Basic Learning Approaches and Complexity Control

CS 6316 – Machine Learning
Fall 2017

OUTLINE

3.0 Objectives

3.1 Terminology and Basic Learning Problems

3.2 Basic Learning Approaches

3.3 Generalization and Complexity Control

3.4 Application Example

3.0 Objectives

1. To quantify the notions of explanation, prediction and model
2. Introduce terminology
3. Describe basic learning methods
   - Past observations ~ data points
   - Explanation (model) ~ function

   → Learning ~ function estimation

   Prediction ~ using estimated model to make predictions

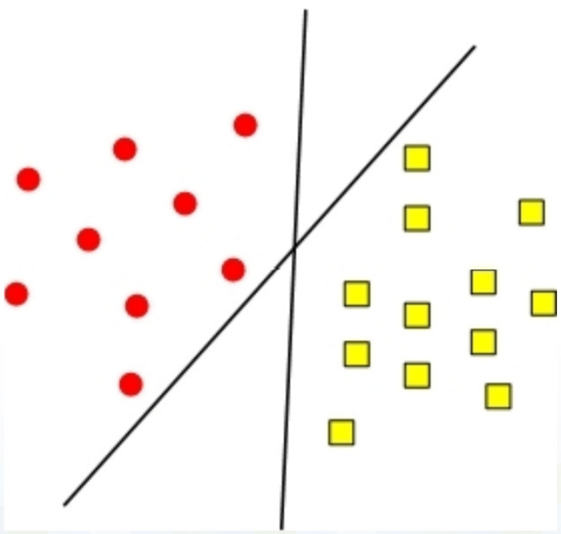
Objectives

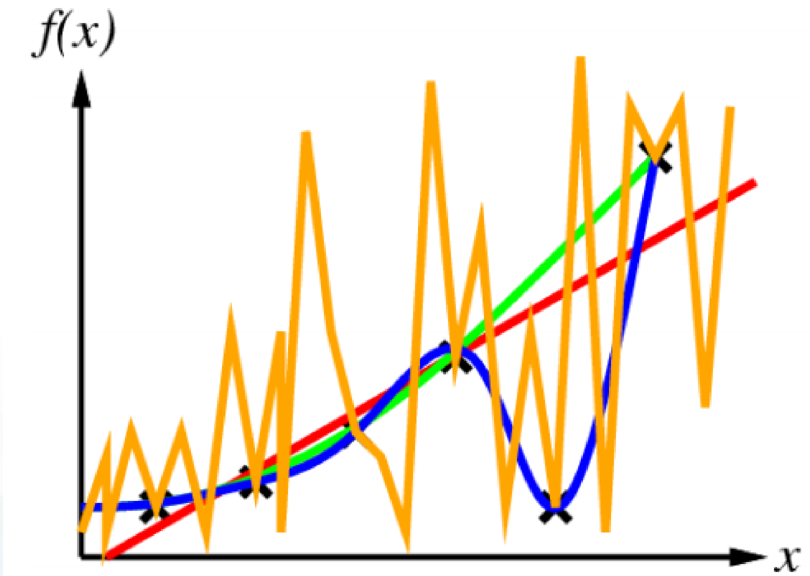
- Example: classification

  training samples, model

  Goal 1: explanation of training data

  Goal 2: generalization (for future data)

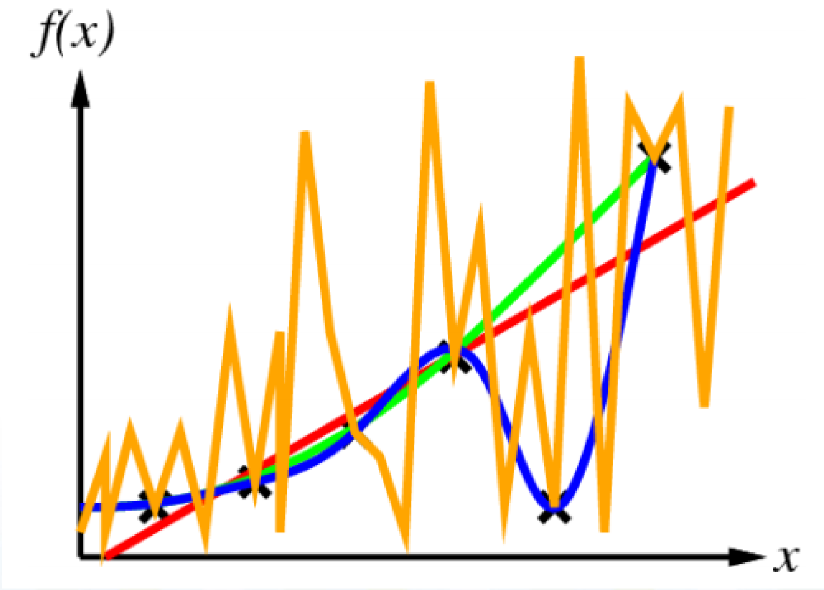




Learning is ill-posed!

