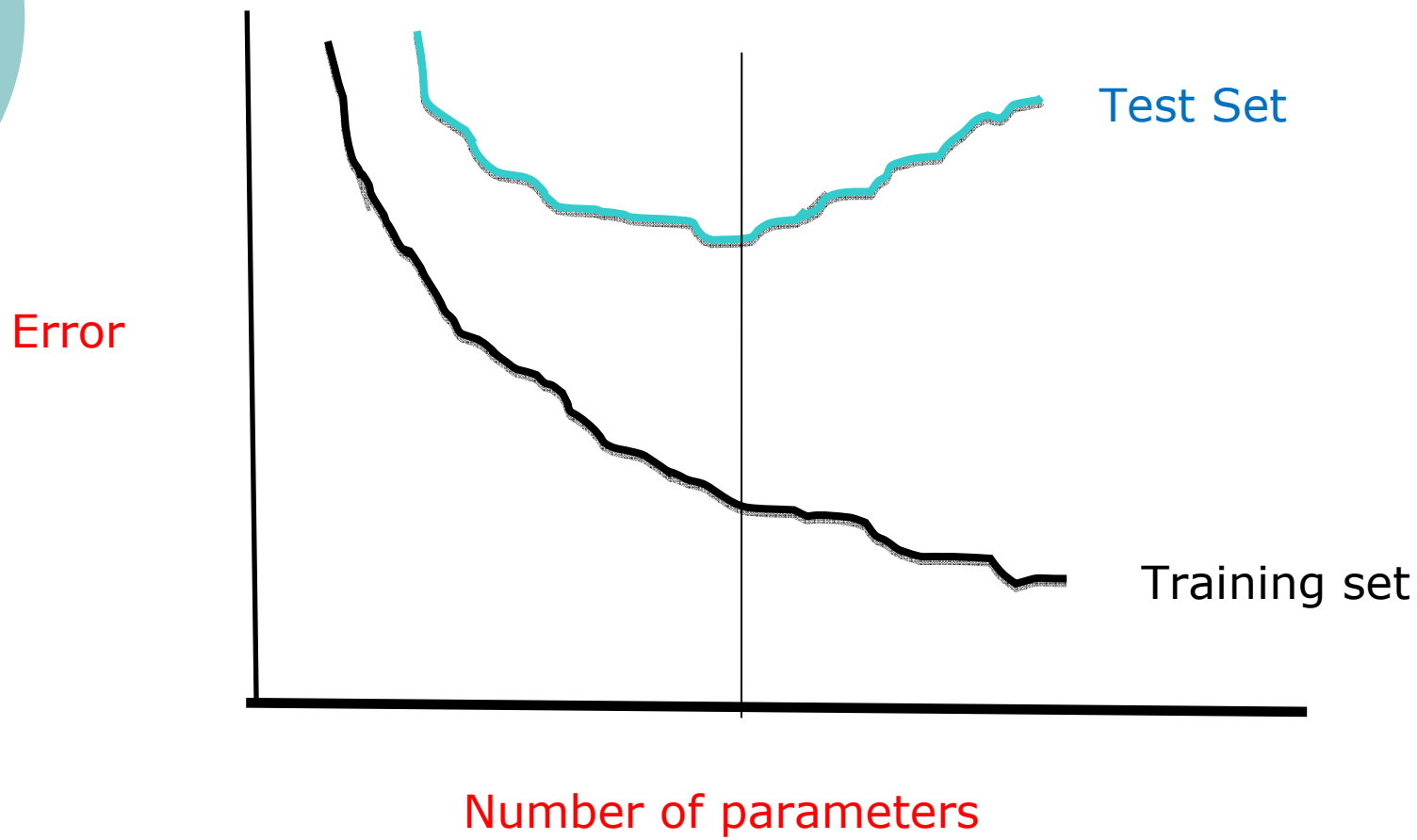
Tree-based Classifiers
Stopping Criteria

- Split until all remaining records belong to one class
- Split until the remaining records have similar attribute values

Linear Regression and Classifiers
The major issue – Over (under) fitting

- Over fitting
  - The model relies solely on the training set
  - The model is too complex than necessary
  - E.g. Too many explanatory variables in the regression model
  - E.g. Too many parameters involved in splitting decision
- Under fitting
  - The model lacks enough precision and knowledge

Tree-based Classifiers
The Over fitting Problem



Linear Regressions
Cross-Validation

- A procedure to identify over fitting problem
- Splitting data into:
  - Training set
  - Test set
- Develop and train a model for the training set
- Test the model with the test set.
  - Pick a model where the error is minimum