CSC 665: Artificial Intelligence

Machine Learning: Intro

Instructor: Pooyan Fazli
San Francisco State University

Some slides borrowed from Andrew Ng

Machine Learning definition

- Arthur Samuel (1959). Machine Learning: Field of study that gives computers the ability to learn without being explicitly programmed.

Machine Learning definition

- Well-posed Learning Problem: A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E. (Tom Mitchell, 1998)

Machine Learning definition

"A computer program is said to *learn* from experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E."

Suppose your email program watches which emails you do or do not mark as spam, and based on that learns how to better filter spam. What is the task T in this setting?

- ✓ Classifying emails as spam or not spam.
- Watching you label emails as spam or not spam.
- The number (or fraction) of emails correctly classified as spam/not spam.
- None of the above—this is not a machine learning problem.

Machine Learning Algorithms

Machine learning algorithms:

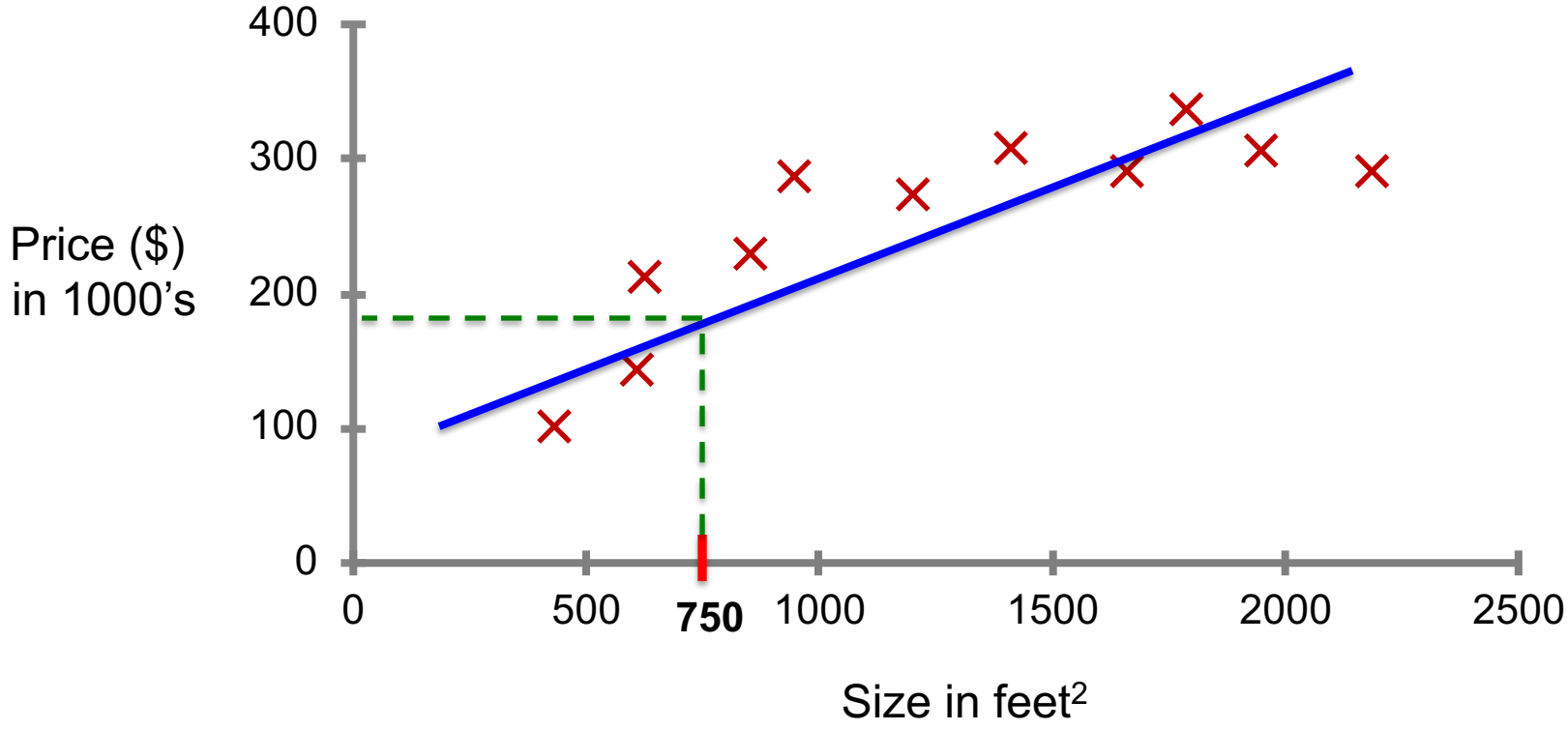
- Supervised learning
- Unsupervised learning
- ✓ Reinforcement learning