Objectives

- What is "ill-posed"?
  - It is possible to find *multiple* hypotheses that are consistent with a given training set
  - How do we choose between these alternative hypotheses
  - This is an example of an ill-posed problem where the **data** by itself is not sufficient to find a unique solution



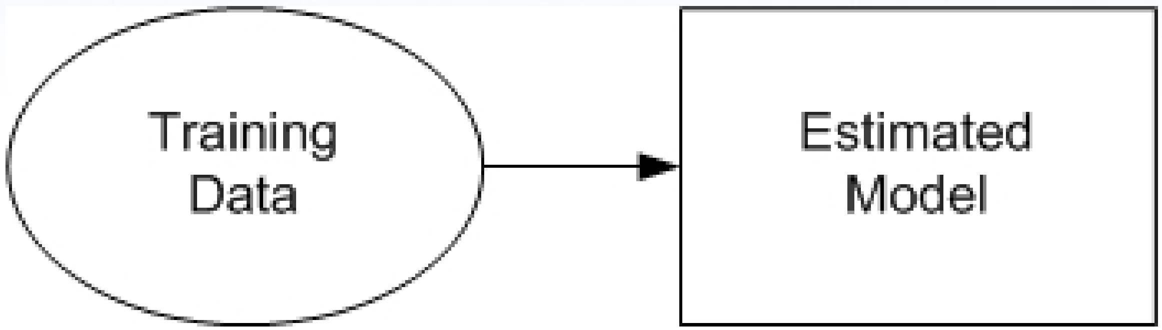
Objectives

- Inductive Learning is an ill posed problem
- Unless we see all possible examples, the data by itself is *not sufficient* for an inductive learning algorithm to find a unique solution
- Therefore learning is challenging

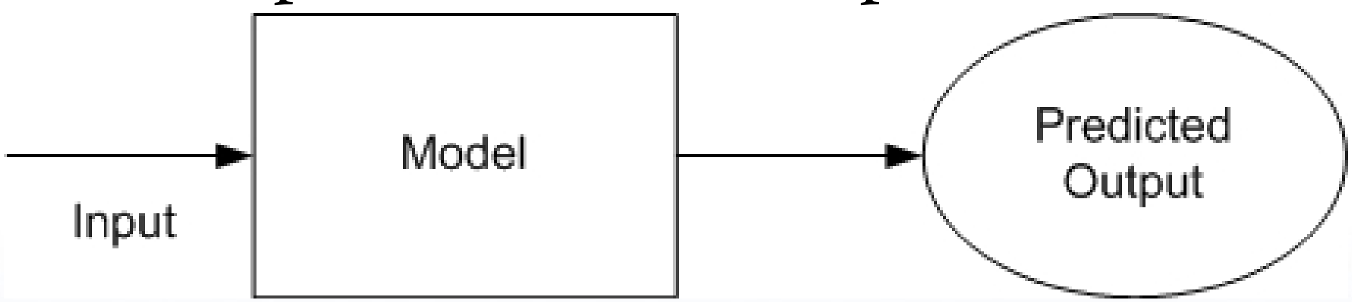
Inductive Learning

"general rule"

- Induction ~ function estimation from data:



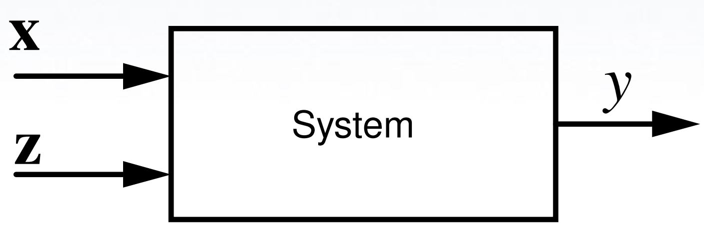
- Deduction ~ prediction for new inputs:



INDUCTIVE LEARNING used for most
ML/statistical/data mining algorithms

3.1 Terminology & Learning Problems

- Input and output variables



- Learning ~ function estimation from noisy samples. Estimating dependency between several input variables (input vector **x**) and output y $F(X)\colon\ X \to y$
- Training data ~ $(\boldsymbol{x}_i, y_i)\ i = 1, 2, \ldots, n$

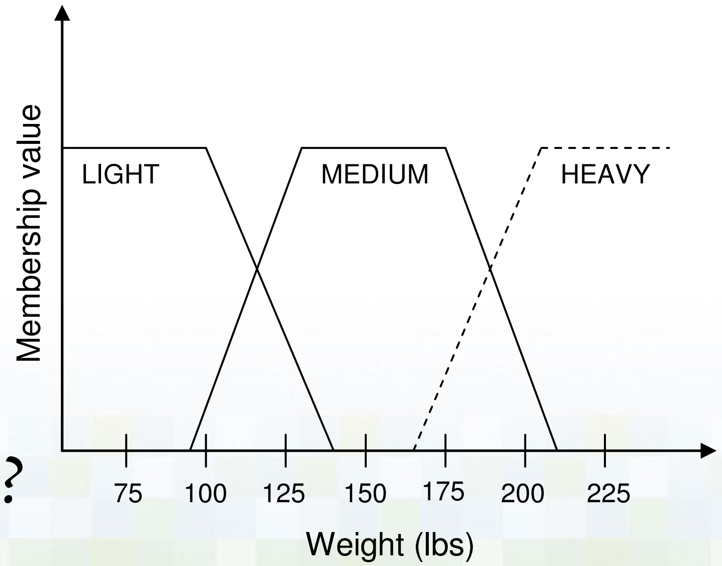
  Past observations of (input, output) samples
- Model ~ The estimated function that is used to predict output values for new inputs

Types of Input and Output Variables

- Numeric: real-valued or integer (*age, speed, length*)
  - For any two feature values there is:
    - Order and a distance relation
- Categorical (class labels): take on certain values (*eye color, gender, country or origin*)
  - For any two feature values there is:
    - No order and no distance relation
  - Categorical outputs occur quite often and represent a class of learning problems known as: **PATTERN RECOGNITION** or **CLASSIFICATION**

Types of Input and Output Variables

- Ordinal (or fuzzy) variables: similar to categorical but there are no crisp boundaries (*gold, silver, bronze medals*)
  - For any two feature values there is:
    - Order but no distance relation
  - Encode numeric values → small set of overlapping intervals corresponding to the values (labels) of an ordinal variable
  - Another example: *young, middle-aged, old*
  - *What does young/old mean to you?*



