Lectures: MWF 11am-noon, CIT 367

Course staff and office hours:
- Greg Shakhnarovich (instructor) gregory@cs CIT 509, Thu 1-3pm
- Dan Grollman (TA) dang@cs CIT 315, Fri 1pm-3pm
- Payman Yadollahpour (TA) pyadolla@cs CIT 303, Mon 1pm-3pm

Website: http://www.cs.brown.edu/courses/cs195-5/

Grading:
- Six (≈ bi-weekly) problem sets, 10% each.
- Midterm, 15% and final, 20%
- 5% for participation
- 200-level credit: a project (in addition to the final)
What is Machine Learning?

- The real question is *what is learning?*
What is Machine Learning?

- The real question is *what is learning?*
  - Using past experiences to improve future performance.
What is Machine Learning?

- The real question is *what is learning?*
  - Using past experiences to improve future performance.
- For a machine, experiences come in the form of *data*.
- What does it mean to *improve performance?*
What is Machine Learning?

- The real question is *what is learning*?
  - Using past experiences to improve future performance.

- For a machine, experiences come in the form of *data*.

- What does it mean to *improve performance*?
  - Learning is guided by an objective, associated with a particular notion of *loss* to be minimized (or, equivalently, *gain* to be maximized).
What is Machine Learning?

- The real question is *what is learning*?
  - Using past experiences to improve future performance.

- For a machine, experiences come in the form of *data*.

- What does it mean to *improve performance*?
  - Learning is guided by an objective, associated with a particular notion of *loss* to be minimized (or, equivalently, *gain* to be maximized).

- Why *machine* learning?
  - We need computers to make informed decisions on new, unseen data.
  - Often it is too difficult to design a set of rules “by hand”.
  - Machine learning is about *automatically extracting relevant information* from data and applying it to analyze new data.
Example 1: visual object categorization

We are given categories for these images:
Example 1: visual object categorization

We are given categories for these images: What are these?
Example 1: visual object categorization

We are given categories for these images: What are these?
From ETH database of object categories, [Leibe & Schiele 2003]

- A *classification* problem: predict category $y$ based on image $x$.
- Little chance to “hand-craft” a solution, without learning.
- Applications: robotics, HCI, web search (a *real* image Google...)

CS195-5 2006 – Lecture 1
Supervised learning

This is an example of supervised learning, which consists of the following basic steps:

- **Data collection** Start with training data for which we know the correct outcome provided by a teacher or oracle. In this case: images for which we know the object category.

- **Representation** Choose how to represent the data.

- **Modeling** Choose a hypothesis class - a set of possible explanations for the connection between images and categories. This is our model of the problem.

- **Estimation** Find best hypothesis you can in the chosen class.

- **Model selection** We may reconsider the class of hypotheses given the outcome.

Each of these steps can make or break the learning outcome.
Example 2: document classification

- A few labeled web pages with categories: faculty, student, department, course etc.

- Need to automatically classify previously unseen web pages.

- What would be good features to represent these data?
Example 2: document classification

- A few labeled web pages with categories: faculty, student, department, course etc.

- Need to automatically classify previously unseen web pages.

- What would be good *features* to represent these data?

- *Feature selection* methods allow us to select from a large set of features those most helpful for the task.
Example 3: binary classification

\[ y = +1 \]

\[ y = -1 \]

- Representation as a vector:

\[ \begin{bmatrix} 0000000000 & 0000001100 & 0001111111 & \ldots & 0001100000 \end{bmatrix}^T \]
Modeling

- Examples are binary vectors of dimension $d = 100$:
  
  \[
  x = \begin{bmatrix}
  0000000000 & 0000001100 & 0001111111 & \ldots & 0001100000
  \end{bmatrix}^T
  \]

- Labels are binary as well, $y \in \{-1, +1\}$.

- We consider the following hypothesis class: \( \hat{y} = \text{sign} (w \cdot x) \)

The “hat” \( \hat{\cdot} \) means “estimated”.

Dot product: \( w \cdot x = w^T x = \sum_{i=1}^{d} w_i x_i \)

- This is a \textit{linear classifier} (based on a linear combination of input components).
  It defines a mapping from the data to labels.
Estimation

\[ \hat{y} = \text{sign} \left( \sum_{i=1}^{d} w_i x_i \right) \]

- How can we set the parameter vector \( \mathbf{w} \) using training data?
  - Start with a random set of weights
  - Iterate through the training data; for each \((\mathbf{x}, y)\), if the current classifier makes a mistake, set
    \[ w_i \leftarrow w_i + y x_i \quad \text{for all } i = 1, \ldots, d. \]
Estimation: mistake-driven algorithm

- Start with a random set of weights

- Iterate through the training data; for each \((x, y)\), if the current classifier makes a mistake, set

\[
   w_i \leftarrow w_i + yx_i \quad \text{for all } i = 1, \ldots, d.
\]

- Magnitude \(|w_i|\) reflects importance (weight) of the \(i\)-th pixel.
  - Negative \(w_i\) means that \(i\)-th pixel being on suggests a 9;
  - Positive \(w_i\) means it suggests a 4.

