Foundations of Machine Learning
Lecture 1

Mehryar Mohri
Courant Institute, NYU
mohri@cs.nyu.edu

Logistics

- Prerequisites: basics in linear algebra, probability, and analysis of algorithms.
- Workload: homework assignments (4-5) + project (topic of your choice).
- Textbooks: no single textbook covering the material presented in this course, lecture slides will be made available electronically.

Introduction to Machine Learning

Machine Learning

- Definition: computational methods using experience to improve performance [e.g., to make accurate predictions].
- Experience: data-driven task [thus statistics, probability].
- Example: use height and weight to predict gender.
- Computer science: need to design efficient and accurate algorithms, analysis of complexity, theoretical guarantees.

Examples of Learning Tasks

- Optical character recognition
- Text or document classification, spam detection
- Morphological analysis, part-of-speech tagging, parsing
- Speech recognition, speech synthesis, speaker verification
- Image recognition, face recognition
- Fraud detection (credit card, telephone), network intrusion
- Games (chess, backgammon)
- Unassisted control of a vehicle (robots, navigation)
- Medical diagnosis

Some Broad Areas of ML

- Classification: assign a category to each object (OCR, text classification, speech recognition; note: the number of categories may be infinite in some difficult tasks).
- Regression: predict a real value for each object (prediction of stock values, variations of economic variables).
- Ranking: order objects according to some criterion (relevant web pages returned by a search engine).
- Clustering: partition data into homogenous groups (analysis of very large data sets).
- Dimensionality reduction: find lower-dimensional manifold preserving some properties of the data (computer vision).

Objectives of Machine Learning

- Algorithms: design of efficient, accurate, and general learning algorithms to
  - deal with large-scale problems (|data| > 1-10M).
  - make accurate predictions (unseen examples).
  - handle a variety of different learning problems.
- Theoretical questions
  - what can be learned efficiently? Under what conditions?
  - how well can it be learned computationally?
- Other: better understanding of (human or animal) learning? Help human learning? Better learning than humans.

This Course

- Several major and mathematically well-studied algorithms, e.g.,
  - support vector machines (SVMs), kernel methods
  - boosting algorithms
  - automata learning algorithms
- Theoretical foundations
  - analysis of algorithms
  - generalization bounds
- Applications
  - illustration of the use of these algorithms

Topics

- Probability, general bounds
- PAC learning model, error bounds, VC-dimension, bounds on sample complexity
- Support vector machines (SVMs), Perceptron, Winnow
- Kernel methods
- Boosting, generalization error, margin
- On-line learning, halving algorithm, weighted majority algorithm, mistake bounds
- Ranking problems and algorithms
- Empirical evaluation, confidence intervals, comparison of learning algorithms
- Learning automata and transducers, Angluin-type algorithms, other algorithms
- Reinforcement learning

Definitions and Terminology

- *Example*: an object, instance of the data used.
- *Features*: the set of attributes, often represented as a vector, associated to an example (e.g., *height* and *weight* for gender prediction).
- *Labels*: in classification, category associated to an object (e.g., *positive* or *negative* in binary classification); in regression real value.
- *Training data*: data used for training learning algorithm (often *labeled data*).
- *Test data*: data used for testing learning algorithm (*unlabeled data*).
- *Unsupervised learning*: no labeled data; *supervised learning*: uses labeled data; *semi-supervised learning*: intermediate situations.

Example - SPAM Detection

- Problem: classify each e-mail message as SPAM or non-SPAM (binary classification problem)
- Potential data: large collection of SPAM and non-SPAM messages (labeled examples)
- Learning stages:
  - divide labeled collection into training and test data.
  - associate relevant features to examples (e.g., presence or absence of some sequences; importance of *prior knowledge*).
  - use training data and features to train machine learning algorithm.
  - predict labels of examples in test data, evaluate algorithm.

Example

- Problem: Predict next symbol (regression problem)
- It is sometimes difficult to find relevant features
- Knowledge about the problem can be very useful!

Example

- Problem: Predict next symbol (regression problem)
- It is sometimes difficult to find relevant features
- Knowledge about the problem can be very useful!

Training		Test data	
n v p d	n	n v d a	-
d n v d	n	n v n b	-
a n v b	b		
d n v p	n		
d a n v	b		

Example

- Problem: Predict next symbol (regression problem)
- It is sometimes difficult to find relevant features
- Knowledge about the problem can be very useful!