Supervised Learning

Supervised Learning

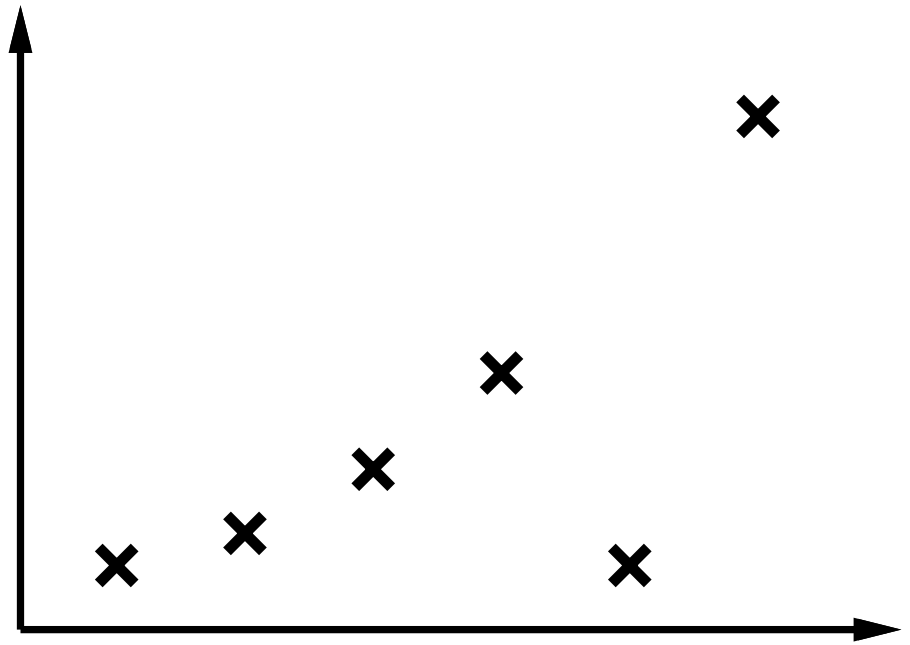
Supervised Learning: “right answers” given

Regression

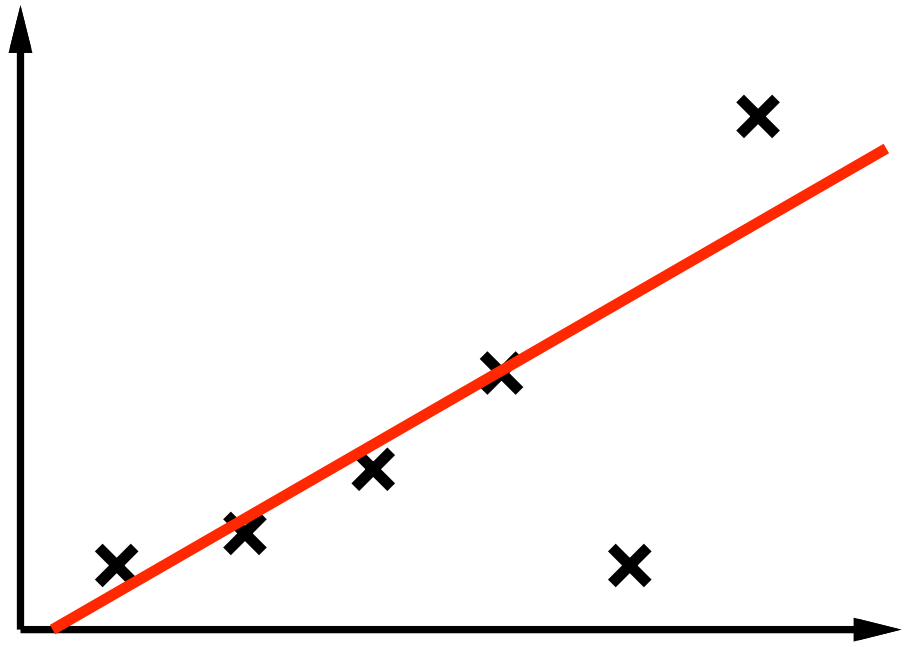


Regression: Predict continuous valued output (price)

