Optimization Methods In Machine Learning

Lecture 1: Introduction

Professor Katya Scheinberg

Lehigh University

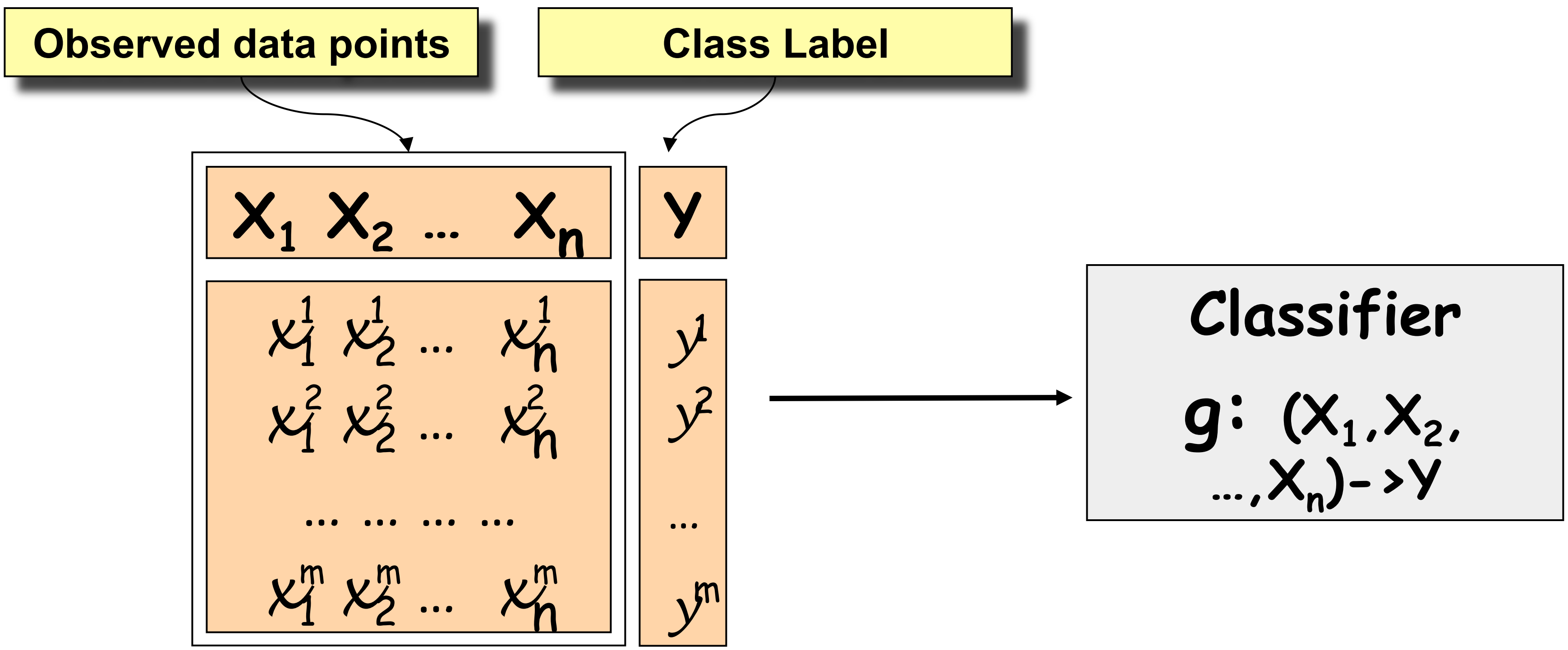
Spring 2016

- Instructor: Katya Scheinberg
- kas410@lehigh.edu
- Office: 486
- Office Hours – by appointment
- Evaluation – homework + project
- We will probably use Matlab for some exercises later in the course.
- Attendance and participation
- Plan for the course

A few words about machine learning..

Warning: This course will offer an optimizers view of machine learning.

Learning a Classifier from Data



Data ≡ set of i.i.d (independent & identically distributed samples)

Examples from image classification

- Optical character recognition
  - Automatically read digits in zip code
    - 256 dim vector of pixels, 10 classes,
    - classification or clustering task
- Face recognition and detection
  - much larger dimension, nonlinear representation,
  - Non-euclidean similarity measures