Data Preprocessing and Scaling

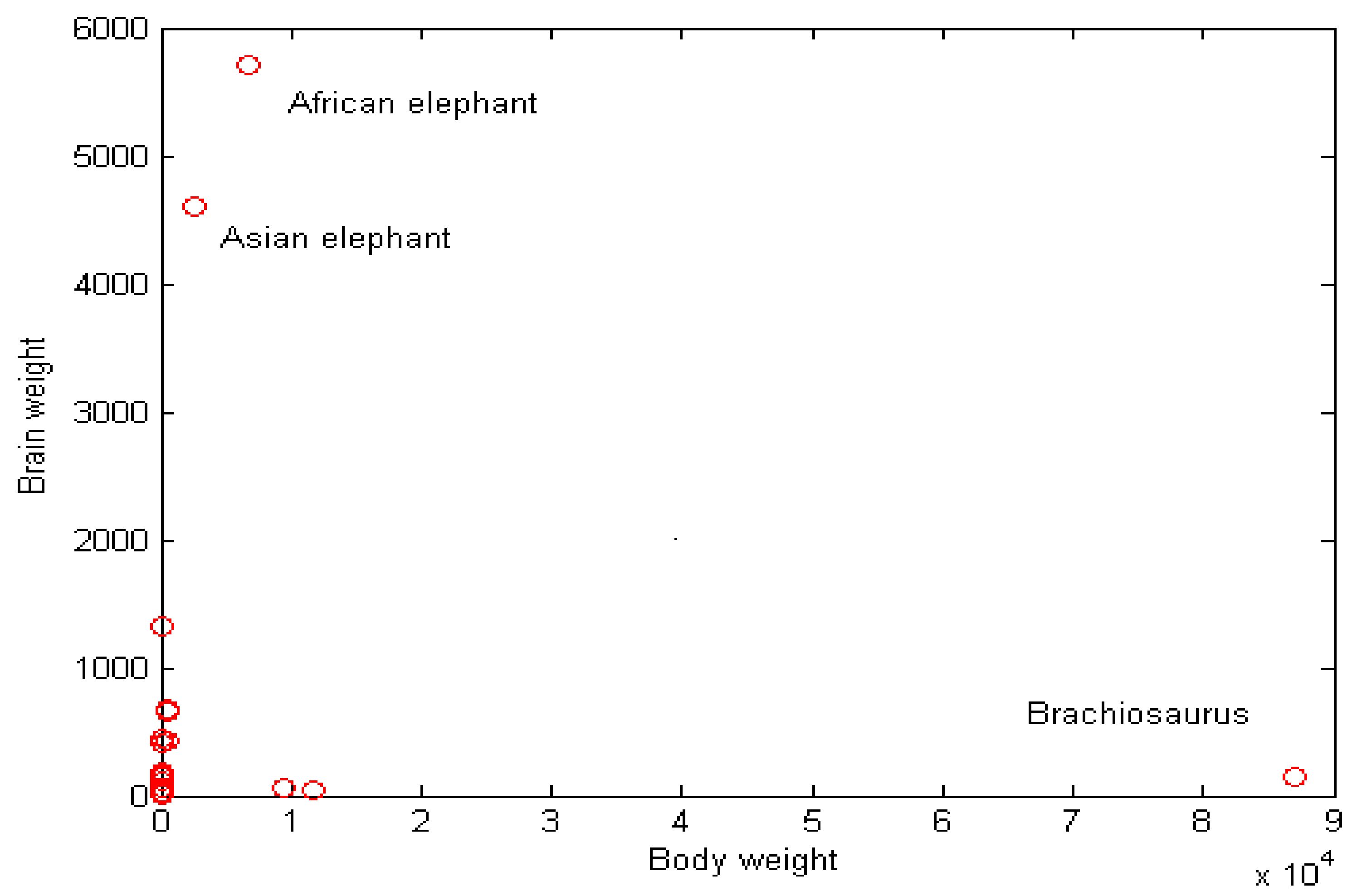
- Preprocessing is required with observational data (step 4 in general experimental procedure) Examples: …
- Basic preprocessing includes
  - summary univariate statistics: mean, st. deviation, min + max value, range, boxplot
    performed independently for each input/output
  - detection (removal) of outliers
  - scaling of input/output variables
    (*may be required for some learning algorithms*)
- Visual inspection of data is tedious but useful