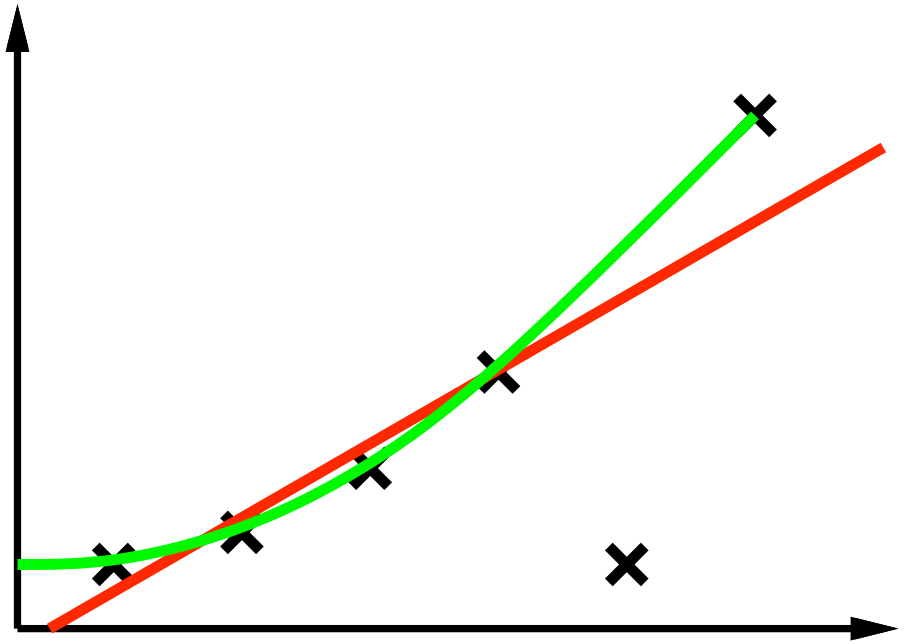
Regression example: Curve fitting



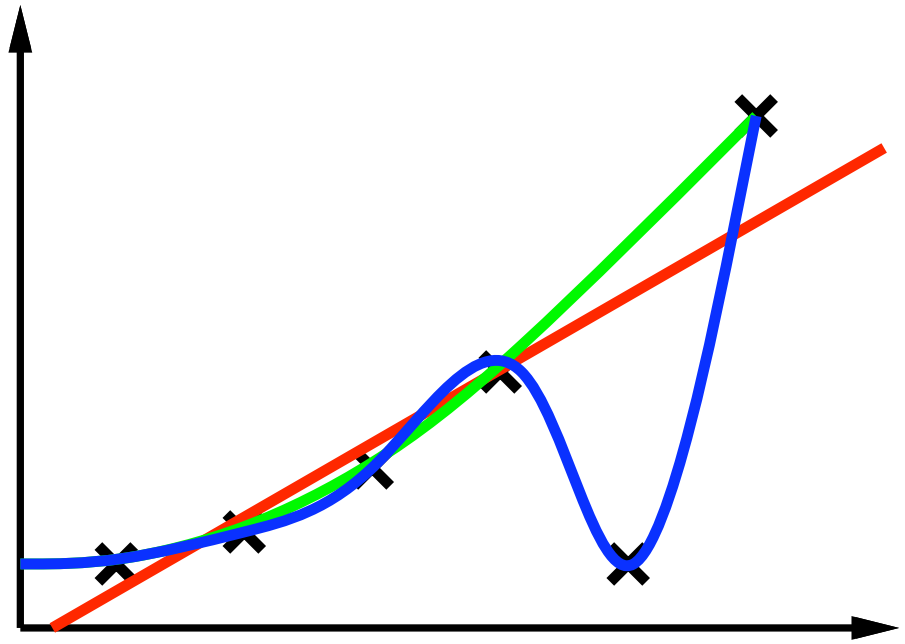
Regression example: Curve fitting



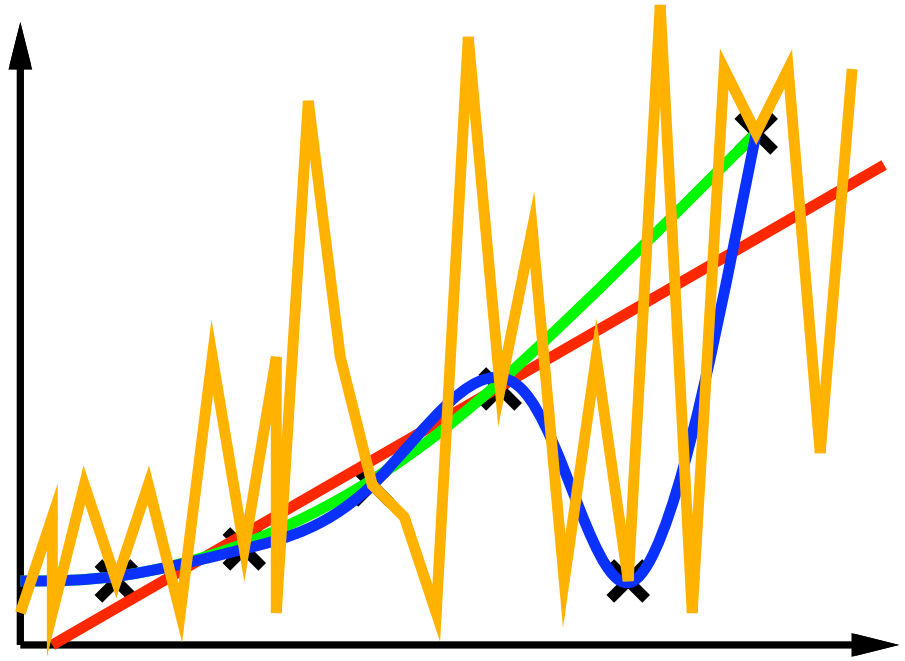
Regression example: Curve fitting