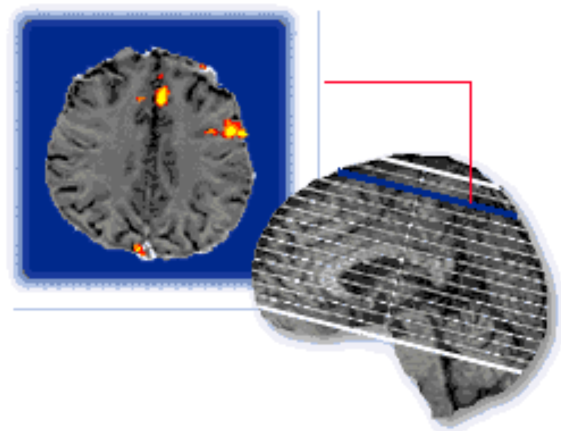
Examples from text and internet

- Text categorization
  - detect spam/nonspam emails
    - Many possible features
    - False positives are very bad, false negatives are OK.
    - Online setting possible, huge data sets.
  - choose articles of interest to individualize news sites
    - Large dimension – size of dictionary, small training set, possibly online setting
    - Only few words are important.
- Ranking
  - Predict a page rank for a given a search query
    - How to do it? Predict relative ranks of each pair of pages?

Examples from Medicine

- Functional Magnetic resonance imaging
  - Uses a standard MRI scanner to acquire functionally meaningful brain activity
  - Measures changes in blood oxygenation
  - Non-invasive, no ionizing radiation
  - Good combination of spatial / temporal resolution
    - Voxel sizes ~4mm
    - Time of Repetition (TR) ~1s

    About 30000 voxels are active and measured.

  - Only a few (probably) contribute to what the subject is “feeling” during the experiment (anger, frustration, boredom..)
- Breast cancer risk patients
  - Take several measurements of a patient and some basic characteristics an predict if the patient is at high risk
  - Low dimensional, but very different attributes. Large scale data.
  - May involve “active learning” – additional labels obtained by involving more tests or a professional.
  - KDD 2008 cup challenge

fMRI image courtesy of fMRI Research Center @ Columbia Unoversity


http://www.netflixprize.com/
kdd cup 2007
McAfee SiteAdvisor
kdd cup 2007
Search
Bookmarks
Check
AutoFill
kdd
Suggested Sites
Get More Add-ons
Page
Safety
Tools
NETFLIX
Netflix Prize
Home
Rules
Leaderboard
Register
Update
Submit
Download
NETFLIX
Browse
Recommendations
Friends
Queue
Buy DVDs
Movies For You
Welcome!
The Netflix Prize seeks to substantially improve the accuracy of predictions about how much someone is going to love a movie based on their movie preferences. Improve it enough and you win one (or more) Prizes. Winning the Netflix Prize improves our ability to connect people to the movies they love.
Read the Rules to see what is required to win the Prizes. If you are interested in joining the quest, you should register a team.
You should also read the frequently-asked questions about the Prize. And

Outline

This lecture is taken from a short course at UT Austin taught by N. Srebro and K. Scheinberg in 2011.

Outline

Introduction

Definitions

What is machine learning?

Machine learning, a branch of artificial intelligence, concerns the construction and study of systems that can learn from data.

The core of machine learning deals with **representation** and **generalization**:

- Representation of data instances and functions evaluated on these instances are part of all machine learning systems

- Generalization is the property that the system will perform well on unseen data instances.

“Machine learning” Vs. “Expert knowledge”

Optical Character Recognition (OCR) for typed characters in the Latin alphabet is a classic problem that has been attacked using both the” machine learning” approach and the “expert knowledge” approach.

- **“Expert knowledge” approach** programs a system with an explicit rule and knowledge of characters, for example identifies the lines in a given character. It would not feed any examples to the system.
- **“Machine-learning” approach** takes lots of data and uses some machine learning method to automatically develop a system. Learning process outputs is essentially a program that takes an image and tells what letter it is.

The advantages of machine learning

- In machine learning the process is much easier, as there is much less programming involved.
- Machine learning approach is definitely more adaptive to changes in the data, For example handwritten Latin characters instead of typed Latin characters.
- In machine learning we can train systems to do things that we dont even know how to do ourselves. For example we can even consider a more extreme example and recognize an alphabet that we have no “expert knowledge” about, such Chinese alphabet or the Arabic alphabet.
- One final advantage of machine learning is that in many applications it simply yields much better performance.

Examples of machine learning

- **Character recognition**

Given an image of a character, correctly identify the character.

- **Spam recognition**

Given an email, correctly identify the email as spam or not-spam.

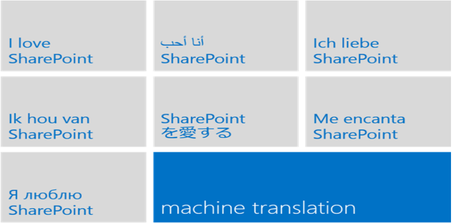
Examples of machine learning

- **Speech recognition**

Given an audio of speech, identify the words being said.