- When do we stop?..
Evaluation

- We can see how well we can predict the labels in the training set:

![Graph showing average error on training set](image)

- Expect the average classification error to go down as we look at more examples. *Do we expect it to reach zero?*

- Baseline: chance performance (random guessing).
Supervised learning beyond classification

- Often the goal is not to classify a data point but to predict some quantitative outcome. This is a regression problem.

- Suppose we want to predict gas mileage of a car based on some characteristics: number of cylinders or doors, weight, horsepower, year etc.
Regression

- Let us look at MPG (miles per gallon) as a function of horsepower only.

- We can fit a straight line to try and explain this behavior:
Regression

- Let us look at MPG (miles per gallon) $y$ as a function of horsepower $x$.

- We can fit a straight line to try and explain this behavior: $\hat{y} = w_1 x + w_0$

- Can we do better than that?
Regression

- We can try to fit a quadratic function: $\hat{y} = w_2 x^2 + w_1 x + w_0$.

- Is it a better fit?
Regression

- We can try to fit a quadratic function: \( \hat{y} = w_2 x^2 + w_1 x + w_0 \).

- Is it a better fit?  *Will it lead to better predictions?*
Generalization - the holy grail of ML

- The ultimate goal is to do as well as possible on new, unseen data (a test set).

- We only have access to labels ("ground truth") for the training set.

- There is a danger of overfitting: learning to predict training labels very well that does not generalize!

- What can we do about it?
  - The most naive approach: minimize training error and keep our fingers crossed.
  - A somewhat more clever approach: if we have enough training data, set some of it aside (holdout) and test on it once learning is done.
  - There are much more powerful, sophisticated and rigorous methods that we will study in this class.
Unsupervised learning

- In *unsupervised learning* the goal is not to predict labels, but to learn some sort of structure in the data.
  - No labels involved!

- Typical problem: *clustering*.

the Old Faithful data
Clustering

- Goal of clustering: discover coherent groups ("clumps") of data.

- Common applications: clustering documents, image segmentation (clustering pixels), activity discovery.
More unsupervised learning

- Other unsupervised tasks:
  - Compression and dimensionality reduction: finding a more parsimonious description for the data (e.g., coding).
  - Detection of outliers/anomalies.
  - Finding correlations between groups of variables.

- The objective is often more vague or subjective than in supervised learning. This is more of an exploratory/descriptive data analysis.
Other learning scenarios

- **Semi-supervised learning**: lots of data available, but only small portion is labeled (e.g. since labeling is expensive).
  - Use unlabeled data to improve learning from the few labeled examples.
  - We will talk about this a bit, time permitting.

- **Reinforcement learning**: action-reward settings.
  - the goal is to find a sequence of actions that maximize expected reward.
  - Probably out of scope of this class...
Learning and probability

• There are many sources of uncertainty with which learning algorithms must cope:
  – Measurement noise
  – Labeling errors
  – Inherent variation in the data

• Probability and statistics provide a robust framework to deal with uncertainty.

• We have already encountered implicitly some basic statistical assumptions:
  – Training data is sampled from the “true” underlying data distribution.
  – Future test data will be sampled from the same distribution.
Applications of learning

• Computer vision and robotics:
  – detection, recognition and categorization of objects
  – face recognition
  – tracking objects (rigid and articulated) in video
  – modeling visual attention

• Speech recognition

• Biology and medicine:
  – drug discovery
  – computational genomics (analysis and design)
  – medical imaging and diagnosis
More applications

- Financial industry:
  - Fraud detection
  - Credit approval
  - Price and market prediction

- Information retrieval, Web search, Google ads...

- Many other applications (and many jobs!)
The machine learning community

- Fast growing representation in CS and similar departments
- Industry: major part of the research focus at Microsoft, MERL, SRI, Yahoo, Google, IBM, ...
- Main annual conferences:
  - International Conf. on Machine Learning (ICML)
  - Conf. on Computational Learning Theory (COLT)
  - Neural Information Processing Systems (NIPS)
  - Large presence at specialized conferences on speech, vision, robotics, graphics, comp. biology, etc.
- Main journals: JMLR, PAMI, Neural Computation, ...
Helpful background

• Basic probability:
  – Random variable; expectation and variance; covariance; independence; probability mass function; Gaussian distribution (scalar and multivariate).

• Basic linear algebra:
  – Linear independence; matrix inverse; eigenvectors; rank.

• Having background in these topics will help, but we will provide a refresher on each concept as we encounter it.
Next time

We will start looking in depth at the regression problem