Introduction to Machine Learning: A Gentle Start

Tirgul 1

Based on slides by Danny Karmon

November 2016
Administration

• 70% final exam
• 30% assignments
  • Theoretical and practical:
    • Python
    • Could appear in the final exam in some form (exact, similar, conclusions).

• tzeviya.biu@gmail.com
Machine Learning in Real Life is a Key role player in a wide range of applications and tasks
E.g.: Search Engines

“bush”

Given a keyword, propose links ranked by relevance
E.g.: Image Recognition

Recognize faces in pictures based on tagged images

Is this Martha?
E.g.: Recommender Systems

Recommend a new item based on past selections
Machine Learning Problems

- Is this cancer?
- What movie should I watch next?
- Who is this?
- What did you say?
- Is it going to rain tomorrow?
- Is this spam?
- etc.
What is Machine Learning?

• “[Machine Learning is the] field of study that gives computers the ability to learn without being explicitly programmed.”

  Arthur Samuel, 1959

• “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, 1997
ML task type distinction: Supervised vs Unsupervised

• Supervised Learning:
  • The program is “trained” on a predefined set of “training examples”, which then facilitate its ability to reach accurate conclusion when given new data.

• Unsupervised Learning:
  • The program is given a bunch of data and must find patterns and relationships therein.
Supervised Learning
Supervised Learning: Example 1

- Housing price prediction

![Graph showing housing price prediction vs size in square feet.](image-url)
Supervised Learning

• Goal:
  • Develop a finely tuned predictor function
    • A.k.a “hypothesis”

• “Learning” consists of using sophisticated mathematical algorithms to optimize this function

• Given input data, it will accurately predict the desired value
Supervised Learning: Example 1

- Housing price prediction

- **Supervised Learning:**
  - "right answer" is given

- **Regression:**
  - Predicts a continuous valued output (the price)
Classification vs Regression

• **Regression machine learning systems:** Systems where the value being predicted falls somewhere on a continuous spectrum. These systems help us with questions of “How much?” or “How many?”.

• **Classification machine learning systems:** Systems where we seek a yes-or-no prediction, such as “Is this tumor cancerous?”, “Does this cookie meet our quality standards?”, and so on.
Supervised Learning: Example 2

- Quality of Cookie
  - Supervised Learning: “right answer” is given
  - Classification: Predicts a yes/no label

yummy cookie?

% Chocolate Chips in cookie

0.2  0.4  0.6  0.8

0  1
Supervised Learning: Example 2 cont.

• Quality of Cookie

![Graph showing the relationship between cookie radius (in centimeters) and the percentage of chocolate chips in a cookie.]
Pivot Example: Student Performance

- Goal: predict a student’s grade in a final exam.
Prediction $\hat{y} \in Y$
Possible Features

- Sex
- Age
- Family Size
- Mother’s Education
- Father’s Education
- Mother’s Job
- Father’s Job
- Home to school travel time

- Weekly study time
- Number of past class failures
- Wants to take higher education?
- Internet access at home?
- Goes out with friends?
- Health status
- Number of school absences
= <weekly study time, past failures, wants higher education>  Performance

<table>
<thead>
<tr>
<th>Numeric</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1: &lt; 2 hours</td>
<td>2: 2-5 hours</td>
<td>3: 5-10 hours</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Numeric</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1: n&lt;3</td>
<td>2: n&gt;=3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Binary</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1: yes</td>
<td>0: no</td>
</tr>
</tbody>
</table>

= < 1, 2, 0>  40

= < 4, 1, 1>  95
#1: Feature Extraction

• Learning algorithms input exist in numerical domains.
• Features → Numerical representation (e.g. vector).
  • Too many → harder to learn (noise, large dimension)
  • Too less → missing critical properties (can’t learn!)

• $x \in \mathbb{R}^d$
• **Goal:** representing distinguishing characteristics

• More: *Luck*\*Art\*Magic. Less: *Science*
Student Performance

#1 Feature
Prediction \( \hat{y} \in Y \)
#2: Learning Algorithm

- **Goal:** finding appropriate weights for the features
  - What’s more\less important?

- Goal (redefined):
  - Find weight vector w
    - A.k.a the *hypothesis* $h(x)$
  - Optimization problem
    - Find w that maximizes the prediction accuracy.

