Machine Learning

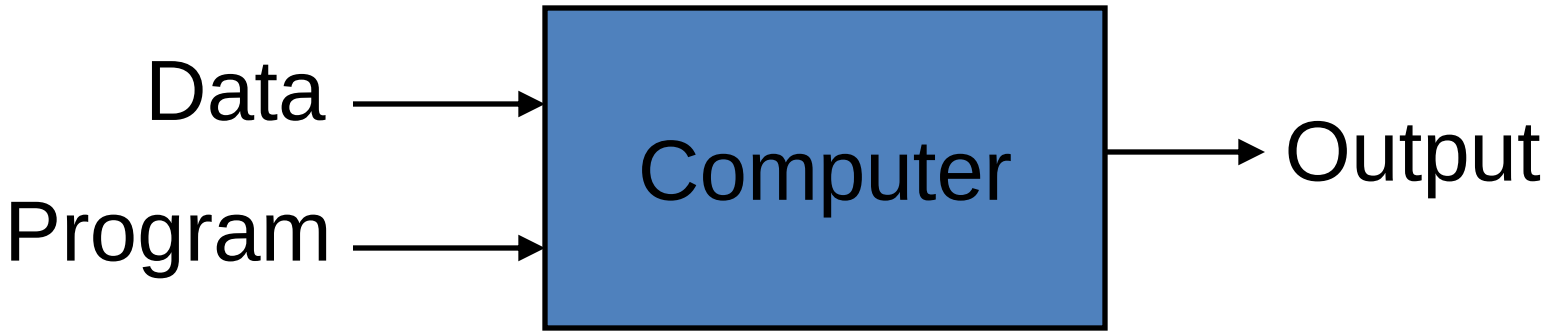
Mausam

(based on slides by Tom Mitchell, Oren Etzioni and Pedro Domingos)

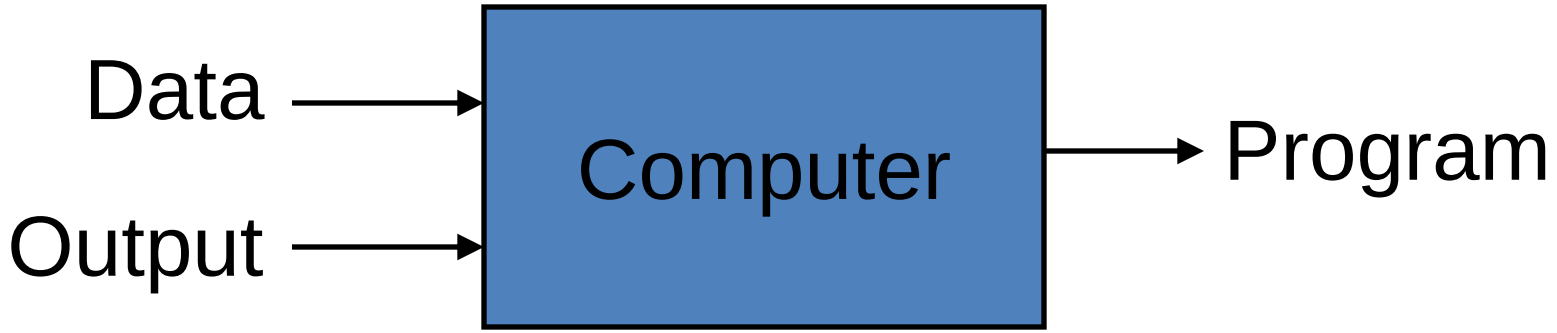
What Is Machine Learning?

A computer program is said to learn from experience E with respect to some class of tasks T and a performance measure P if it improves performance on T (according to P) with more E.

Traditional Programming



Machine Learning



Why Bother with Machine Learning?