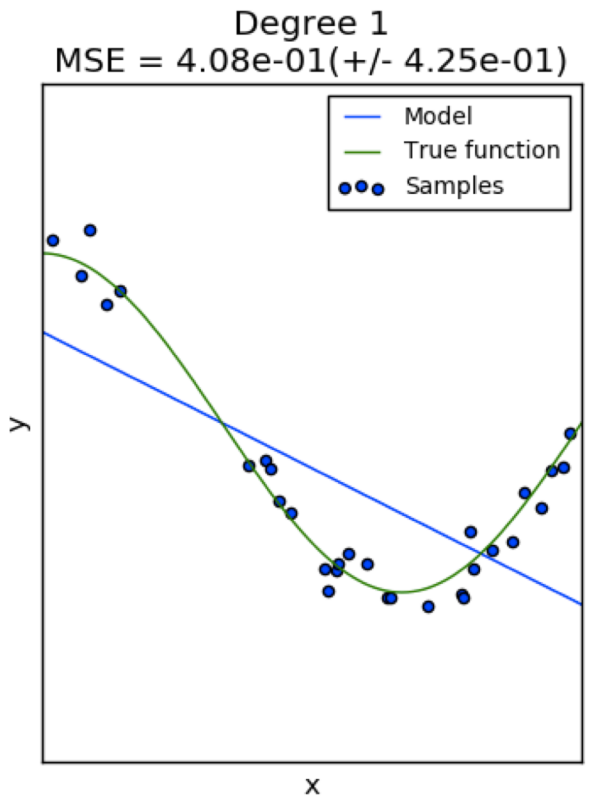
Regression example: Curve fitting

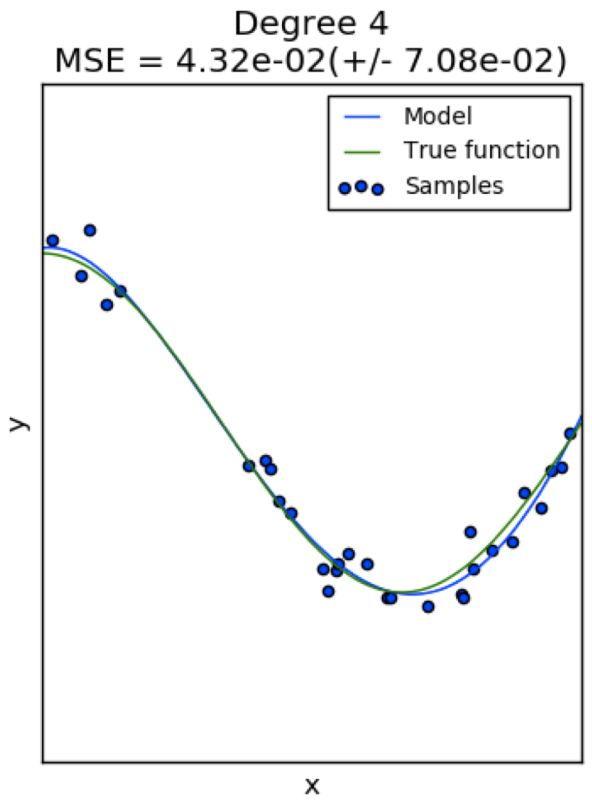


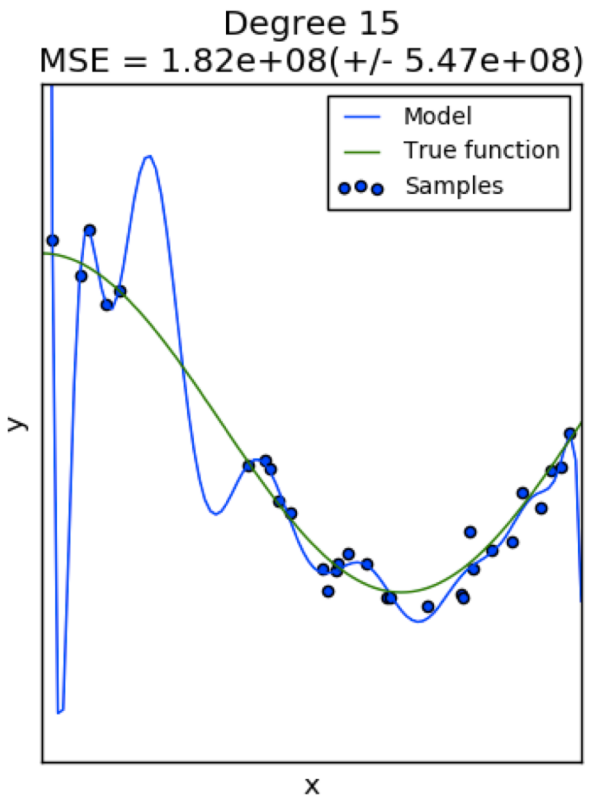
Regression example: Curve fitting



Underfitting vs Overfitting







Underfitting

Overfitting