Example Data Set: Animal body & brain weight

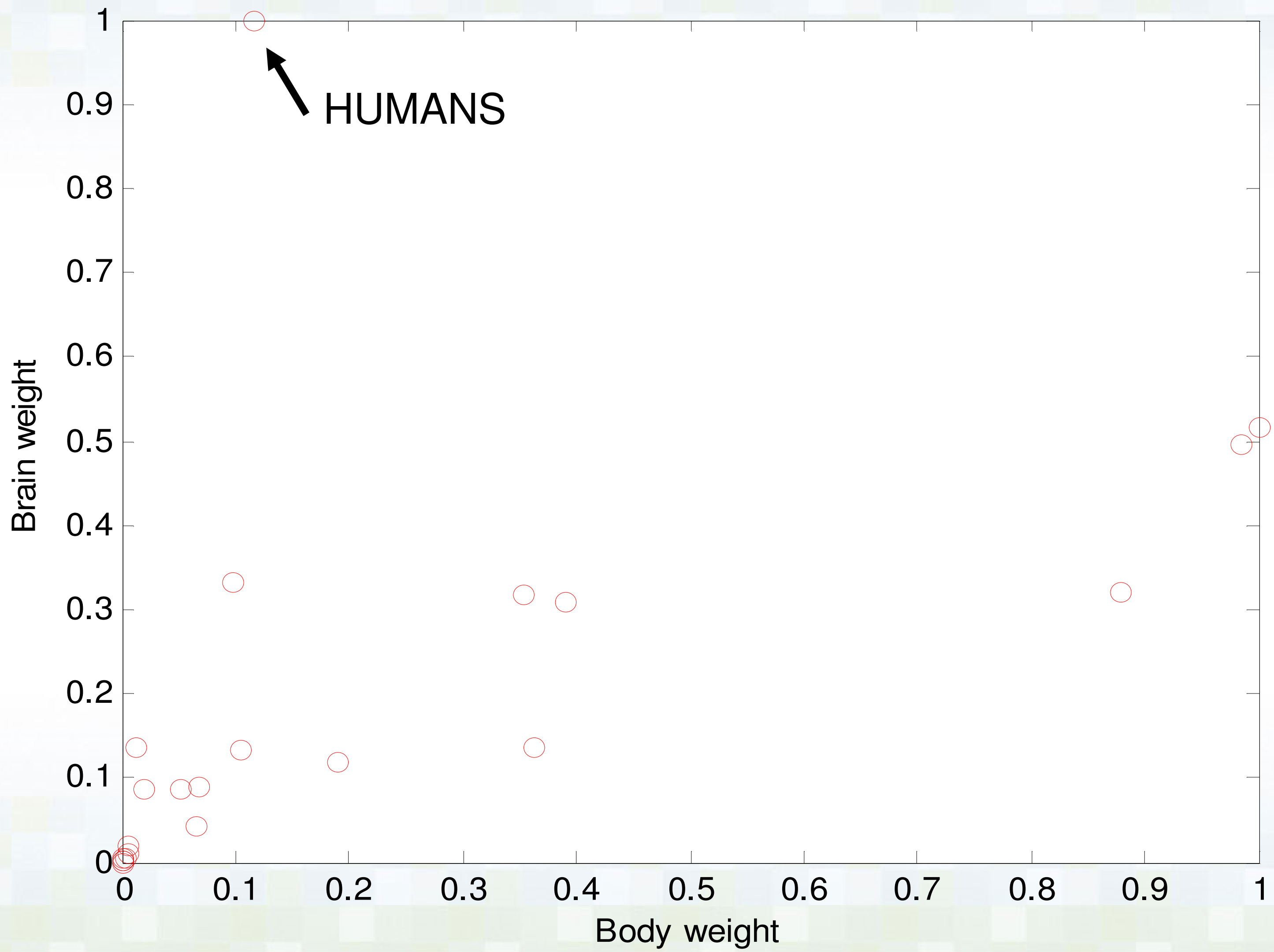
		kg	gram
1	Mountain beaver	1.350	8.100
2	Cow	465.000	423.000
3	Gray wolf	36.330	119.500
4	Goat	27.660	115.000
5	Guinea pig	1.040	5.500
6	Diplodocus	11700.000	50.000
7	Asian elephant	2547.000	4603.000
8	Donkey	187.100	419.000
9	Horse	521.000	655.000
10	Potar monkey	10.000	115.000
11	Cat	3.300	25.600
12	Giraffe	529.000	680.000
13	Gorilla	207.000	406.000
14	Human	62.000	1320.000

		kg	gram
15	African elephant	6654.000	5712.000
16	Triceratops	9400.000	70.000
17	Rhesus monkey	6.800	179.000
18	Kangaroo	35.000	56.000
19	Hamster	0.120	1.000
20	Mouse	0.023	0.400
21	Rabbit	2.500	12.100
22	Sheep	55.500	175.000
23	Jaguar	100.000	157.000
24	Chimpanzee	52.160	440.000
25	Brachiosaurus	87000.000	154.500
26	Rat	0.280	1.900
27	Mole	0.122	3.000
28	Pig	192.000	180.000

Original Unscaled Animal Data: *what points are outliers?*



Animal Data: with outliers removed and scaled to [0,1] range



Learning System

Two modes of operation: ("Goals of Learning")

1. Learning or estimation ("explanation of training data")
   - Goal is to select the "best' model (or function) from a large set of possible models
   - Training data is used for model selection
2. Test or prediction ("prediction of new (test) data")
   - An estimated model is used for predicting the outputs y for new (or *test*) inputs **x**

Many Kinds of Learning

- Supervised Learning
  - Training data are a set of n samples $(\boldsymbol{x}_i, y_i)$ used to estimate a model f(x)
  - Called *supervised learning* because the training data includes correct or "*ground truth*" output values (teacher)
  - Two types of supervised learning problems
    - Regression
    - Classification
  - Quality: as per goals of learning
    - Empirical Risk (avg. training error)
    - Prediction Risk (avg. test error)

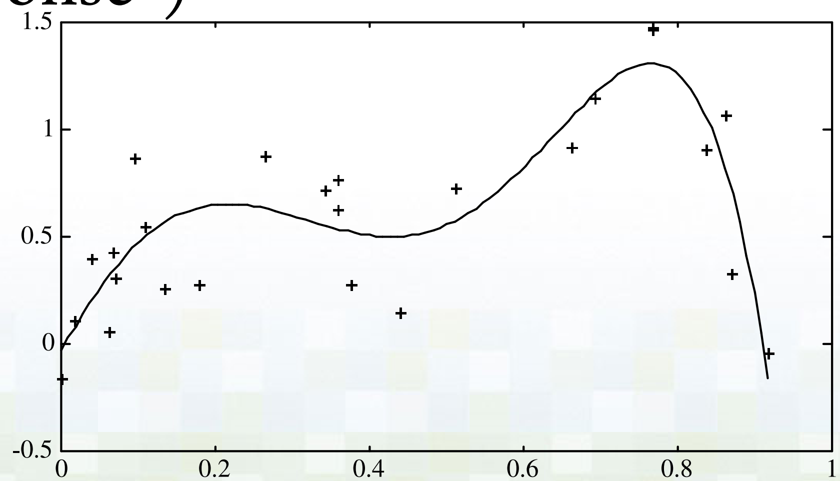
Many Kinds of Learning

- Unsupervised Learning
  - Available training data is in the form of n multivariate input samples $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ in d-dimensional sample space
  - These samples originate from unknown distribution
  - Called *unsupervised learning* because the output (response) values are not present in the data
  - Goal: approximate unknown distribution so samples produced by the approximation model are 'close' to samples from the generating distribution
  - Quality: approximation accuracy for training data only

