# EECS 395/495 Lecture 4: Machine Learning (in 25 minutes or less)

Doug Downey

# Machine Learning

- "The study of computer programs that improve automatically with experience"
  T. Mitchell *Machine Learning*, 1998
- Used heavily in:
  - Bioinformatics, robotics, marketing/advertising, recommendations systems, information retrieval, fraud detection, handwriting/speech recognition, etc., etc…

# Example Machine Learning Tasks

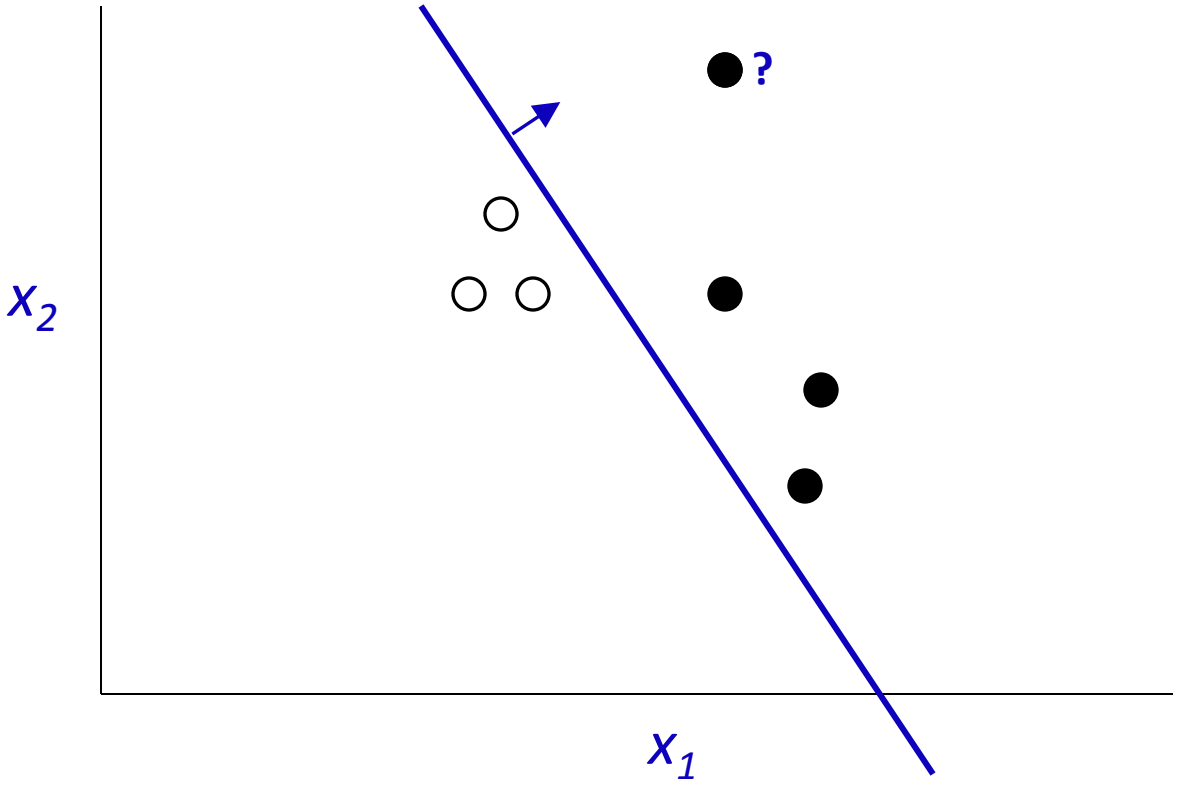
- How likely is person **x** to default on a loan?
- What is the location of robot **x**?
- Is a given Web page **x** about “baseball?”

# Learning a function from examples

- **Given:** examples of a function $f$ for various inputs **x**:
  - $\{(\mathbf{x}_1, f(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, f(\mathbf{x}_n))\}$
- **Goal:** Estimate $f$
  - Input **x** = $(x_1, \ldots, x_d)$; individual features $x_i$
  - Output $f(\mathbf{x})$

- Probably the most common machine learning task formulation (though there are others)

Learn function from $\mathbf{x} = (x_1, \ldots, x_d)$ to $f(\mathbf{x}) \in \{0, 1\}$
given labeled examples $(\mathbf{x}, f(\mathbf{x}))$



# How to do Machine Learning

1) Pick a feature representation for your task
2) Compile data
3) Choose a machine learning algorithm
4) Train the algorithm
5) Evaluate the results
6) *Probably: go to (1)*

# How to do Machine Learning

1) Pick a feature representation for your task
2) Compile data
3) Choose a machine learning algorithm
4) Train the algorithm
5) Evaluate the results
6) *Probably: go to (1)*

# Representation

- In general, inputs and outputs can be
  - Nominal (e.g. Gender)
  - Ordinal (e.g. small, medium, large)
  - Numeric (e.g. Years of Education, probability of credit default, etc.)
- Predicting a nominal output: classification
  - Thus, predicting whether document is about politics or sports is an instance of **Text Classification**
- Predicting a numeric output: regression (typically continuous)

# Feature Engineering

- The art of machine learning
  - Features should be predictive and (relatively) conditionally independent
- How likely is person **x** to default on a loan?
  - FICO score
  - Income
  - Education Level
  - Assets
  - ~~Social Security Number~~
  - …

# How to do Machine Learning

1) Pick a feature representation for your task
2) **Compile data**
3) Choose a machine learning algorithm
4) Train the algorithm
5) Evaluate the results
6) *Probably: go to (1)*

# What the right task (for the class)?

- What you'll be graded on:
  - Utility/Interestingness of the task (15%)
  - Completion of the homework problems (85%)
- So…choose something interesting, but also something you can get done!
- Things to consider:
  - Availability of data
  - "Munging" required
  - Your knowledge of the domain

# Examples

- Something from your research
- The $ ones:
  - Price prediction (e.g. stock market)
  - Box office success
  - The “next big sound” see: nextbigsound.com
  - Sports contests
- UCI Repository
  - Tons of tasks, wines, mushrooms, text…

# Examples

- Text ideas:
  - Improve TextRunner
  - Predict Blog “Anger”
- More…
  - Data.gov – US State data (agriculture, spending, etc.), census data
    - Also: NYC Big Apps
  - Satellite data
  - Customer reviews (summarization, deception detection…)
    - Other item attributes from review?
  - Word choice vs. semantic content
    - Use out-of-text information
  - Smart DJ

# Features?

- Smart DJ

# Something to remember

- A side effect of the models we're building is that they can be used for multiple tasks
  - E.g., not just box office gross prediction, but what's P(Horror movie | Ben Affleck) ?
- So it's interesting to have multiple possible questions in mind for a given data set…

# Brainstorming project ideas

- What's your *second* best project idea?