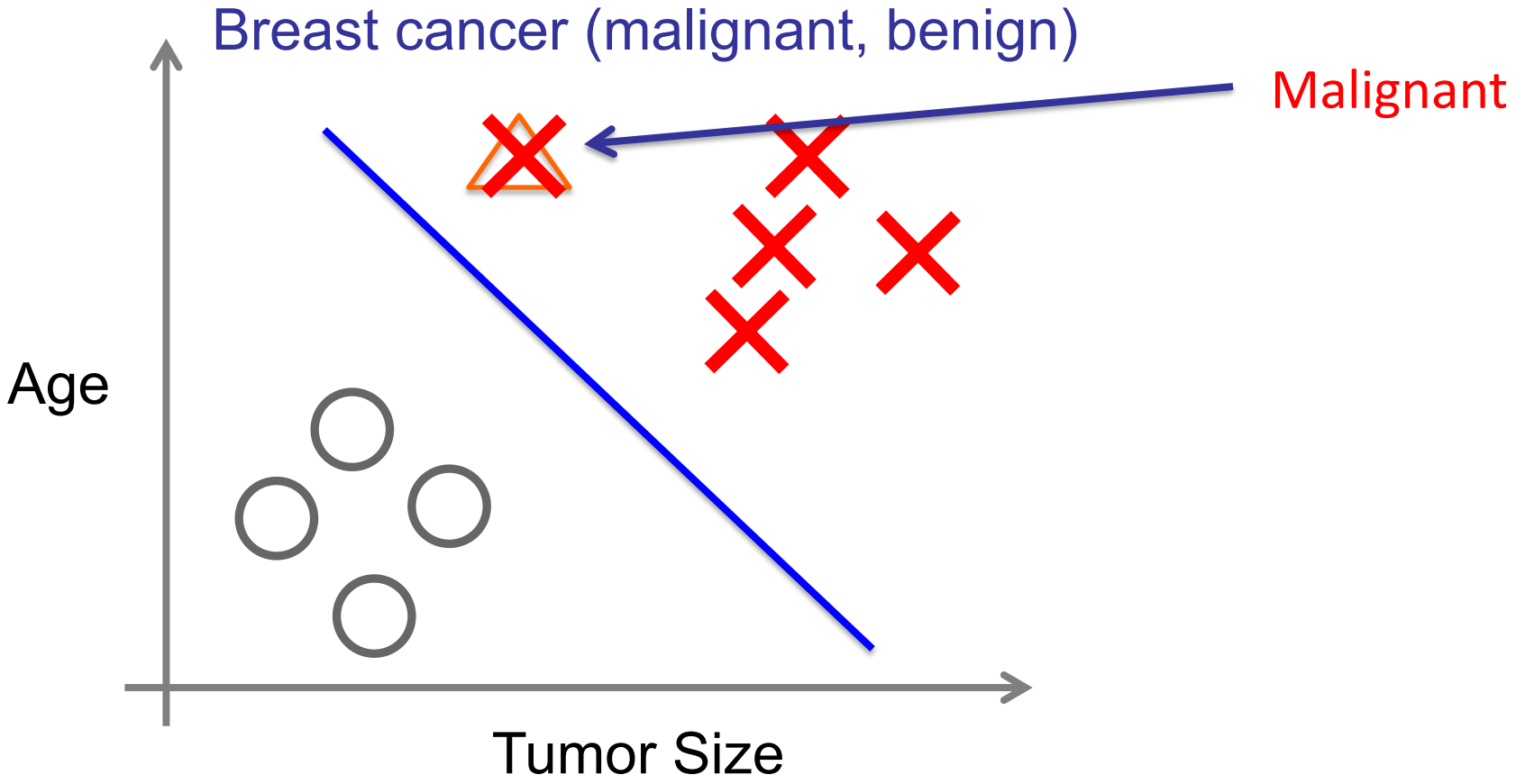
Image Ref: scikit-learn

Underfitting vs Overfitting

- Overfitting means the model is too complex/strong for the limited amount of data. It fits the training data very closely, but does not generalize well.