Supervised Learning: Regression

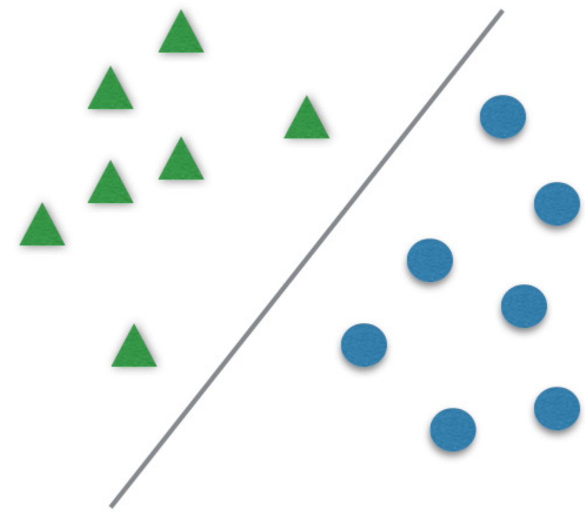
- The prediction task concerned with estimation of numeric (real-valued) outputs is called regression (real-valued function estimation problem)
- Data in the form $(\boldsymbol{x}, y)$ where
  - $\boldsymbol{x}$ is multivariate input (i.e. vector)
  - y is univariate output (“response”)
- **Regression**: y is real-valued

- Estimation of real-valued function $\boldsymbol{x} \rightarrow y$



Supervised Learning: Classification

- The prediction task concerned with estimation of categorical (class label) outputs is called classification
- Data in the form $(\boldsymbol{x}, y)$ where
  - $\boldsymbol{x}$ is multivariate input (i.e. vector)
  - y is univariate output ("response")
- **Classification**: y is categorical (class label)



- Estimation of indicator function $\boldsymbol{x} \to y$

Quality of Prediction: Supervised Learning

- Squared Loss
  - Regression: $L(y, f(x) = \left(y - f(x)\right)^2$
  - Classification (0/1 loss function):

$$L\left(y, f(x)\right) = \begin{cases} 0 \; if \; y = f(x) \\ 1 \; if \; y \neq f(x) \end{cases}$$

- Empirical Risk

  (*measure quality of explanation* – avg. training error)
  - $R_{emp} = \frac{1}{n}\sum_{i=1}^{n} L(y_i, f(x_i))$
- Prediction Risk

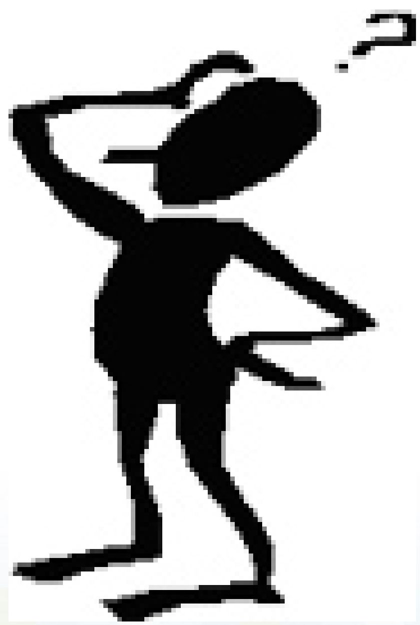
  (*measure quality of prediction* – avg. test error)
  - $R = \frac{1}{T}\sum_{t=1}^{T} L(y_i, f(x_i))$, T=# test samples

Question

- Does minimizing training error improve prediction accuracy on new or testing data??

Question

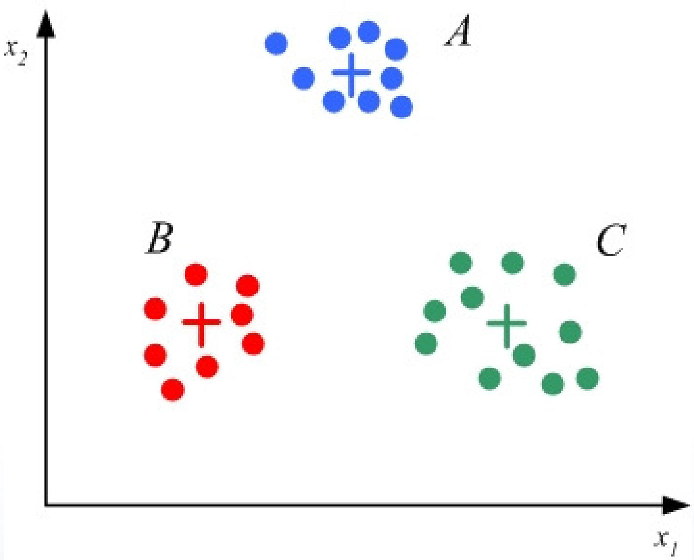
- Does minimizing training error improve prediction accuracy on new or testing data??