Btw, Machine Learning ~ Data Mining

- Necessary for AI
- Learn concepts that people don't have time for ("drowning in data…starved for knowledge")
- Mass customization (adapt software to each)
- Super-human learning/discovery

Quotes

- “A break through in machine learning would be worth ten Microsofts” (Bill Gates)
- “Machine learning is the next Internet” (Tony Tether, Former Director, DARPA)
- Machine learning is the hot new thing” (John Hennessy, President, Stanford)
- “Web rankings today are mostly a matter of machine learning” (Prabhakar Raghavan, Dir. Research, Yahoo)
- “Machine learning is going to result in a real revolution” (Greg Papadopoulos, CTO, Sun)

Inductive Learning

- **Given** examples of a function *(X, F(X))*
- **Predict** function *F(X)* for new examples *X*
  - Discrete *F(X)*: Classification
  - Continuous *F(X)*: Regression
  - *F(X)* = Probability(*X*): Probability estimation

Training Data Versus Test

- Terms: ‘data’, ‘examples’, and ‘instances’ used interchangeably

- Training data: data where the labels are given
- Test data: data where the labels are known but not given

Which do you use to measure performance?

Cross validation…

Basic Setup

- **Input:**
  - Labeled training examples
  - Hypothesis space H

- **Output:** hypothesis h in H that is consistent with the training data & (hopefully) correctly classifies test data.

The ‘new’ Machine Learning

Old	New
Small data sets (100s of examples)	Massive (10^6 to 10^10)
Hand-labeled data	Automatically labeled; semi supervised; labeled by “crowds”
Hand-coded algorithms	WEKA package downloaded over 1,000,000 times

ML in a Nutshell

- 10^5 machine learning algorithms
- Hundreds new every year
- Every algorithm has three components:
  - **Hypothesis space—possible outputs**
  - **Search strategy---strategy for exploring space**
  - **Evaluation**

Hypothesis Space (Representation)

- Decision trees
- Sets of rules / Logic programs
- Instances
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- Etc.

Metrics for Evaluation

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Etc.

Based on Data

Search Strategy

- Greedy (depth-first, best-first, hill climbing)
- Exhaustive
- Optimize an objective function
- More…

Types of Learning

- **Supervised (inductive) learning**
  - Training data includes desired outputs
- **Unsupervised learning**
  - Training data does not include desired outputs
- **Semi-supervised learning**
  - Training data includes a few desired outputs
- **Reinforcement learning**
  - Rewards from sequence of actions

Supervised Learning

- **Given:** Training examples $\langle \mathbf{x}, f(\mathbf{x}) \rangle$ for some unknown function f.
- **Find:** A good approximation to f.

Example Applications

- **Credit risk assessment**

  $\mathbf{x}$: Properties of customer and proposed purchase.

  $f(\mathbf{x})$: Approve purchase or not.

- **Disease diagnosis**

  $\mathbf{x}$: Properties of patient (symptoms, lab tests)

  $f(\mathbf{x})$: Disease (or maybe, recommended therapy)

- **Face recognition**

  $\mathbf{x}$: Bitmap picture of person's face

  $f(\mathbf{x})$: Name of the person.

- **Automatic Steering**

  $\mathbf{x}$: Bitmap picture of road surface in front of car.

  $f(\mathbf{x})$: Degrees to turn the steering wheel.

Appropriate Applications for Supervised Learning

- **Situations where there is no human expert**

  $\mathbf{x}$: Bond graph for a new molecule.

  $f(\mathbf{x})$: Predicted binding strength to AIDS protease molecule.

- **Situations where humans can perform the task but can't describe how they do it.**

  $\mathbf{x}$: Bitmap picture of hand-written character

  $f(\mathbf{x})$: Ascii code of the character

- **Situations where the desired function is changing frequently**

  $\mathbf{x}$: Description of stock prices and trades for last 10 days.

  $f(\mathbf{x})$: Recommended stock transactions

- **Situations where each user needs a customized function f**

  $\mathbf{x}$: Incoming email message.

  $f(\mathbf{x})$: Importance score for presenting to user (or deleting without presenting).

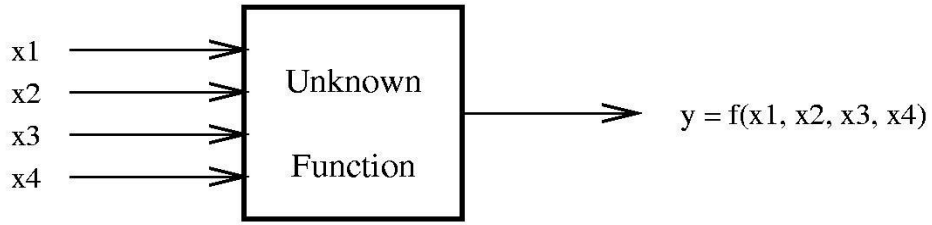
Why Learning?

