Åbo Akademi University

Classification and logistic regression

József Mezei

Business Analytics I

Machine learning

- Learning from past experiences
- A computer system learns from data of an application domain
- Now: predict the values of a discrete class attribute, e.g., approve or not-approved, high-risk or low risk.
- Supervised learning, classification

Supervised vs. Unsupervised Learning

Supervised learning (classification)

- Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations
- New data is classified based on the training set

Unsupervised learning (clustering)

- The class labels of training data is unknown
- Given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data

Åbo Akademi University

Classification vs. Prediction

- Classification:
  - predicts categorical class labels
  - classifies data (constructs a model) based on the training set and the values (class labels) in a classifying attribute and uses it in classifying new data
- Prediction:
  - models continuous-valued functions

Åbo Akademi University

Classification

Typical Applications

- credit approval
- target marketing
- medical diagnosis
- treatment effectiveness

Example 1.

- Patients in a hospital: numerous variables (e.g., blood pressure, age, etc.) of newly admitted patients
- Decision to make: assigning patient to the intensive-care unit or not
- Prioritizing more severe patients
- Task: identify high-risk patients

Example 2.

- A credit card company receives thousands of applications for new cards
- Information about an applicant
  - Age, marital status, salary, debts, credit rating, etc.
- Task: classify applications into two groups, approve and reject

Classification—A Two-Step Process
Model Construction

Model construction: describing a set of predetermined classes

- Each observation is assumed to belong to a predefined class
- The set of observations used for model construction: training set
- The model is represented as classification rules, decision trees, mathematical formulae, etc.

Åbo Akademi University

Classification—A Two-Step Process
Model Usage

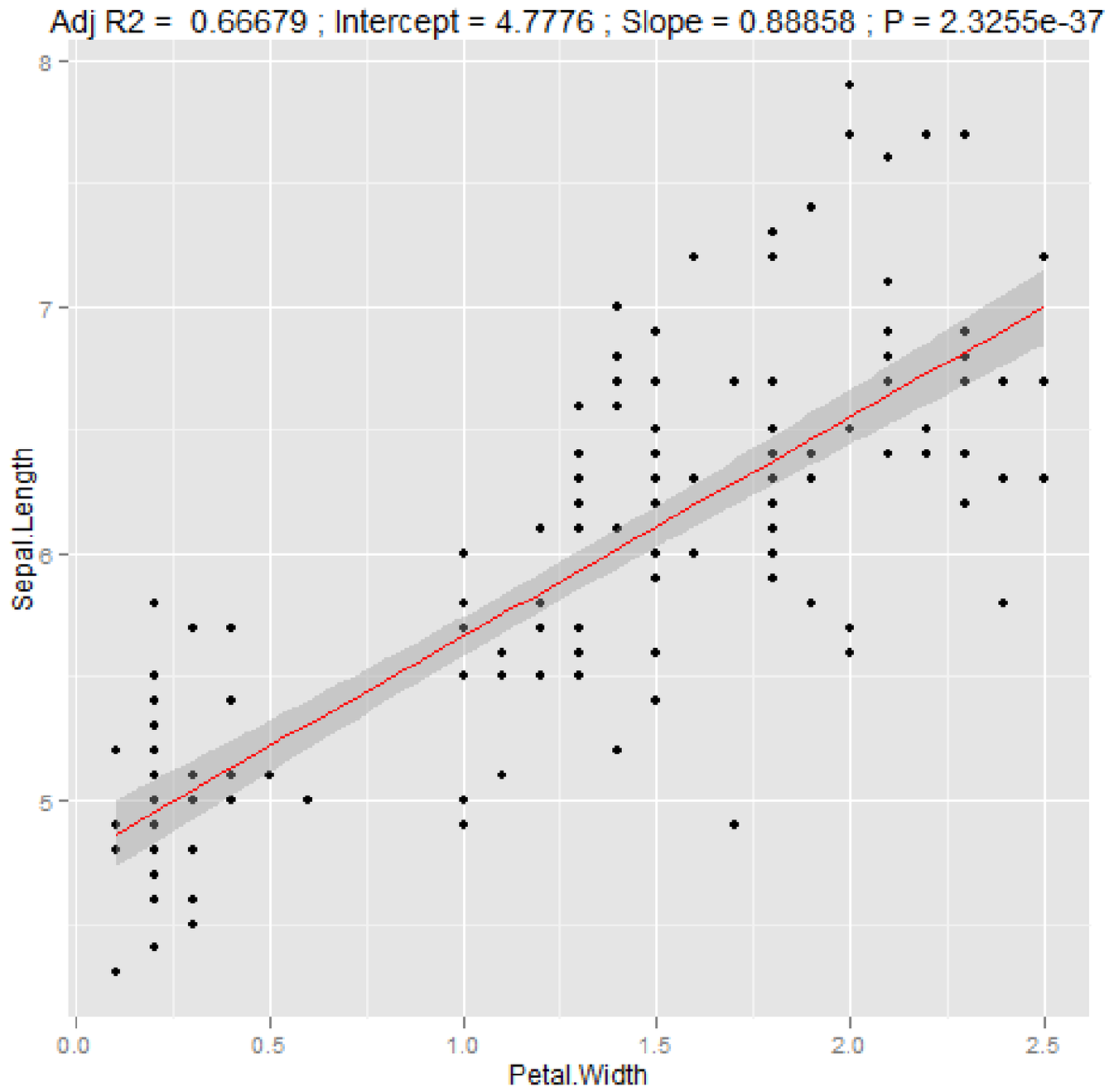
Model usage: for classifying future or unknown objects