- Underfitting means the model is too simple/weak to express the structure of the data, even can't fit the training data.

Classification



Classification: Discrete valued output (0 or 1, malignant or benign)

Example: Image Classification

x      

f(x) giraffe giraffe giraffe dog dog dog

X= 

f(x)=?

Example: Spam Filter

- Input: an email
- Output: spam/ham

- Setup:
  - Get a large collection of example emails, each labeled “spam” or “ham” (by hand)
  - Want to Learn to predict labels of new incoming emails
  - Classifiers reject 200 billion spam emails per day

- Features: The attributes used to make the ham / spam decision
  - Words: FREE!
  - Text Patterns: $dd, CAPS
  - Non-text: SenderInContacts, …
  - …

Dear Sir.

First, I must solicit your confidence in this transaction, this is by virture of its nature as being utterly confidencial and top secret. …

TO BE REMOVED FROM FUTURE MAILINGS, SIMPLY REPLY TO THIS MESSAGE AND PUT "REMOVE" IN THE SUBJECT.

99 MILLION EMAIL ADDRESSES
FOR ONLY $99

Ok, Iknow this is blatantly OT but I'm beginning to go insane. Had an old Dell Dimension XPS sitting in the corner and decided to put it to use, I know it was working pre being stuck in the corner, but when I plugged it in, hit the power nothing happened.

Example: Digit Recognition

- Input: images / pixel grids
- Output: a digit 0-9

- Setup:
  - Get a large collection of example images, each labeled with a digit
    - Note: someone has to hand label all this data!
  - Want to learn to predict labels of new digit images

- Features: The attributes used to make the digit decision
  - Pixels: (6,8)=ON
  - Shape Patterns: NumComponents, AspectRatio, NumLoops
  - …


0


1


2


1


??

Other Classification Tasks

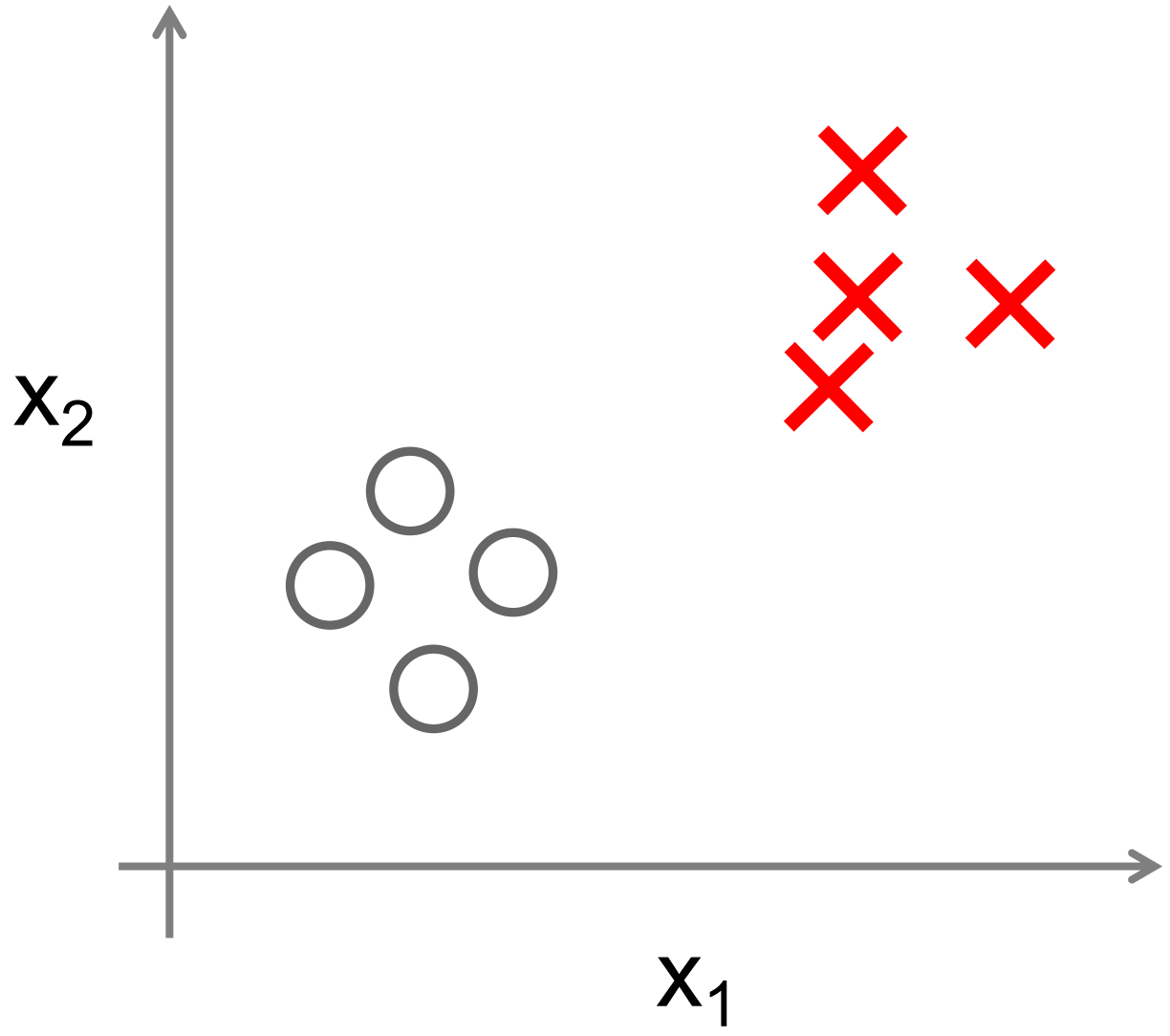
- Medical diagnosis
  - input: symptoms
  - output: disease
- Fraud detection
  - input: account activity
  - output: fraud / no fraud
- Email routing
  - input: customer complaint email
  - output: which department needs to ignore this email
- Fruit and vegetable inspection
  - input: image (or gas analysis)
  - output: moldy or OK