Generalization

Unsupervised Learning

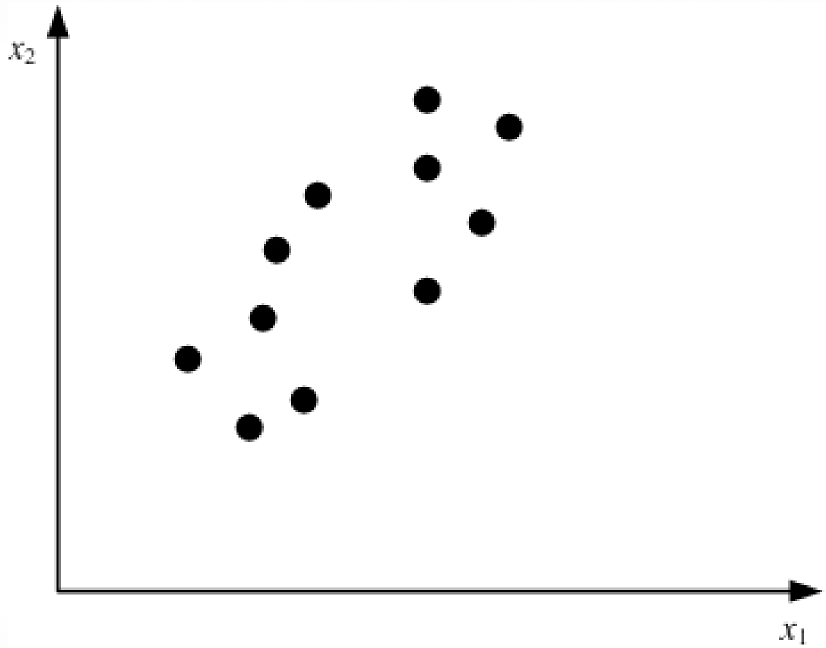
- Data in the form $(\boldsymbol{x}, y)$ where
  - $\boldsymbol{x}$ is multivariate input (i.e. vector)
- Goal 1: CLUSTERING or data reduction

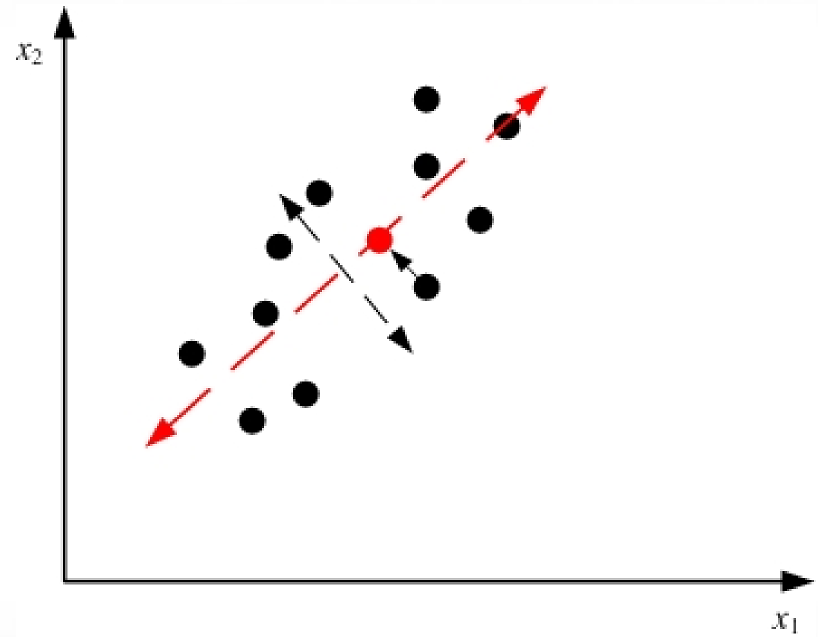


- Clustering = estimation of mapping $\boldsymbol{x} \rightarrow c$

Unsupervised Learning

- Goal 2: DIMENSIONALITY REDUCTION





- Finding low-dimensional model of the data
- (other: multiple model estimation)

Quality of Prediction: Unsupervised Learning

- Loss function
  - $L(y, f(x)) = \|x - f(x)\|^2$
  - The double bars is notation for distance
  - Goal: minimizing the squared distance between training points and their projections (mappings) onto a model space:

$$R_{emp} = \frac{1}{n}\sum_{i=1}^{n} L(x_i, f(x_i))$$

$$= \frac{1}{n}\sum_{i=1}^{n} \|x_i - f(x_i)\|^2$$

3.2 Basic Learning Approaches

Outline:

- Parametric Modeling
- Non-parametric Modeling
- Data Reduction

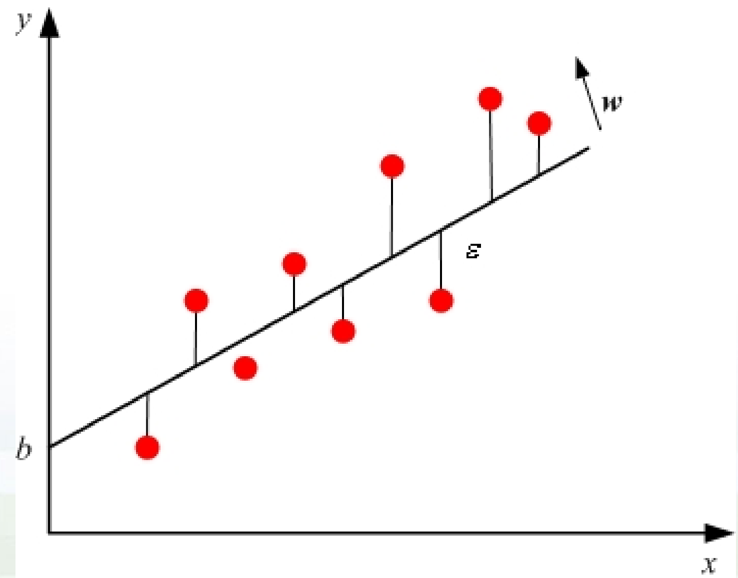
Parametric Modeling

- Given training data $(\boldsymbol{x}_i, y_i), i = 1,2, \ldots, n$

1. Specify parametric model
2. Estimate its parameters (via fitting to data)

- Example: Linear regression $F(\boldsymbol{x}) = (\boldsymbol{w} \,.\, \boldsymbol{x}) + b$

$$\sum_{i=1}^{n} [y_i - (w\, x_i) - b]^2 \rightarrow min$$



Parameters are estimated via minimization of the mean-squared-error fitting error for training data