- Estimate accuracy of the model
- The known label of test sample is compared with the classified result from the model
- Accuracy rate is the percentage of test set samples that are correctly classified by the model
- Test set is independent of training set, otherwise over-fitting will occur

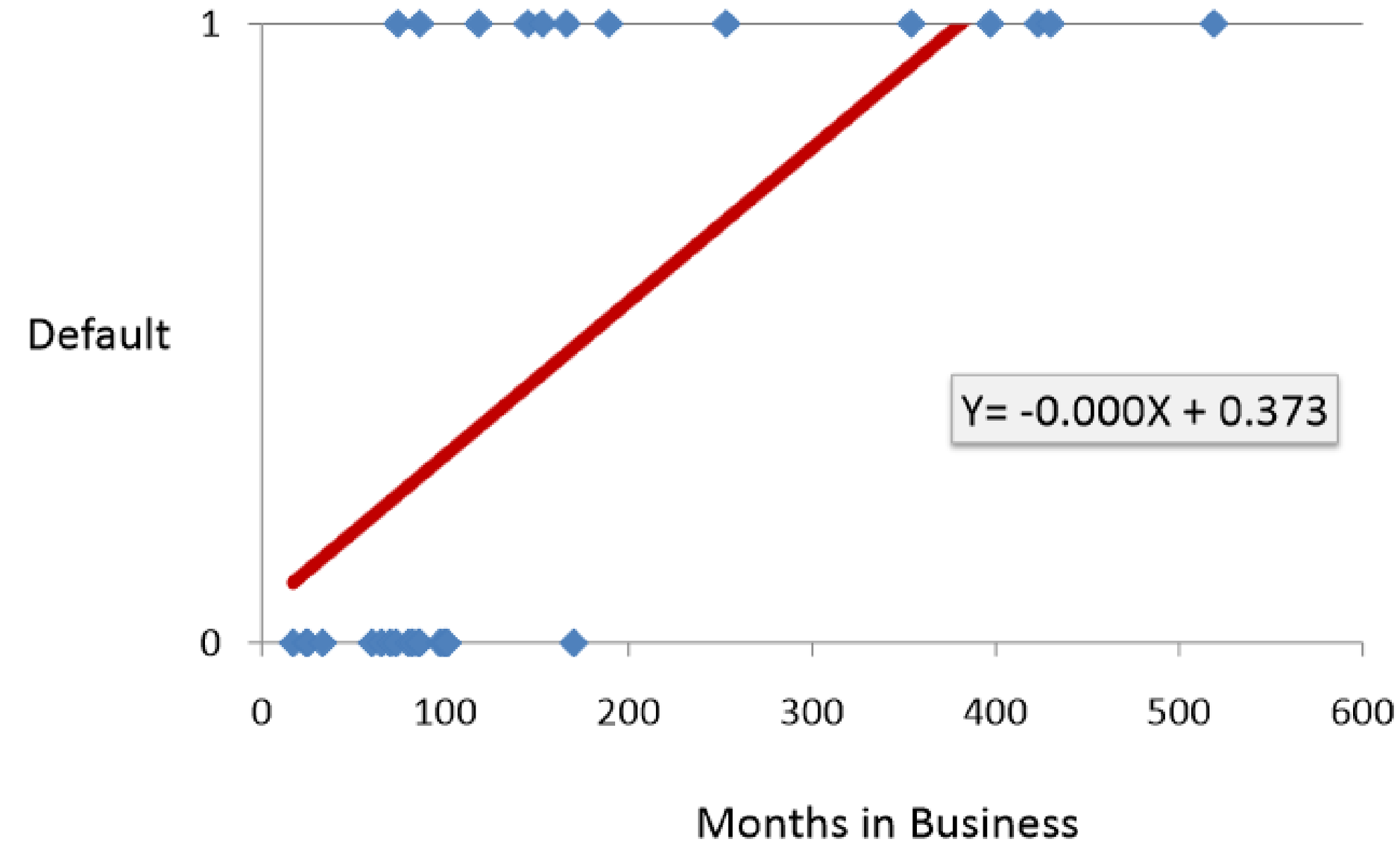
The data and the goal

- Data: A set of data records described by
  - k attributes: A_1, A_2, … A_k
  - a class: Each example is labelled with a pre-defined class
- Goal: To learn a classification model from the data that can be used to predict the classes of new cases

