Lecture 14
Machine Learning. K-means, kNN
Contents

K-means clustering
K-Nearest Neighbour
Understanding states in the power system is established through observation of inputs and outputs without regard to the physical electrotechnical relations between the states.

Adding knowledge about the electrotechnical rules means adding heuristics to the learning.

Given a set of examples (the learning set (LS)) of associated input/output pairs, derive a general rule representing the underlying input/output relationship, which may be used to explain the observed pairs and/or predict output values for any new unseen input.
Supervised learning - a preview

In the scope of this course, we will be studying three forms of supervised learning.

• Decision Trees
  *Overview and practical work on exercise session.*

• Statistical methods – k-Nearest Neighbour
  *Overview and practical work on exercise session. Also included in Project Assignment*
  *kNN algorithm can also be used for unsupervised clustering.*

• Artificial Neural Networks
  *Overview only, no practical work.*
Lazy vs. Eager learning

In Eager learning, the Training set is pre-classified. All objects in the Learning set are clustered with regards to their neighbours. Ex.: Artificial Neural Networks.

In Lazy learning, only when a new object is input to the algorithm, the distance is calculated. Ex.: k Nearest Neighbours.
Contents

K-means clustering
K-Nearest Neighbour
K-means clustering

K-means clustering involves creating clusters of data. It is iterative and continues until no more clusters can be created. It requires the value of k to be defined at start.

Consider for instance a table like the following:

<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>
Plotted the data looks something like
K-means clustering (continued)

In k means clustering, first pick k mean points randomly in the space

Calculate the Euclidean distance from each data point in dataset to the mean points

Assign a data point to its closest mean point

Recalculate means

Once ended, we have k clusters
Contents

K-means clustering

k-Nearest Neighbour
The k Nearest Neighbour algorithm

The k Nearest Neighbour algorithm is a way to classify objects with attributes to its nearest neighbour in the Learning set.

In kNN method, the k nearest neighbours are considered.

”Nearest” is measured as distance in Euclidean space.
k-Nearest Neighbour classification

Assuming instead a table like this where we have labels to "clusters"

<table>
<thead>
<tr>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
<th>Species</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.1</td>
<td>3.5</td>
<td>1.4</td>
<td>0.2</td>
<td>iris setosa</td>
</tr>
<tr>
<td>4.9</td>
<td>3.0</td>
<td>1.4</td>
<td>0.2</td>
<td>iris setosa</td>
</tr>
<tr>
<td>4.7</td>
<td>3.2</td>
<td>1.3</td>
<td>0.2</td>
<td>iris setosa</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>7.0</td>
<td>3.2</td>
<td>4.7</td>
<td>1.4</td>
<td>iris versicolor</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>6.3</td>
<td>3.3</td>
<td>6.0</td>
<td>2.5</td>
<td>iris virginica</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
K-Nearest Neighbour algorithm

Given a new set of measurements, perform the following test:

Find (using Euclidean distance, for example), the $k$ nearest entities from the training set. These entities have known labels. The choice of $k$ is left to us.

Among these $k$ entities, which label is most common? That is the label for the unknown entity.
Example from Automatic Learning techniques in Power Systems

One Machine Infinite Bus (OMIB) system

- Assuming a fault close to the Generator will be cleared within 155 ms by protection relays
- We need to identify situations in which this clearing time is sufficient and when it is not
- Under certain loading situations, 155 ms may be too slow.

Source: Automatic Learning techniques in Power Systems, L. Wehenkel
OMIB – further information

In the OMIB system the following parameters influence security

- Amount of active and reactive power of the generator (P, Q)
- Amount of load nearby the generator (PI)
- Voltage magnitudes at the load bus and at the infinite bus
- Short-circuit reactance $X_{inf}$, representing the effect of variable topology in the large system represented by the infinite bus.

In the example, Voltages at generator and Infinite bus are assumed similar and constant for simplicity.
Our database of objects with attributes

In a simulator, we randomly sample values for $P_u$ and $Q_u$ creating a database with 5000 samples (objects) and for each object we have a set of attributes $(P_u, Q_u, V_1, P_1, V_{inf}, X_{inf}, CCT)$ as per below.

Source: Automatic Learning techniques in Power Systems, L. Wehenkel
Plot of database content

Source: Automatic Learning techniques in Power Systems, L. Wehenkel
In the OMIB example database

Sample 4984, and its neighbours
Error in the 1-NN classification

Distribution of approximation errors in the test set: CCT(SBS)-CCT(1NN) ms

Maximal absolute error = 10 ms
Mean absolute error = 2 ms