Training		Test data	
n v p d	n	n v d a	-
d n v d	n	n v n b	-
a n v b	b		
d n v p	n		
d a n v	b		

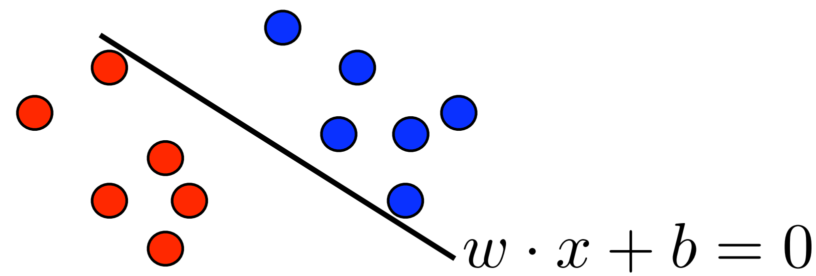
Training data		Test data	
noun verb prep det	noun	noun verb det a	noun
det noun verb det	noun	noun verb noun adv	adv
a noun verb adv	adv		
det noun verb prep	noun		
det a noun verb	adv		

Generalization

- Definition: a learning algorithm is a consistent learner when it commits no error on examples from the training data

- Naive consistent learners are poor predictors, e.g.,

  - Arbitrary linear separation:



  - Learning DNF formulas: the disjunction of all positive examples is a consistent learner, but learning *k*-term DNF is NP-complete!

    $\bigvee_{i=1}^{k} a_i(X_1) \wedge \cdots \wedge a_i(X_n)$, with $a_i(X_j) \in \{X_j, \overline{X_j}, 1\}$.

- Problem: poor generalization, closer to memorization, computational complexity.

Probability Review

Probabilistic Model

- Sample space: Ω, set of all outcomes or *elementary events* possible in a trial, e.g., casting a die or tossing a coin.

- Event: subset $A \subseteq \Omega$ of sample space. The set of all events must be closed under complementation and countable union and intersection.

- Probability distribution: mapping $\Pr$ from the set of all events to $[0, 1]$ such that $\Pr[\Omega] = 1$, and for all mutually exclusive events,
$$\Pr[A_1 \cup \ldots \cup A_n] = \sum_{i=1}^{n} \Pr[A_i].$$

Random Variables

- Definition: a random variable is a function $X : \Omega \to \mathbb{R}$ such that for any interval I, the subset of the sample space $\{A : X(A) \in I\}$ is an event. Such a function is said to be measurable.

- Example: the sum of the values obtained when casting a die.

- Probability density function of random variable X: function $f : x \mapsto f(x) = \Pr[X = x]$.

- Joint probability density function of X and Y:
$$f : (x, y) \mapsto f(x, y) = \Pr[X = x \wedge Y = y].$$

Conditional Probability and Independence

- Conditional probability of event A given B:

$$\Pr[A \mid B] = \frac{\Pr[A \wedge B]}{\Pr[B]},$$

when $\Pr[B] \neq 0.$

- Independence: two events A and B are *independent* when

$$\Pr[A \wedge B] = \Pr[A]\Pr[B].$$

Equivalently, $\Pr[A \mid B] = \Pr[A]$, when $\Pr[B] \neq 0.$

Some Probability Formulae

- Sum rule:

$$\Pr[A \vee B] = \Pr[A] + \Pr[B] - \Pr[A \wedge B].$$

- Union bound:

$$\Pr[\bigvee_{i=1}^{n} A_i] \leq \sum_{i=1}^{n} \Pr[A_i].$$

- Bayes formula:

$$\Pr[X \mid Y] = \frac{\Pr[Y \mid X] \Pr[X]}{\Pr[Y]} \quad (\Pr[Y] \neq 0).$$

Some Probability Formulae

- Chain rule:

$$\Pr[\bigwedge_{i=1}^{n} X_i] = \Pr[X_1] \Pr[X_2 \mid X_1] \Pr[X_3 \mid X_1 \wedge X_2] \\ \ldots \Pr[X_n \mid \bigwedge_{i=1}^{n-1} X_i].$$

- Theorem of total probability: assume that

$$\Omega = A_1 \cup A_2 \cup \ldots \cup A_n, \text{ with } A_i \cap A_j = \emptyset \text{ for } i \neq j;$$

then for any event B,

$$\Pr[B] = \sum_{i=1}^{n} \Pr[B \mid A_i] \Pr[A_i].$$

Application - Maximum a Posteriori

- Problem formulation: given some observation *O*, determine the most likely outcome out of a set of hypotheses *H*:

$$\hat{h} = \operatorname*{argmax}_{h \in H} \Pr[h \mid O] = \operatorname*{argmax}_{h \in H} \frac{\Pr[O|h]\Pr[h]}{\Pr[O]} = \operatorname*{argmax}_{h \in H} \Pr[O|h]\Pr[h]$$