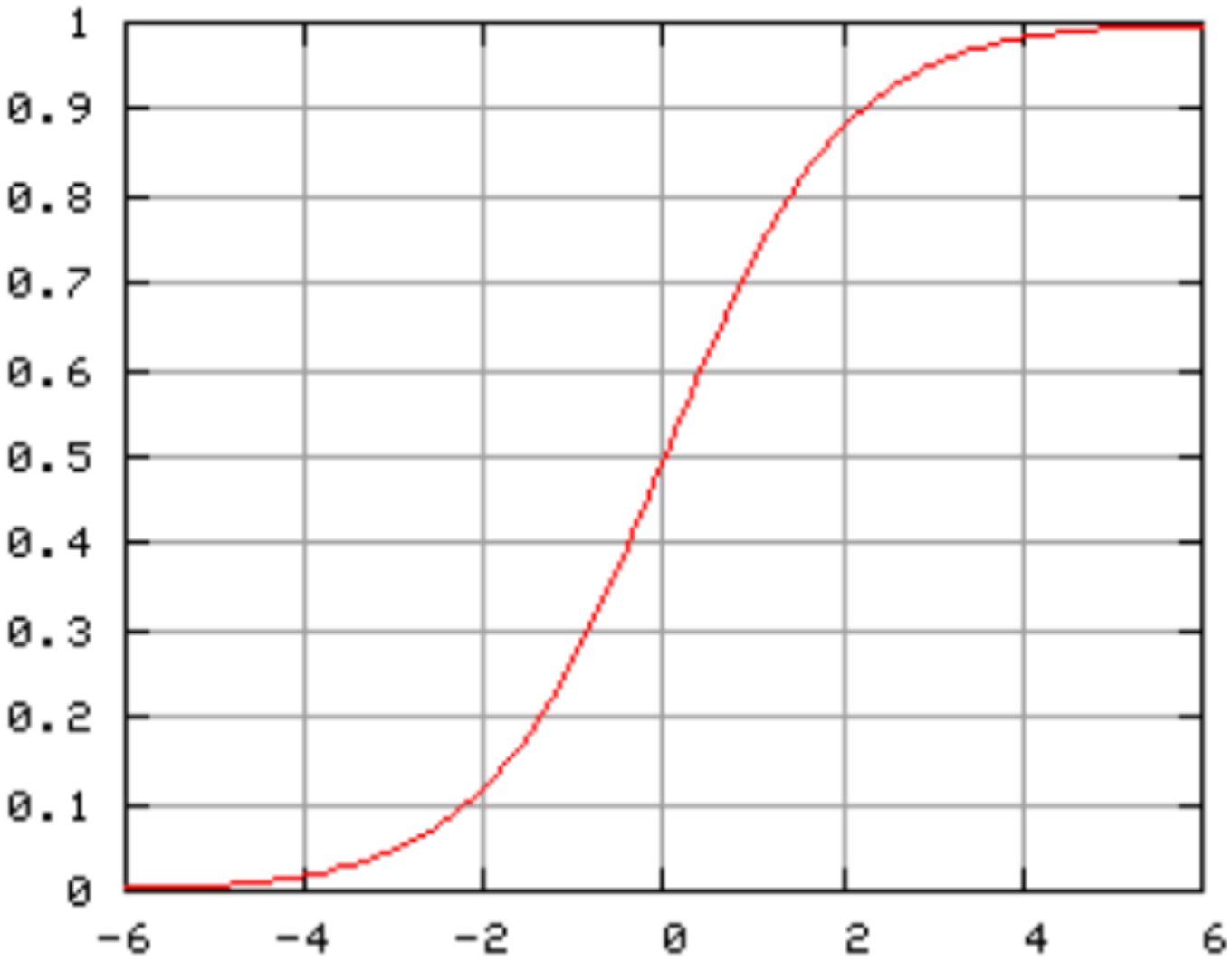
Linear versus logistic regression



Linear versus logistic regression



Logistic function



$$f(z) = \frac{1}{1+e^{-z}}$$

Example data

ID	Age	Has_Job	Own_House	Credit_Rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

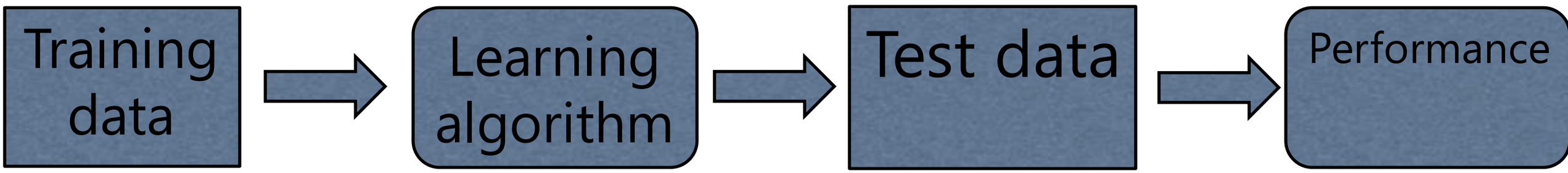
Predicting class for new cases

- Learn a classification model from the data
- Use the model to classify future loan applications into
  - Yes (approved) and
  - No (not approved)
- What is the class for following case/instance?

Age	Has_Job	Own_house	Credit-Rating	Class
young	false	false	good	?