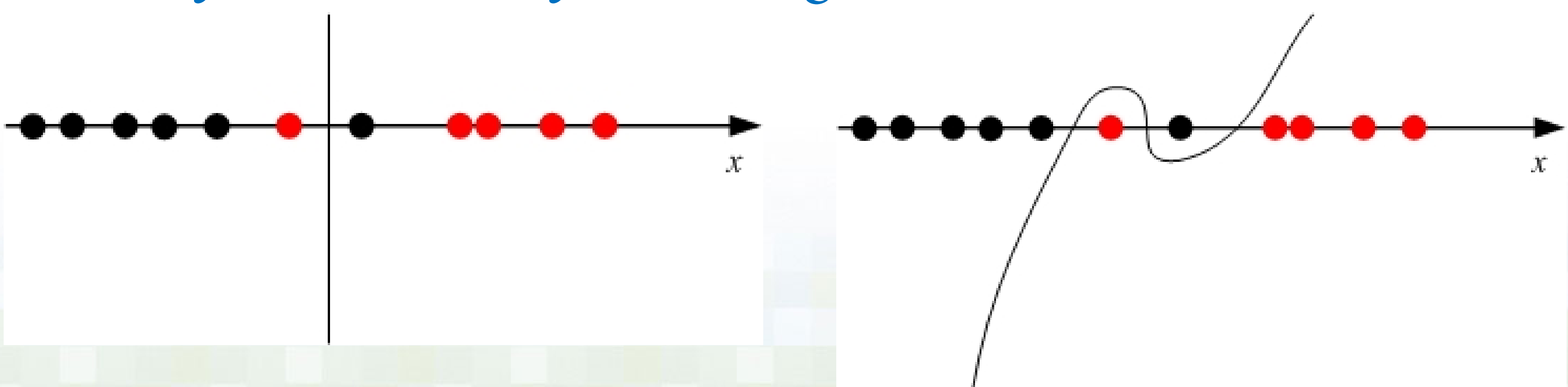
Parametric Modeling

- Given training data $(\boldsymbol{x}_i, y_i), i = 1,2, \dots, n$

1. Specify parametric model
2. Estimate its parameters (via fitting to data)

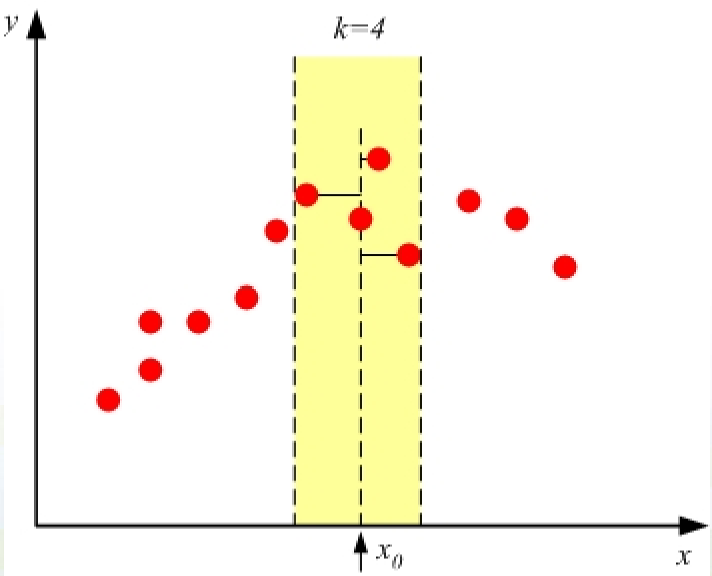
Univariate classification:

First order and third order model (decision boundary) with parameters estimated via minimization of empirical risk (classification error) for training data



Non-Parametric Modeling

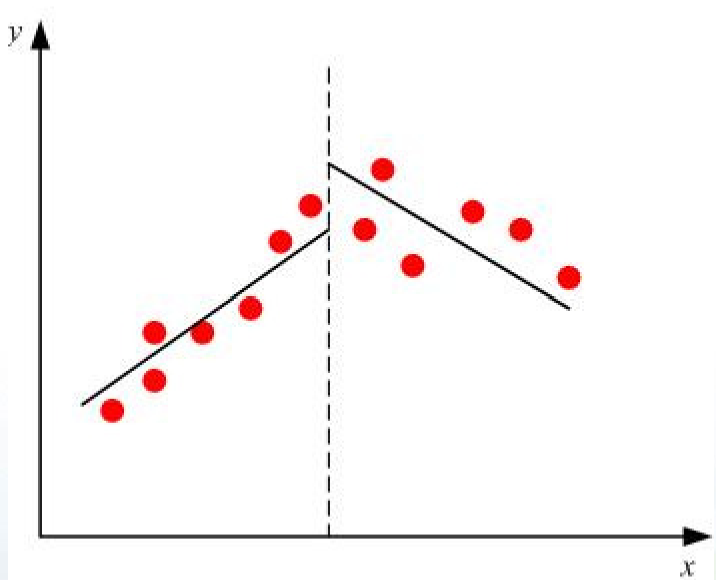
- Given training data $(\boldsymbol{x}_i, y_i), i = 1,2, \dots, n$
- Estimate the model (for given $\boldsymbol{x}_0$) as "local average" of the data ("local estimation modeling")
- Note: need to define "local" and "average"
- Example: k-nearest neighbors regression