Example - medical diagnosis: laboratory test with two results *O*={*Positive*, *Negative*} used to determine if a patient has specific disease *d*, thus *H* = {*d*, *no-d*}. Assumptions:

- Pr[*d*] = .005 (a priori probability of *d*);
- Pr[*Positive* | *d*] = .98 (probability of true positive)
- Pr[*Negative* | *no-d*] = .95 (probability of true negative)
- If the test is *Positive*, what should be the diagnosis?

Pr[*Positive* | *d*] Pr[*d*] = .98 x .005 = .0049

Pr[*Positive* | *no-d*] Pr[*no-d*] = (1 - .95) x (1 - .005) = .04975 > .0049

Expectation

- Definition: the *expectation* (or *mean*) of a random variable X is

$$\mathrm{E}[X] = \sum_x x \Pr[X = x].$$

- Properties:
    - linearity, $\mathrm{E}[aX + bY] = a\mathrm{E}[X] + b\mathrm{E}[Y]$.
    - if X and Y are independent,

$$\mathrm{E}[XY] = \mathrm{E}[X]\mathrm{E}[Y].$$

Expectation

- **Theorem** (Markov's inequality): let X be a non-negative random variable with $\mathrm{E}[X] < \infty$, then for all $t > 0$,

$$\Pr[X \geq t\mathrm{E}[X]] \leq \frac{1}{t}.$$

- **Proof**:

$$\begin{aligned}
\Pr[X \geq t\mathrm{E}[X]] &= \sum_{x \geq t\mathrm{E}[X]} \Pr[X = x] \\
&\leq \sum_{x \geq t\mathrm{E}[X]} \Pr[X = x] \frac{x}{t\mathrm{E}[X]} \\
&\leq \sum_{x} \Pr[X = x] \frac{x}{t\mathrm{E}[X]} \\
&= \mathrm{E}\Big[\frac{x}{t\mathrm{E}[X]}\Big] = \frac{1}{t}.
\end{aligned}$$

Variance

- Definition: the *variance* of a random variable X is

$$\mathrm{Var}[X] = \sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2].$$

σ_X is called the *standard deviation* of the random variable X.

- Properties:
  - $\mathrm{Var}[aX] = a^2 \mathrm{Var}[X]$.
  - if X and Y are independent,

$$\mathrm{Var}[X + Y] = \mathrm{Var}[X] + \mathrm{Var}[Y].$$

Variance

- **Theorem** (Chebyshev's inequality): let X be a random variable with $\mathrm{Var}[X] < \infty$, then for all $t > 0$,
$$\Pr[|X - \mathrm{E}[X]| \geq t\sigma_X] \leq \frac{1}{t^2}.$$
- **Proof**: Observe that

$$\Pr[|X - \mathrm{E}[X]| \geq t\sigma_X] = \Pr[(X - \mathrm{E}[X])^2 \geq t^2\sigma_X^2].$$

The result follows Markov's inequality.

Application

- Experiment: roll a pair of fair dice n times. Can we give a good estimate of total value of the n rolls?

- Mean: $7n$, variance: $35/6\ n$; thus by Chebyshev's inequality, the final sum will lie between

$$7n - 10\sqrt{\frac{35}{6}n} \text{ and } 7n + 10\sqrt{\frac{35}{6}n}$$

in at least 99% of all experiments. The odds are better than 99 to 1 that the sum be roughly between 6.976M and 7.024M after 1M rolls.

Weak Law of Large Numbers

- **Theorem**: let $(X_n)_{n \in \mathbb{N}}$ be a sequence of random variables with the same mean μ and variance $\sigma^2 < \infty$ and let $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, then for any $\epsilon > 0$,

$$\lim_{n \to \infty} \Pr[|\overline{X}_n - \mu| \geq \epsilon] = 0.$$

- **Proof**: Since the variables are independent,

$$\mathrm{Var}[\overline{X}_n] = \sum_{i=1}^n \mathrm{Var}[\frac{X_i}{n}] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.$$

- Thus, by Chebyshev's inequality,

$$\Pr[|\overline{X}_n - \mu| \geq \epsilon] \leq \frac{\sigma^2}{n\epsilon^2}.$$