Introduction of Machine Learning

Filippo Federici Canova and Martha Arbayani Zaidan
16 March 2017

Spring School in Computational Chemistry 2017 @CSC

Aalto Science Institute (ASci)
Centre of Excellence in Computational Nanoscience (COMP)
School of Science, Aalto University
Outline

✓ Introduction of machine learning concept
✓ Motivation: machine learning for computational chemistry
✓ Machine learning approaches:
  1. Principal Component Analysis (PCA)
  2. Artificial Neural Network (ANN)
  3. K-means clustering
  4. K-Nearest Neighbors (KNN)
  5. Kernel Ridge Regression (KRR)
  6. Genetic Algorithm (GA)
✓ Demo: ANN + GA for video games
Introduction: Machine Learning concept
What is Machine Learning?

Algorithms that can learn patterns from data, and use the information to analyse new data, predict future outcomes, or to perform other kinds of decision making under uncertainty.

Why “Learn”?  

• Learning is used when:     
  – Human expertise does not exist (navigating on Mars)  
  – Humans are unable to explain their expertise (speech recognition)  
  – Solution changes in time (financial market)  
  – Solution needs to be adapted to particular cases (user biometrics)
Growth of Machine Learning

- Machine learning has been used in:
  - Speech recognition, natural language processing
  - Computer vision
  - Medical outcomes analysis
  - Robot control
  - Computational biology
  - Computational physics/chemistry
## Machine Learning Problems

<table>
<thead>
<tr>
<th>Supervised Learning</th>
<th>Unsupervised Learning</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discrete</strong></td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td>Clustering</td>
</tr>
<tr>
<td><strong>Continuous</strong></td>
<td></td>
</tr>
<tr>
<td>Regression</td>
<td>Dimensionality Reduction</td>
</tr>
</tbody>
</table>
Unsupervised learning

• Input data without labels.
• Learning “what normally happens”, e.g. with clustering methods.

Examples:

• Customer segmentation in Customer Relationship Management (CRM)
• Bioinformatics: Learning motifs

Clustering aims to discover groups of similar elements within the data.

similarity metric determines the results!
Supervised learning

• Data consists of inputs $x$ and *correct* outputs $y$.
• Methods: classification and regression.

**Classification:**
- data: people descriptor + color output
- *where do I go?*
- predict association of new people

**Regression:**
- Same... with continuous variables.
Machine Learning for Computational Chemistry
Common Chemistry Problems

Find composition/shape of clusters and molecules

prediction of their properties
Principal Component Analysis for Dimensionality Reduction
Principal Component Analysis

- Principal Component Analysis (PCA) is a dimensionality-reduction technique that is often used to transform a high-dimensional dataset into a smaller-dimensional subspace prior to running a machine learning algorithm on the data.

- The PCA transformation is given by:

  \[ X V = Y \]

- \( X \) is the original dataset, \( V \) is the Principal Components (PCs) and \( Y \) is the re-representation of the original dataset. The PCs are the eigenvectors of the covariance matrix of the original dataset, where \( \text{cov} (X) \) is defined as:

  \[
  \text{cov}(X) = \frac{1}{N} \sum_{n}^{N} (x_n - \mu_x)(x_n - \mu_x)
  \]

- The PCs (eigenvectors) correspond to the direction with the greatest variance in the data whereas the corresponding eigenvalue is a number that indicates how much variance there is in the data along that PC.
(1) Original data

(2) Identify the first principal component that has explains the highest amount of variance.

(3) Projecting the data onto that line (the line is the first PC)

(4) After being projected onto a single dimension corresponding to the first PC.
From previous result, some information may be lost in the transformation. However, the most important axis has been kept, which incorporates information from both x1 and x2.

The second PC does its best to capture the variance in the data that is not captured by the first PC.

If PCA was performed on this dataset and project the original dataset onto the first two principal components, then no information would be lost.