- Learning is essential for unknown environments
  - e.g., when designer lacks omniscience
- Learning is necessary in dynamic environments
  - Agent can adapt to changes in environment not foreseen at design time
- Learning is useful as a system construction method
  - Expose the agent to reality rather than trying to approximate it through equations etc.
- Learning modifies the agent's decision mechanisms to improve performance

Terminology

- **Training example.** An example of the form $\langle \mathbf{x}, f(\mathbf{x}) \rangle$.
- **Target function (target concept).** The true function f.
- **Hypothesis**. A proposed function h believed to be similar to f.
- **Concept**. A boolean function. Examples for which $f(\mathbf{x}) = 1$ are called **positive examples** or **positive instances** of the concept. Examples for which $f(\mathbf{x}) = 0$ are called **negative examples** or **negative instances.**
- **Classifier**. A discrete-valued function. The possible values $f(\mathbf{x}) \in \{1, \ldots, K\}$ are called the **classes** or **class labels**.
- **Hypothesis Space**. The space of all hypotheses that can, in principle, be output by a learning algorithm.
- **Version Space**. The space of all hypotheses in the hypothesis space that have not yet been ruled out by a training example.

A Learning Problem



Example	x_1	x_2	x_3	x_4	y
1	0	0	1	0	0
2	0	1	0	0	0
3	0	0	1	1	1
4	1	0	0	1	1
5	0	1	1	0	0
6	1	1	0	0	0
7	0	1	0	1	0

Hypothesis Spaces

- **Complete Ignorance.** There are $2^{16} = 65536$ possible boolean functions over four input features. We can't figure out which one is correct until we've seen every possible input-output pair. After 7 examples, we still have 2^9 possibilities.

x_1	x_2	x_3	x_4	y
0	0	0	0	?
0	0	0	1	?
0	0	1	0	0
0	0	1	1	1
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	?
1	0	0	0	?
1	0	0	1	1
1	0	1	0	?
1	0	1	1	?
1	1	0	0	0
1	1	0	1	?
1	1	1	0	?
1	1	1	1	?

Hypothesis Spaces (2)

- **Simple Rules.** There are only 16 simple conjunctive rules.

Rule	Counterexample
$\Rightarrow y$	1
$x_1 \Rightarrow y$	3
$x_2 \Rightarrow y$	2
$x_3 \Rightarrow y$	1
$x_4 \Rightarrow y$	7
$x_1 \wedge x_2 \Rightarrow y$	3
$x_1 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \Rightarrow y$	3
$x_2 \wedge x_4 \Rightarrow y$	3
$x_3 \wedge x_4 \Rightarrow y$	4
$x_1 \wedge x_2 \wedge x_3 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3
$x_1 \wedge x_2 \wedge x_3 \wedge x_4 \Rightarrow y$	3

No simple rule explains the data. The same is true for simple clauses.

Hypothesis Space (3)

- m-**of-**n **rules**. There are 32 possible rules (includes simple conjunctions and clauses).

	Counterexample			
variables	1-of	2-of	3-of	4-of
$\{x_1\}$	3	–	–	–
$\{x_2\}$	2	–	–	–
$\{x_3\}$	1	–	–	–
$\{x_4\}$	7	–	–	–
$\{x_1, x_2\}$	3	3	–	–
$\{x_1, x_3\}$	4	3	–	–
$\{x_1, x_4\}$	6	3	–	–
$\{x_2, x_3\}$	2	3	–	–
$\{x_2, x_4\}$	2	3	–	–
$\{x_3, x_4\}$	4	4	–	–
$\{x_1, x_2, x_3\}$	1	3	3	–
$\{x_1, x_2, x_4\}$	2	3	3	–
$\{x_1, x_3, x_4\}$	1	***	3	–
$\{x_2, x_3, x_4\}$	1	5	3	–
$\{x_1, x_2, x_3, x_4\}$	1	5	3	3