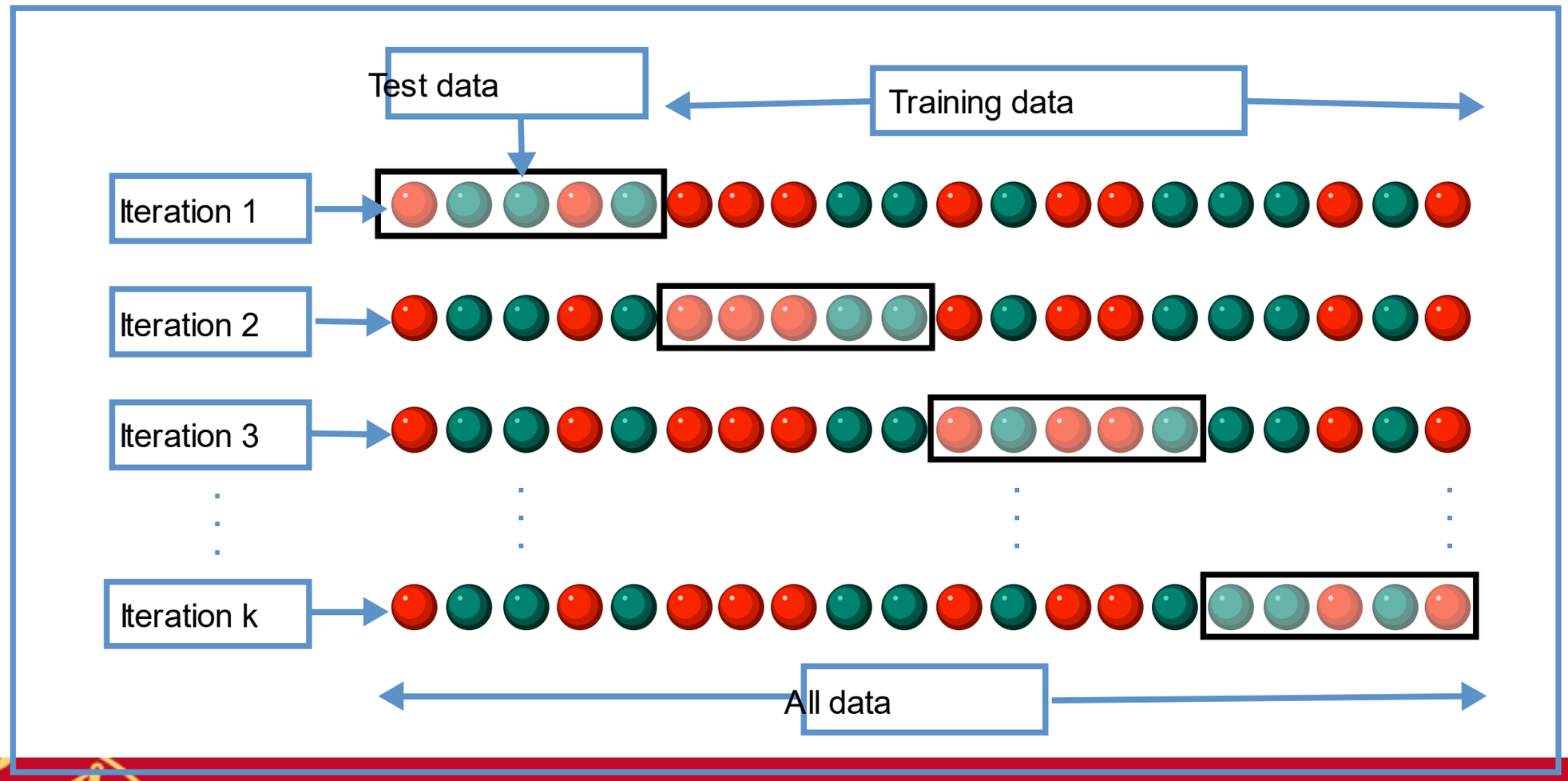
Supervised learning in two steps

- Learning (training): Learn a model using the training data
  - Testing: Test the model using unseen test data to assess the model accuracy



Cross-validation

- Instead of having one training and test set, we divide our dataset into k parts (k folds) and choose the model that has the best average perfomance across all folds



Fundamental assumption of learning

- Assumption: The distribution of training examples is identical to the distribution of test examples (including future unseen examples).
- In practice, this assumption is often violated to certain degree.
- Strong violations will clearly result in poor classification accuracy.
- To achieve good accuracy on the test data, training examples must be sufficiently representative of the test data.

Evaluation methods

- The available data set *D* is divided into two disjoint subsets,
  - the *training set* (for learning a model)
  - the *test set* (for testing the model)
- Training set should not be used in testing and the test set should not be used in learning.
  - Unseen test set provides a unbiased estimate of accuracy.
- This method is mainly used when the data set *D* is large (for smaller datasets: cross-validation)

Classification measures

- Accuracy is only one measure (error = 1-accuracy).
- **Accuracy is not suitable in some applications**
- In classification involving imbalanced data, we are interested only in the minority class.
- High accuracy can be achieved by missing many cases of the smaller class
- The class of interest is commonly called the **positive class**, and the rest **negative classes.**

Confusion matrix

		Predicted Class: Positive	Predicted Class: Negative
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)

Performance evaluation

		Predicted Class: Positive	Predicted Class: Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	**Negative Predictive Value** $\frac{TN}{(TN+FN)}$	**Accuracy** $\frac{TP+TN}{(TP+TN+FP+FN)}$

Naïve Rule

Naïve rule: classify all records as belonging to the most prevalent class

- Often used as benchmark: we hope to do better than that

Confusion Matrix

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	201	85
0	25	2689

201 1’s correctly classified as “1”

85 1’s incorrectly classified as “0”

25 0’s incorrectly classified as “1”

2689 0’s correctly classified as “0”