- **Machine translation**

Given a sample of text in one language, produce text in another language with the same meaning.



Examples of machine learning

- **Computer vision**

Starting with some seminal work on face recognition and continuing to the present with almost every other application in vision, vision has been turned into a largely learning-base field.

Instead of trying to figure out geometrically what geometry makes the face, we just give the computer a bunch of faces and let it figure out "In these images, this is what makes up a face".

- **Ranking web search results**

Given a search query return a ranking of web pages by relevance/"goodness".

- **Recommender systems**

For example "Netflix movie recommender system".

Outline

Data

- **Data and Labels**

In learning we seek a mapping from the initial data $\mathcal{X}$ (the domain of abstract input objects) to some label set $\mathcal{Y}$ (anything we want to predict).

Example: In character recognition $\mathcal{X}$ consists of possible images of letters and $\mathcal{Y}$, consists of the twenty-six letters of the Latin alphabet.

Note: For simplicity we will use binary labels $\{+1, -1\}$. Whether something is the letter "G" (+1) or not the letter "G" (-1), or whether a given image contains a face (+1) or does not contain a face (-1).

Joint distribution

- **Joint distribution** $p_{X,Y}(x,y)$

Future data is coming from some unknown source joint distribution $p_{X,Y}$ over input objects and their corresponding labels, which we write as the joint distribution $p_{X,Y}(x,y)$, where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$.

Example: Character recognition source distribution would assign much more probability to ("image containing a circular shape", "O") than to ("image containing a circular shape","T").

Conditional probability

- **Conditional probability** $p_{Y|X}(y|x)$

We can define source joint distribution as really having two components:

$$p_{X,Y}(x,y) = p_{Y|X}(y|x) \cdot p_X(x)$$

Where $p_{Y|X}(y|x)$ is the conditional probability of the label random variable Y given the appearance random variable X and $p_X(x)$ is marginal probability of the input image.

Note: $p_{Y|X}(y|x)$, is defined as "correctness" for a predictor.

Example: In character recognition we may have:

$p_{Y|X}(Y = \text{"A"} | X = \text{"image of an } A\text{"}) = 1$

$p_{Y|X}(Y = \text{"C"} | X = \text{"image of an A"}) = 0$

Hypothesis and Loss function

- **Hypothesis** h

A hypothesis (a predictor) h is a function from X to Y, $h : X \mapsto Y$.

- **Loss function** $loss_{01}\left(h\left(x\right), y\right)$

How we can evaluate the performance of h on a given (input,label) pair (x, y)?

If the label $h\left(x\right)$ does not match the provided label y, we incur a loss of 1 and if the prediction $h\left(x\right)$ does match the provided label y, we incur 0 loss.

The loss function that represents this measure of performance is called the 01 loss and defined as:

$$loss_{01}(h(x), y) = \begin{cases} 1 & \text{if } \ h(x) \neq y \\ 0 & \text{if } \ h(x) = y \end{cases}$$

Expected Risk

- **Expected Risk** $R_{01}\left[h\left(\cdot\right)\right]$

How well we expect to do (on average) over the entire (admittedly unknown) source joint distribution $p_{X,Y}\left(x,y\right)$?

The expected risk $R_{01}\left[h\left(\cdot\right)\right]$ of a hypothesis h on that distribution $p_{X,Y}\left(x,y\right)$, measures the performance of this hypothesis by evaluating its expected loss over pairs (x,y) drawn from the distribution:

$$R_{01}[h(.)] = \mathbb{E}_{(X,Y)\sim P_{X,Y}}[loss_{01}(h(X),Y)] = \sum_{X,Y} P(x,y)\; loss(h(x),y)$$

Note: Other terms with the same meaning are *expected loss*, *generalization error*, or *source-distribution risk*.

Additional property of Expected Risk

- A predictor h is “good” on a particular source joint distribution if it has low risk $R[h(.)]$ on that distribution.

- The 01 risk $R_{01}[h(.)]$ is the probability that the predictor h will incorrectly predict the label for any pair (x, y) drawn at random from the source joint distribution:

$$R_{01}[h(.)] = \mathbb{E}_{(X,Y)\sim P_{X,Y}}[loss_{01}(h(X), Y)] = \mathbb{P}_{(X,Y)\sim P_{X,Y}}\{h(X) \neq Y\}$$

- This equivalence between the risk and the probability of incorrect label prediction holds only for the 01 loss.

Note: We will be assuming that the source joint distribution is fixed.