Inductive Bias

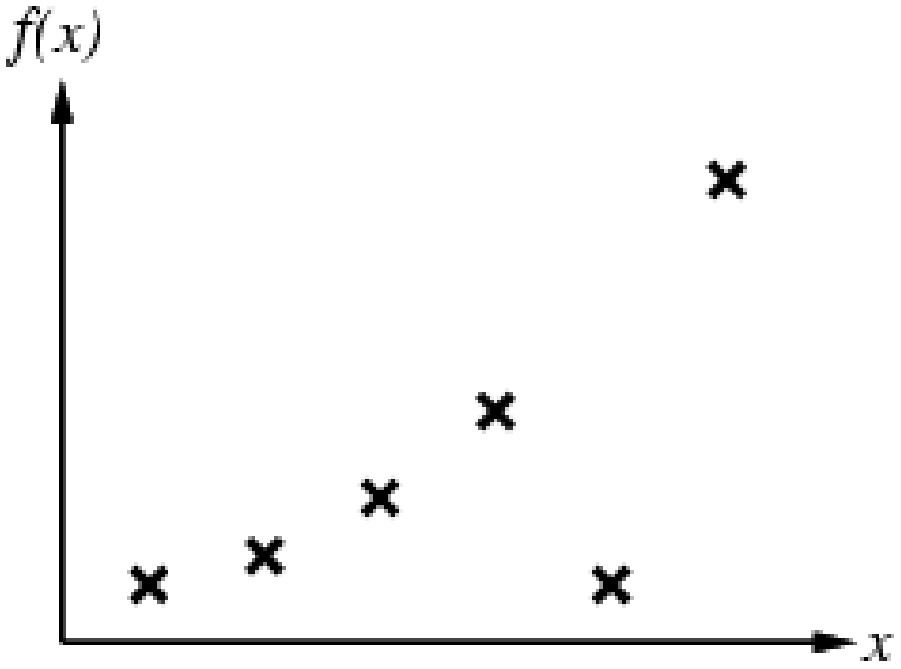
- Need to make assumptions
  - Experience alone doesn't allow us to make conclusions about unseen data instances

- Two types of bias:
  - Restriction: Limit the hypothesis space

    (e.g., naïve Bayes)
  - Preference: Impose ordering on hypothesis space (e.g., decision tree)

Inductive learning example

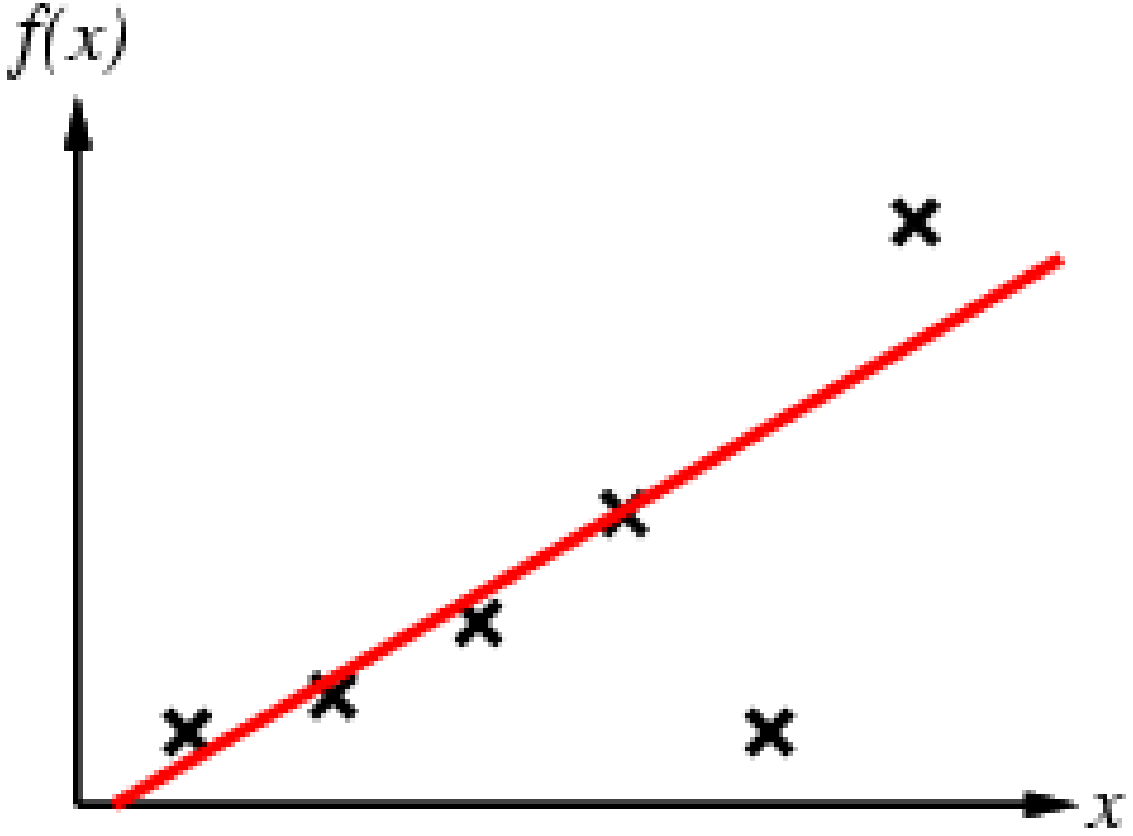
- Construct *h* to agree with *f* on training set
  - *h* is consistent if it agrees with *f* on all training examples
- E.g., curve fitting (regression):