- Inference:
  - $w \cdot x \in R = \text{score}$
  - Dot product definition:
    - $\vec{w} \cdot \vec{x} = \sum_{i=1}^{n} w_i x_i = w_1 x_1 + w_2 x_2 + \ldots + w_n x_n$
Weights * features

\[ w \cdot x = w_1 \cdot \text{weekly study time} + w_2 \cdot \text{past failures} + w_3 \cdot \text{wants higher education} \]

W = \langle w_1, w_2, w_3 \rangle

coefficient vector: weighting the importance of each feature

○ Ignoring useless\ noisy features (very small weight)
○ Focusing on distinguishing features
Classification vs Regression

• **Regression machine learning systems**: Systems where the value being predicted falls somewhere on a continuous spectrum. These systems help us with questions of “How much?” or “How many?”.
  - \( w \cdot x = \text{score} \rightarrow \text{target output} \)

• **Classification machine learning systems**: Systems where we seek a yes-or-no prediction, such as “Is this tumor cancerous?”, “Does this cookie meet our quality standards?”, and so on.
  - \( w \cdot x = \text{score} \rightarrow \text{prediction class} \)
Course Focus

Prediction $\hat{y} \in Y$
#2: The process flow

- **Train** predictor according to given dataset:
  - \((x,y)\) where:
    - \(x\) is an instance (a vector)
    - \(y\) is its target value

- **Test** the predictor on new data:
  - Evaluate performance
Training Mode

• Training and modifying the predictor function according to a given data set (i.e. training set)

• **Goal**: well trained weight vector on target distribution (not specific on training set)
Training Mode: Epoch

• An Epoch:
  • A single pass on the entire training set

• Usually a single epoch is not enough for suitable training.
Training Mode: Batch vs. Iterative

**Batch Learning**

for each epoch:
  for $i = 1$ to $|\text{Training Set}|$:
    1. pick a pair $(x_i, y_i) \in (X,Y)$
    2. predict $\hat{y}_i$ using hypothesis
    3. compare performance $\hat{y}_i$ vs. $y_i$
    Update hypothesis accordingly
Return hypothesis

**Iterative Learning**

for each epoch:
  for $i = 1$ to $|\text{Training Set}|$:
    1. pick a pair $(x_i, y_i) \in (X,Y)$
    2. predict $\hat{y}_i$ using hypothesis
    3. compare performance $\hat{y}_i$ vs. $y_i$
    4. update $w$ accordingly
Return averaged version of hypothesis

\[ \text{Suffer the loss (cost function)} \]
\[ \text{Correct weight vector} \]
Training Mode: Batch vs. Iterative

**Batch Learning**
- The hypothesis stays constant while computing the error associated with each sample in the input.
- More efficient in terms of # of computations.

**Iterative Learning**
- Constantly updating the hypothesis, its error calculation uses different weights for each input sample.
- More practical in case the big data sets.

Both converge to the same minimum (explained in future lessons)
**Training Mode: Batch vs. Online**

**Batch Learning**

for each epoch:
  for i = 1 to |Training Set|:
    1. pick a pair \((x_i, y_i)\) \(\in\) \((X, Y)\)
    2. predict \(\hat{y}_i\) using hypothesis
    3. compare performance \(\hat{y}_i\) vs. \(y_i\)
  Update hypothesis accordingly

Return hypothesis

**Online Learning**

for \(t = 1\) to \(T\):
  1. pick the pair \((x_i, y_i)\) \(\in\) \((X, Y)\)
  2. predict \(\hat{y}\) using the hypothesis
  3. compare performance \(\hat{y}_i\) vs. \(y_i\)
  4. update hypothesis

Always updating hypothesis
Performance evaluated according to current hypothesis
Test Mode

• Once the training process is done - our trained predictor model (i.e. hypothesis) is ready for action!

• Evaluate the performance

Performance Evaluation
Accuracy/Error rate

\[ x \rightarrow \hat{y} \rightarrow y \]
Test Mode

for $i = 1$ to $|\text{Test Set}|$:
1. pick the pair $(x_i, y_i) \in (X, Y)$
2. predict $\hat{y}_i$ using the hypothesis
3. compare performance $\hat{y}_i$ vs. $y_i$
return average performance (error rate)
Unsupervised Learning
Unsupervised Learning

Supervised Learning

Unsupervised Learning
Unsupervised Learning

- **Task:** find relationships, patterns and correlations within data
- No training examples used in this process.
E.g.: Google News
Summary

• What is Machine Learning?
• Supervised Learning
  • Classification vs. Regression
  • Two Steps:
    • Feature Extraction
    • Learning – finding w
  • The process flow:
    • Train mode and test mode
    • Train mode: batch/iterative/online
• Unsupervised Learning
  • clustering