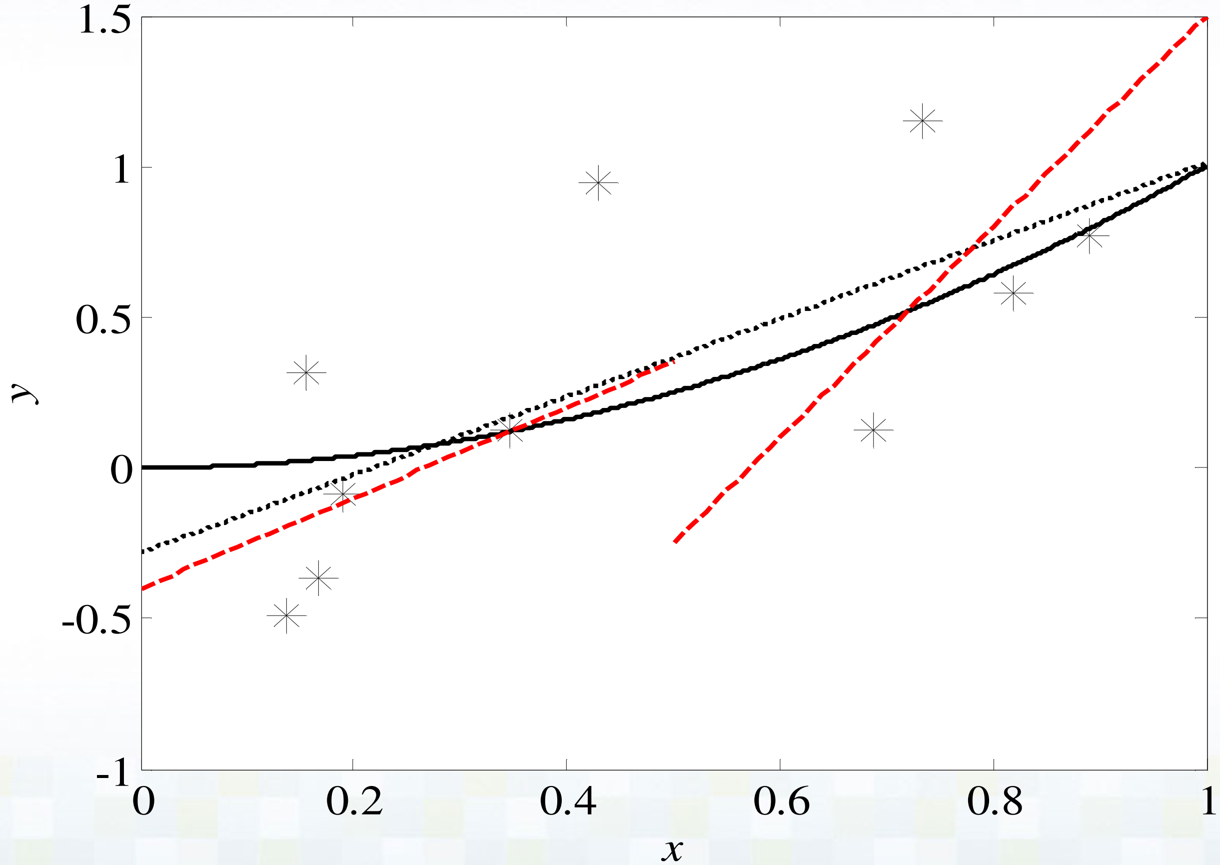
$$f(\boldsymbol{x}_0) = \frac{\sum_{j=1}^{k} y_j}{k}$$

Data Reduction Approach

- Given training data estimate the model as "compact encoding" of the data
- Note: "compact" ~ # of bits to encode the model
- Example: piece-wise linear regression



Example: piece-wise linear regression vs. linear regression

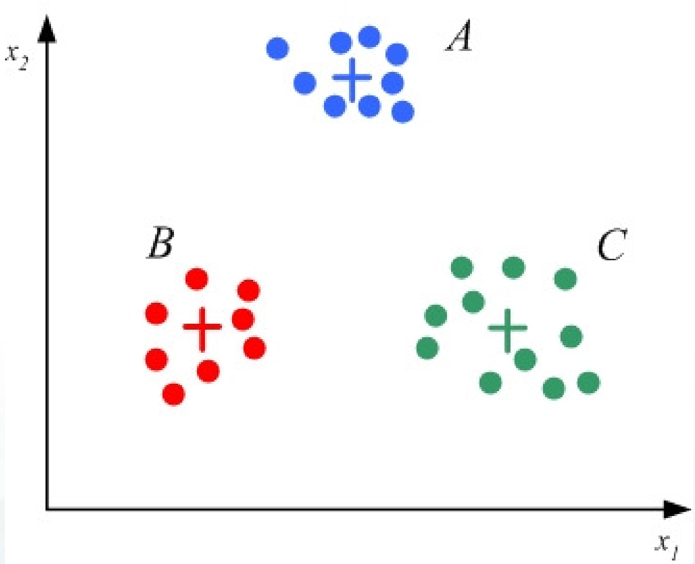


Data Reduction Approach (cont'd)

Data Reduction approaches are commonly used for unsupervised learning tasks.

Example: clustering.

Training data encoded by 3 points (cluster centers)



Issues:

- How to find centers?
- How to select the number of clusters?

Things to Think About

- Induction and Deduction in Philosophy:

  All observed swans are white (data samples).

  Therefore, all swans are white.
- Model estimation ~ inductive step, i.e. *estimate function from data samples.*
- Prediction ~ deductive step

  → **Inductive Learning Setting**
- Which of the 3 modeling approaches follow inductive learning?
- Do humans implement inductive inference?