x = Input data point (training example)

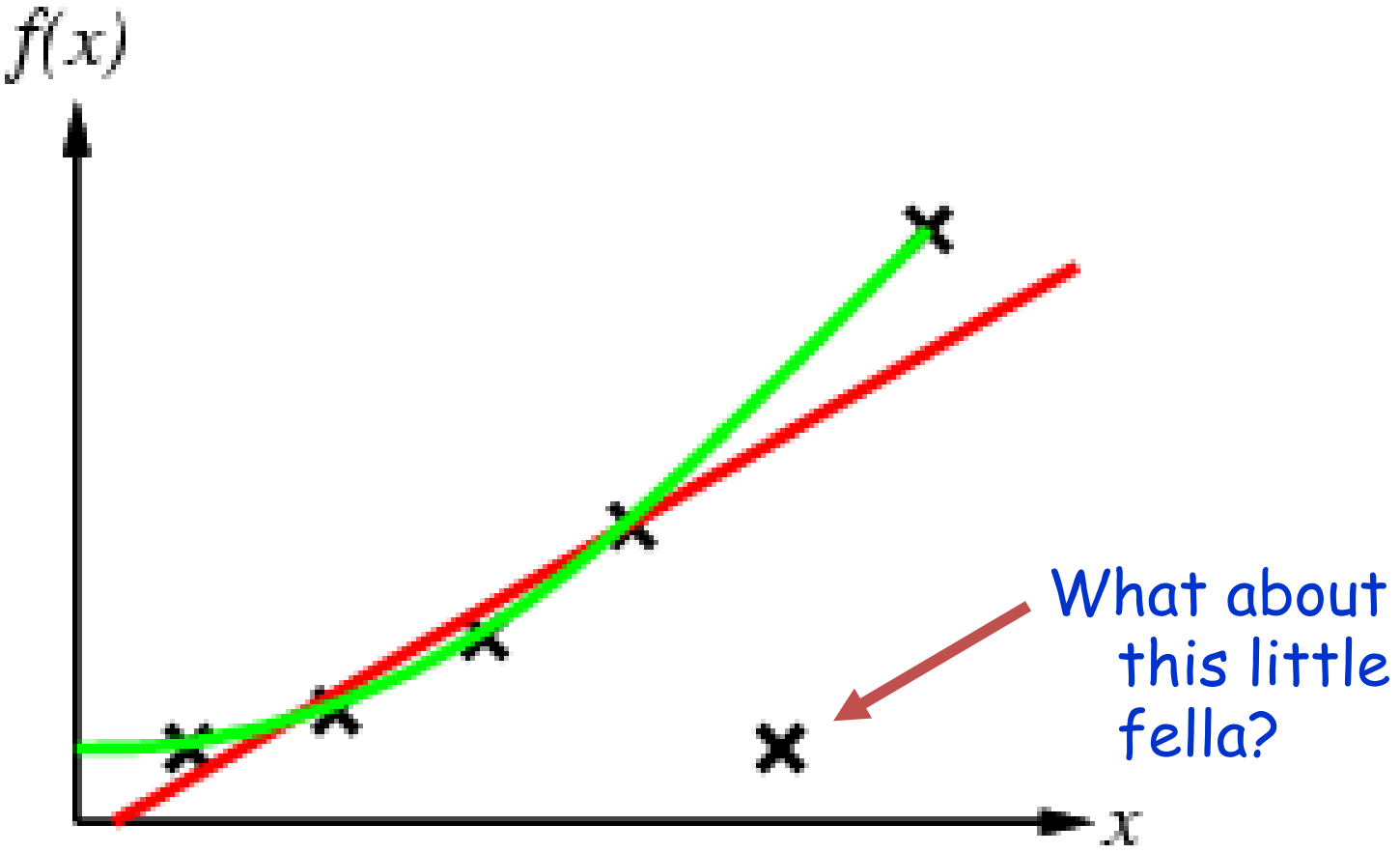
Inductive learning example