- … many more

Unsupervised Learning

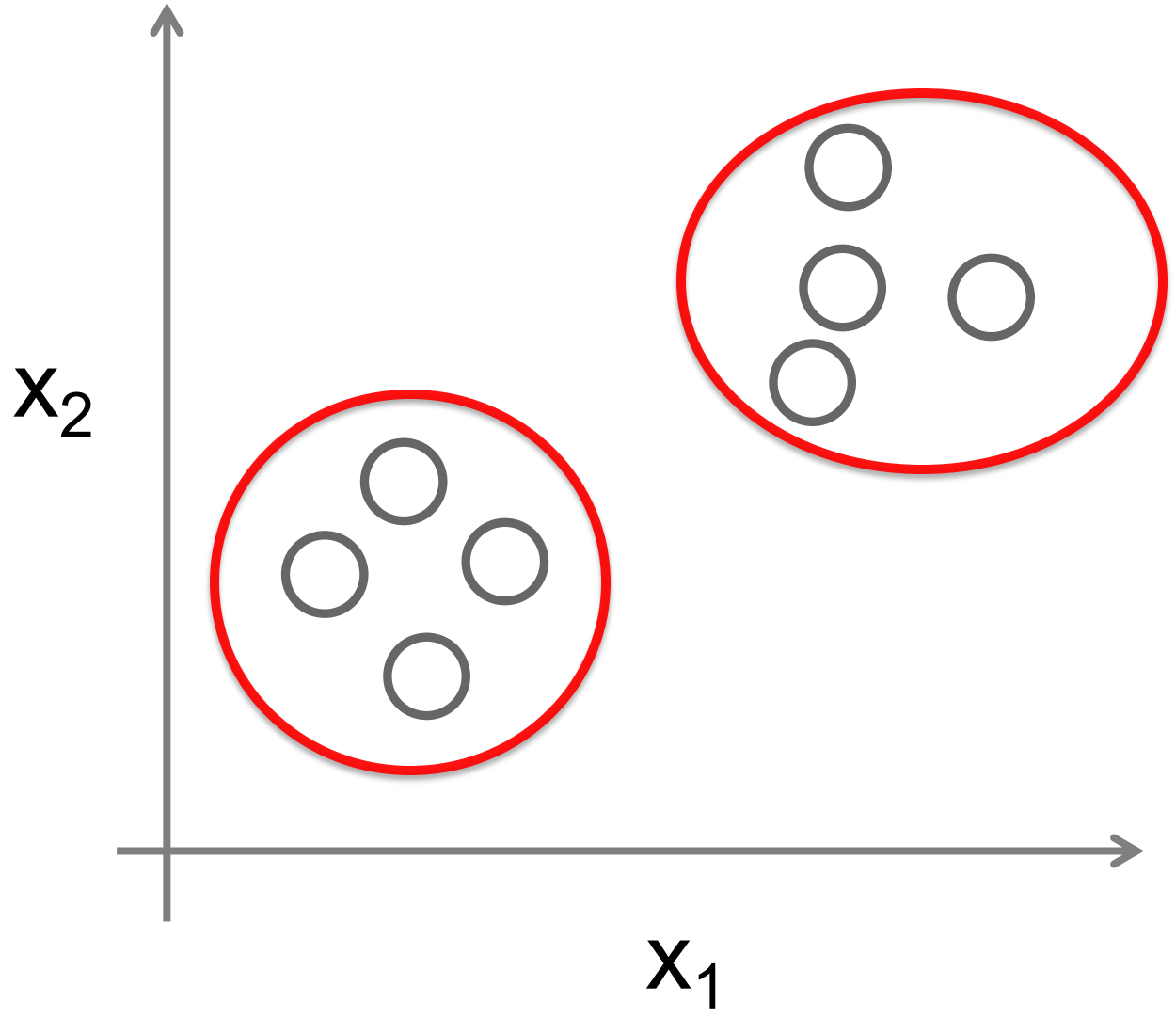
Supervised Learning

Labeled!



