Learning: Overview and Supervised Learning

Agents that learn rather than programming them by hand, and agents that learn to improve their performance over time.

Applications

- Medicine
  - diagnosis
- Text classification
  - spam filtering
- Game playing
  - move heuristics
- Vision
  - face recognition
- Speech
  - speech understanding
- Character recognition
  - handwriting recognition

How can “learning” be defined?

- Learning is an increase in “knowledge”
- Depends on definition of “knowledge”

Taxonomy of learning systems

- Rote learning (learning by being told)
  - memorize new facts (e.g., database)
- Speedup learning (learning by practice)
  - become more efficient over time (e.g., caching)
- Inductive learning (learning by generalizing)
  - make plausible inferences from partial information

Evaluating learning systems

- Rote learning (learning by being told)
  - evaluate by system's ability to exploit the new knowledge
- Speedup learning (learning by practice)
  - evaluate by measuring increase in efficiency
- Inductive learning (learning by generalizing)
  - evaluate by correctness of the new knowledge

Inductive learning:
Supervised learning of an unknown function *f*

- Input: Collection of training examples

  $((\bar{x}_1, y_1), \ldots, (\bar{x}_m, y_m))$

  for some unknown function $y_i = f(\bar{x}_i)$, where

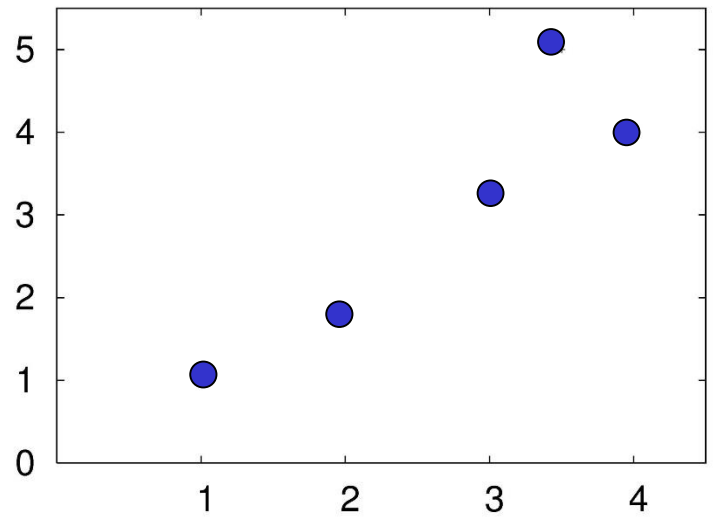
  $\bar{x}_i$ are tuples of features/attributes

  y_i is some outcome;

  Hypothesis space H

- Output: An hypothesis $h \in H$ that approximates f
  (a method of classifying subsequent, unseen examples)

Example: regression



Training set
(1, 1)
(2, 1.75)
(3, 3.25)
(3.5, 5)
(4, 4)

Possible hypothesis spaces H ?

Possible best hypothesis $h \in H$?

Jeeves is a valet to Bertie Wooster. On some days, Bertie likes to play tennis and asks Jeeves to lay out his tennis things and book the court. Jeeves would like to be able to predict whether Bertie will play tennis (and so be a better valet). Each morning over the last two weeks, Jeeves has recorded whether Bertie played tennis on that day and various attributes of the weather.

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Can Jeeves learn to predict Bertie's tennis playing?

Jeeves would like to evaluate the classifier he has come up with for predicting whether Bertie will play tennis. Each morning over the next two weeks, Jeeves records the following data.

Day	Outlook	Temp	Humidity	Wind	Tennis?
1	Sunny	Mild	High	Strong	No
2	Rain	Hot	Normal	Strong	No
3	Rain	Cool	High	Strong	No
4	Overcast	Hot	High	Strong	Yes
5	Overcast	Cool	Normal	Weak	Yes
6	Rain	Hot	High	Weak	Yes
7	Overcast	Mild	Normal	Weak	Yes
8	Overcast	Cool	High	Weak	Yes
9	Rain	Cool	High	Weak	Yes
10	Rain	Mild	Normal	Strong	No
11	Overcast	Mild	High	Weak	Yes
12	Sunny	Mild	Normal	Weak	Yes
13	Sunny	Cool	High	Strong	No
14	Sunny	Cool	High	Weak	No

How well does Jeeves predict Bertie's tennis playing?

Features / outcomes

- real-valued
- discrete, finite set
  - also called categorical or nominal
  - binary
  - ordered
  - unordered

- outcome is real-valued
  - regression
- outcome is discrete, finite set
  - classification

Outline

- Issues in learning
- Learning Bayesian networks
- Learning decision trees
- Learning neural networks
- Ensembles of classifiers
- Summary

Issues in learning functions: Choice of training data

- Is it representative?
- Choice of features/attributes
- Constructing new features
- Balance of positive and negative examples

Issues in learning functions: Choice of hypothesis space *H*

- Cost of learning the hypothesis *h*
- Amount of training data needed to learn *h*
- Does representation make sense given domain knowledge and data (is it capable of fitting / representing the data?)

Issues in learning functions: Choice of hypothesis $h \in H$

- Simplicity of learned function h (Ockham's razor)
- Performance on training set
- Performance on test set (i.e., generalization)
- Judging performance (measuring errors)

Issues in learning functions: Judging performance (measuring errors)

- Collection of training examples (or test examples)

  $((\bar{x}_1, y_1), \ldots, (\bar{x}_m, y_m))$

  for some unknown function $y_i = f(\bar{x}_i)$
- Minimize total error

$$\sum_{i=1}^{m} \text{error}_i$$

where error_i is the error on the i^{th} example

Issues in learning functions: Judging performance (measuring errors)

Discrete or classification error

$$\text{error}_i = \begin{cases} 0 & \text{if } h(\bar{x}_i) = y_i \\ 1 & \text{if } h(\bar{x}_i) \neq y_i \end{cases}$$

	y_i = no	y_i = yes
$h(\bar{x}_i)$ = no	0	1
$h(\bar{x}_i)$ = yes	1	0

false negative

false positive

Issues in learning functions: Judging performance (measuring errors)

Continuous error

$$\text{error}_i = |\, h(\bar{x}_i) - y_i \,| \qquad (L_1 \text{ norm})$$

or

$$\text{error}_i = (h(\bar{x}_i) - y_i)^2 \qquad (L_2 \text{ norm})$$

Note: if total error is 0, *h* is said to be *consistent* with the data.

Issues in learning functions: Choice of hypothesis $h \in H$ (con't)

- Defn: Given a hypothesis space H, a hypothesis h is said to <u>overfit</u> the training data if there exists some hypothesis $h' \in H$ such that

  h has smaller error than h' over the training set

  <u>but</u> h' has smaller error than h over the entire distributions of instances