- *h* = Straight line?



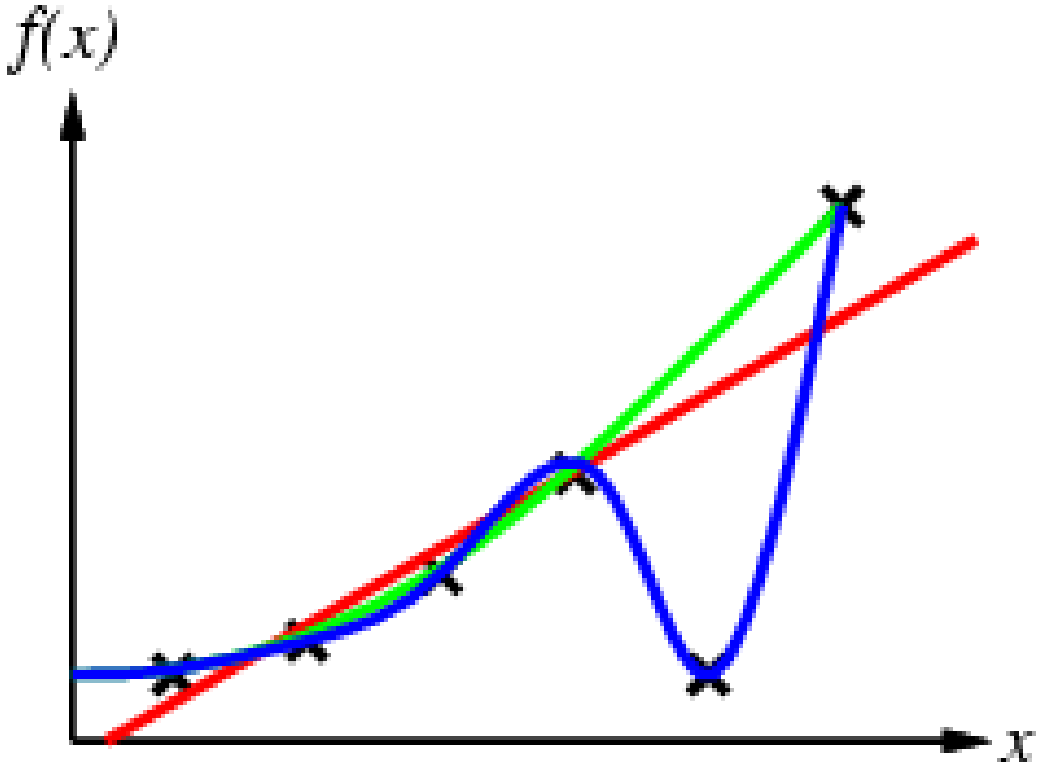
Inductive learning example

- What about a quadratic function?