The dataset after projection onto the first two principal components is the transformation of the original dataset.
PCA for compressing wine data

There are 178 wine samples.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alcohol</td>
</tr>
<tr>
<td>2</td>
<td>Malic acid</td>
</tr>
<tr>
<td>3</td>
<td>Ash</td>
</tr>
<tr>
<td>4</td>
<td>Alkalinity of ash</td>
</tr>
<tr>
<td>5</td>
<td>Magnesium</td>
</tr>
<tr>
<td>6</td>
<td>Total phenols</td>
</tr>
<tr>
<td>7</td>
<td>Flavanoids</td>
</tr>
<tr>
<td>8</td>
<td>Nonflavanoid phenols</td>
</tr>
<tr>
<td>9</td>
<td>Proanthocyanins</td>
</tr>
<tr>
<td>10</td>
<td>Color intensity</td>
</tr>
<tr>
<td>11</td>
<td>Hue</td>
</tr>
<tr>
<td>12</td>
<td>OD280/OD315 of diluted wines</td>
</tr>
<tr>
<td>13</td>
<td>Proline</td>
</tr>
</tbody>
</table>

Table 1.1: Chemical analysis of wine data

The objective is to reduce the number of descriptors in this data.
PCA for compressing chemical wine data

![Graph showing the cumulative sum of eigenvalues for PCA compression of chemical wine data.](image)
Artificial Neural Networks
Artificial Neural Networks

The main idea is to break down a complex function into many simple ones combined.

Simple multilayer perceptron:
- there can be many hidden layers
- and many output neurons
- feed forward!

Learning:
- given known inputs and outputs, minimize the error w.r.t. weights!

Weighted sum of inputs $\rightarrow$ Sigmoid
Artificial Neural Networks

More complex designs can include feedback...

recurrent network:
• time series
• speech recognition
• ...

or more trickeries → convolutional networks for image recognition
K-means clustering
Clustering: illustration (1/4)

How many clusters do you expect?
Clustering: illustration (2/4)
Clustering: illustration (3/4)
Clustering: illustration (4/4)

Search for outlier?
K-means clustering requires the definition of the number of clusters (K) that the data needs to be partitioned into.

Flowchart for K-means clustering:

1. Start
2. Divide dataset into K clusters
3. Assign data points to the initial clusters
4. Compute mean for each cluster
5. Assign each data point to closest centroids
6. Change in assignment of clusters
   - Yes: Go back to step 4
   - No: Stop

Flowchart for K-means clustering (1/3)
K-means clustering (2/3)

1. **Initial data**
2. **Partitioning of data into three random clusters**
3. **Change in cluster assignment**
4. **No change in cluster assignment**
5. **Final clustering of data**
6. **Re-partitioning of data into three clusters based on the proximity to centroids**
7. **Calculation of centroids (mean) of each cluster**
K-means clustering (3/3)

• This animation illustrates how $k$-means algorithm clusters data.

Randomly initialized centroids of 4 clusters.

Evaluating clustering algorithm on wine data

The Davies–Bouldin index (DBI) is one of metrics for evaluating clustering algorithms.

![Davies Bouldin Index (DBI) vs Number of clusters (K)](image)
K-means to cluster wine data
K Nearest Neighbor classifier
KNN Classifier

Given three classes, find the class for $x_u$

In this case use the Euclidean distance and a value of $K = 5$
KNN advantages and disadvantages

Advantages:
- Simple in implementation
- Better than “only” nearest neighbor method
- Transparent and understandable
- Works well for simple problems

Disadvantages:
- Selecting K value
- Distance metrics may not be optimal for high dimensional data
- Expensive in computation
Genetic Algorithm
Genetic Algorithm

Sometimes the problem is so complex we cannot learn it through gradients...

... so we can just evolve the solution!

1. system descriptor: DNA
2. create random population
3. evaluate **fitness**
4. mix & mutate → new population
5. goto 3 – repeat forever!