A TUTORIAL ON MACHINE LEARNING WITH WEKA

STEFANO PIO ZINGARO
PH.D. STUDENT IN COMPUTER SCIENCE AND ENGINEERING
STEFANOPIO.ZINGARO [AT] UNIBO.IT
CS.UNIBO.IT/~STEFANOPIO.ZINGARO

Except where otherwise noted, content on this site is licensed under a Creative Commons Attribution 4.0 International license.
OVERVIEW

ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt.

MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time.

DEEP LEARNING
Subset of ML in which multilayered neural networks learn from vast amounts of data.
BASIC ASSUMPTIONS FOR THE TUTORIAL

SOME KNOWLEDGE ON... BUT NO KNOWLEDGE ON...

‣ The concept of independent and dependent variables;
‣ The concept of error in statistic;
‣ The concepts of Supervised and Unsupervised Learning;
‣ Boolean variables (TRUE and FALSE).

‣ Programming Languages;
‣ Coding;
‣ Machine Learning techniques;
‣ Neural Networks;
‣ Minimisation or Maximisation of a function.
MACHINE LEARNING VS. DEEP LEARNING

CLASSIC MACHINE LEARNING
How do you engineer the best features?

FEATURES
- Roundness of face;
- Distance between eyes;
- Nose width;
- Eye socket depth;
- Cheek bone structure;
- Jaw line length…

CLASSIFIER ALGORITHMS
- SVM;
- Random Forest;
- Naïve Bayes;
- Decision Tree;
- Regression
- …many more…

DEEP LEARNING
How do you guide the model to find the best features?

DEEP NEURAL NETWORK

Jim Carrey
DEEP LEARNING BREAKTHROUGHS

IMAGE RECOGNITION  SPEECH RECOGNITION

MACHINES ABLE TO MEET OR EXCEED HUMAN IMAGE & SPEECH RECOGNITION (TO SOME EXTEND...)
Deep Learning in Practice

Healthcare: Tumor detection
- Normal
- Tumor

Industry: Agricultural Robotics
- Plant
- Weed

Energy: Oil & Gas

Automotive: Speech interfaces

Finance: Document Classification

Genomics: Sequence analysis
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING
SUPERVISED TRAINING

Bicycle
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING

LABELLED DATA

Human
Bicycle
Strawberry

Bicycle
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING

LABELLED DATA

MODEL WEIGHTS
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING

LABELLED DATA

MODEL WEIGHTS

PREDICTION

Bicycle

Human

Bicycle

Strawberry

Forward Propagation

Strawberry
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING

LABELLED DATA

MODEL WEIGHTS

PREDICTION

Bicycle

Human

Bicycle

Strawberry

Forward Propagation

Strawberry

True OR False

7
FEED FORWARD NEURAL NETWORKS

Supervised Training

Labelled Data

Model Weights

Prediction

Forward Propagation

Supervision

Human

Bicycle

Strawberry

Supervision

Strawberry

True OR False

Supervision
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING

LABELLED DATA
Human
Bicycle
Strawberry

MODEL WEIGHTS
Forward Propagation

PREDICTION
Strawberry

True OR False
ERROR

We are learning here!
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING

LABELLED DATA

MODEL WEIGHTS

PREDICTION

Supervision

Forward Propagation

Backward Propagation

We are learning here!

True OR False

ERROR

Bicycle

Strawberry

Human

Bicycle

Strawberry
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING

LABELLED DATA

MODEL WEIGHTS

PREDICTION

Supervision

Forward Propagation

Backward Propagation

True OR False

ERROR

We are learning here!

BLIND TEST
FEED FORWARD NEURAL NETWORKS

SUPervised Training

Labelled Data

Model Weights

Prediction

Blind Test

We are learning here!

True OR False

Forward Propagation

Backward Propagation

Supervision

Bicycle

Human

Bicycle

Strawberry

Strawberry
FEED FORWARD NEURAL NETWORKS

SUPERVISED TRAINING

LABELLED DATA

MODEL WEIGHTS

PREDICTION

PREDICTION

Forward Propagation

Backward Propagation

We are learning here!

Supervision

TRAINED MODEL

Forward Propagation

True OR False

ERROR

BLIND TEST

?????

Supervision

Bicycle

Human

Bicycle

Strawberry

Strawberry

Human

Bicycle

Strawberry

Bicycle

?????
FEED FORWARD NEURAL NETWORKS

**Supervised Training**

- **Labelled Data**
  - Human
  - Bicycle
  - Strawberry

- **Model Weights**
  - Forward Propagation
  - Backward Propagation

- **Prediction**
  - Strawberry

**Blind Test**

- **Trained Model**
  - Forward Propagation

- **Prediction**
  - Bicycle

**We are learning here!**

- **True OR False**
  - ????
FEED FORWARD NEURAL NETWORKS

LABELLED DATA

SUPERVISED TRAINING

MODEL WEIGHTS

Forward Propagation

Backward Propagation

TRAINED MODEL

BLIND TEST

PREDICTION

Supervision

We are learning here!