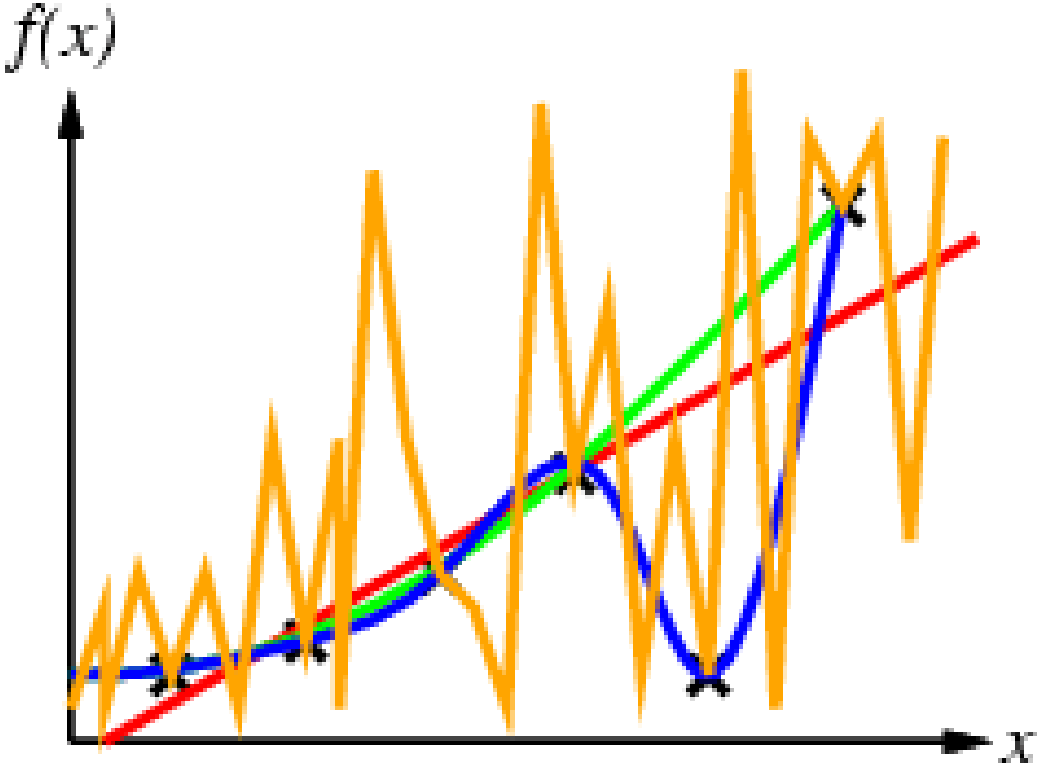
Inductive learning example

Finally, a function that satisfies all!

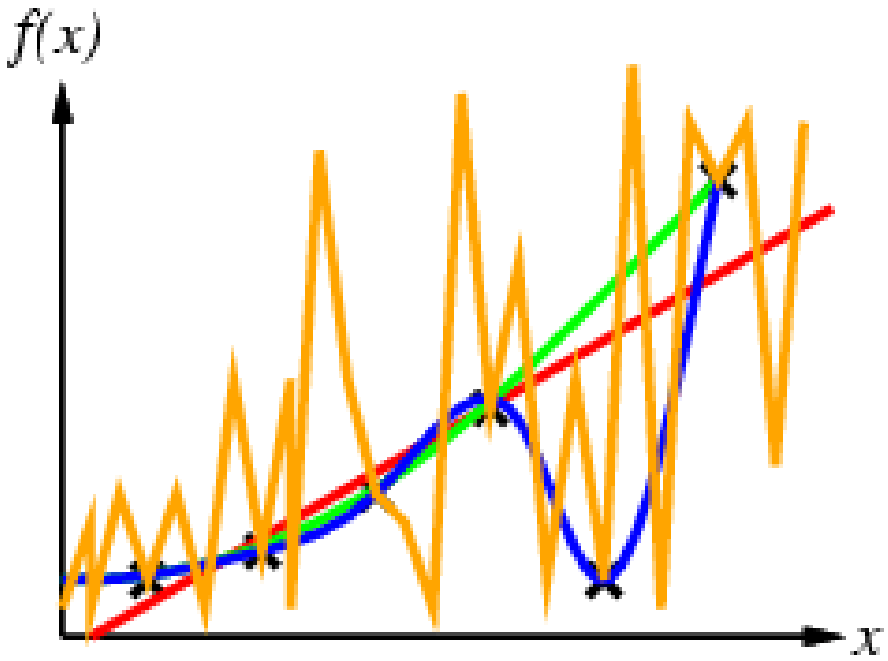


Inductive learning example

- But so does this one…



Ockham's Razor Principle



Ockham's razor: prefer the simplest hypothesis consistent with data

Related to KISS principle ("keep it simple stupid")

Smooth blue function preferable over wiggly yellow one

If noise known to exist in this data, even linear might be better (the lowest x might be due to noise)