Unsupervised Learning

Not Labeled!



Example: Image Clustering

x






f(x)

X=

f(x)=?

Example: Google News

Google News

Search News | Search the Web | Advanced news search | Preferences

Search and browse 25,000 news sources updated continuously.

World » edit

Heavy Fighting Continues As Pakistan Army Battles Taliban
Voice of America - **10 hours ago**
By Barry Newhouse Pakistan's military said its forces have killed 55 to 60 Taliban militants in the last 24 hours in heavy fighting in Taliban-held areas of the northwest.
Pakistani troops battle Taliban militants for fourth day guardian.co.uk
Army: 55 militants killed in Pakistan fighting The Associated Press
Christian Science Monitor - CNN International - Bloomberg - New York Times
all 3,824 news articles »


ABC News

Sri Lanka admits bombing safe haven
guardian.co.uk - **3 hours ago**
Sri Lanka has admitted bombing a "safe haven" created for up to 150000 civilians fleeing fighting between Tamil Tiger fighters and the army.
Chinese billions in Sri Lanka fund battle against Tamil Tigers Times Online
Huge Humanitarian Operation Under Way in Sri Lanka Voice of America
BBC News - Reuters - AFP - Xinhua
all 2,492 news articles »


WA today

U.S. » edi

Weekend Opinionator: Souter, Specter and the Future of the GOP
New York Times - **48 minutes ago**
By Tobin Harshaw An odd week. While Barack Obama celebrated his 100th day in office, the headlines were pretty much dominated by the opposition party, albeit not in the way many Republicans would have liked.
US Supreme Court Vacancy An Early Test For Sen Specter Wall Street Journal
Letters: Arlen Specter, Notre Dame, Chrysler Houston Chronicle
The Associated Press - Kansas City Star - Philadelphia Inquirer - Bangor Daily News
all 401 news articles »


FOXNew

Joe Biden, the Flu and You
New York Times - **48 minutes ago**
By GAIL COLLINS The swine flu scare has made it clear why Barack Obama picked Joe Biden for vice president. David Brooks and Gail Collins talk between columns.
After his flu warning, Biden takes the train home The Associated Press
Biden to visit Balkan states in mid-May Washington Post
AFP - Christian Science Monitor - Bizjournals.com - Voice of America
all 1,506 news articles »


TIME

Business » edit

Buffett Calls Investment Candidates' 2008 Performance Subpar
Bloomberg - **2 hours ago**
By Hugh Son, Erik Holm and Andrew Frye May 2 (Bloomberg) -- Billionaire Warren Buffett said all of the candidates to replace him as chief investment officer of Berkshire Hathaway Inc. failed to beat the 38 percent decline of the Standard & Poor's 500 ...
Buffett offers bleak outlook for US newspapers Reuters
Buffett: Limit CEO pay through embarrassment MarketWatch
CNBC - The Associated Press - guardian.co.uk
all 1,454 news articles » BRK.A

Chrysler's Fall May Help Administration Reshape GM
New York Times - **5 hours ago**
Auto task force members, from left: Treasury's Ron Bloom and Gene Sperling, Labor's Edward Montgomery, and Steve Rattner. BY DAVID E. SANGER and BILL VLASIC WASHINGTON - Fresh from pushing Chrysler into bankruptcy, President Obama and his economic team ...
Comment by Gary Chaison Prof. of Industrial Relations, Clark University
Bankruptcy reality sets in for Chrysler, workers Detroit Free Press
Washington Post - Bloomberg - CNNMoney.com
all 11,028 news articles » OTC:FIATY – BIT:FR – GM


guardian.co.uk

Top-level categories:
supervised classification

Story groupings:
unsupervised clustering

Question

Of the following examples, which would you address using an unsupervised learning algorithm? (Check all that apply.)

1. Given email labeled as spam/not spam, learn a spam filter.

✓ 2. Given a set of news articles found on the web, group them into set of articles about the same story.

✓ 3. Given a database of customer data, automatically discover market segments and group customers into different market segments.

4. Given a dataset of patients diagnosed as either having diabetes or not, learn to classify new patients as having diabetes or not.

Reading

- Read Sections 18.1, 18.2 in the AIMA textbook (Third Edition)