True OR False

ERROR
WEKA — A DATA MINING TOOL

WEKA is developed by the University of Waikato (New Zealand) under the GNU General Public License (GPL).

It is written in the Java™ object-oriented programming language and provides a GUI for interacting with data files and producing visual results.
WEKA is developed by the University of Waikato (New Zealand) under the GNU General Public License (GPL).

It is written in the Java™ object-oriented programming language and provides a GUI for interacting with data files and producing visual results.

FOR THE AIM OF THE TUTORIAL WE WILL USE JUST THE “Explorer” TOOL
<table>
<thead>
<tr>
<th>Classification</th>
<th>Model</th>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Multivariate Linear Regression</td>
<td>House Pricing</td>
</tr>
<tr>
<td></td>
<td>Decision Tree</td>
<td>Campaign</td>
</tr>
<tr>
<td>Clustering</td>
<td><em>k</em>-means</td>
<td>Behaviour Analysis</td>
</tr>
<tr>
<td></td>
<td>Neural Network VS. KNN</td>
<td>Breast Cancer</td>
</tr>
</tbody>
</table>
A SIMPLE EXAMPLE OF PREDICTION

The price of the house (the dependent variable) is the result of many independent variables:

› the square footage of the house;
› the size of the lot;
› whether granite is in the kitchen;
› bathrooms are upgraded;
› etc.

We create a model based on other comparable houses in the neighbourhood and what they sold for, then put the values of our own house into this model to produce an expected price.
### COLLECTING THE DATA

<table>
<thead>
<tr>
<th>House size (square feet)</th>
<th>Lot size</th>
<th>Bedrooms</th>
<th>Granite</th>
<th>Upgraded bathroom?</th>
<th>Selling price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3529</td>
<td>9191</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>205000</td>
</tr>
<tr>
<td>3247</td>
<td>10061</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>224900</td>
</tr>
<tr>
<td>4032</td>
<td>10150</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>197900</td>
</tr>
<tr>
<td>2397</td>
<td>14156</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>189900</td>
</tr>
<tr>
<td>2200</td>
<td>9600</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>195000</td>
</tr>
<tr>
<td>3536</td>
<td>19994</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>325000</td>
</tr>
<tr>
<td>2983</td>
<td>9365</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>230000</td>
</tr>
<tr>
<td>3198</td>
<td>9669</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>???</td>
</tr>
</tbody>
</table>
## COLLECTING THE DATA

<table>
<thead>
<tr>
<th>House size (square feet)</th>
<th>Lot size</th>
<th>Bedrooms</th>
<th>Granite</th>
<th>Upgraded bathroom?</th>
<th>Selling price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3529</td>
<td>9191</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>205000</td>
</tr>
<tr>
<td>3247</td>
<td>10061</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>224900</td>
</tr>
<tr>
<td>4032</td>
<td>10150</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>197900</td>
</tr>
<tr>
<td>2397</td>
<td>14156</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>189900</td>
</tr>
<tr>
<td>2200</td>
<td>9600</td>
<td>4</td>
<td>0</td>
<td>1</td>
<td>195000</td>
</tr>
<tr>
<td>3536</td>
<td>19994</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>325000</td>
</tr>
<tr>
<td>2983</td>
<td>9365</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>230000</td>
</tr>
<tr>
<td>9669</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td></td>
<td>???</td>
</tr>
</tbody>
</table>
## COLLECTING THE DATA

<table>
<thead>
<tr>
<th>House size (square feet)</th>
<th>Lot size</th>
<th>Bedrooms</th>
<th>Granite</th>
<th>Upgraded bathroom?</th>
<th>Selling price</th>
</tr>
</thead>
<tbody>
<tr>
<td>3529</td>
<td>9191</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>205000</td>
</tr>
<tr>
<td>3247</td>
<td>10061</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>224900</td>
</tr>
<tr>
<td>4032</td>
<td>10150</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>197900</td>
</tr>
<tr>
<td>2397</td>
<td>14156</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>189900</td>
</tr>
<tr>
<td>2200</td>
<td>9600</td>
<td>4</td>
<td>0</td>
<td></td>
<td>195000</td>
</tr>
<tr>
<td>3536</td>
<td>19994</td>
<td>6</td>
<td>1</td>
<td>1</td>
<td>325000</td>
</tr>
<tr>
<td>2983</td>
<td>9365</td>
<td>5</td>
<td>0</td>
<td>1</td>
<td>230000</td>
</tr>
<tr>
<td>3198</td>
<td>9669</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

**TESTING DATA**
PREPROCESS THE DATA WITH WEKA
LOAD THE DATASET

- Preprocess Tab
PREPROCESS THE DATA WITH WEKA

- Preprocess Tab
- Open File:
  - CSV format
  - XLS format
  - ARFF format
PREPROCESS THE DATA WITH WEKA

LOAD THE DATASET

- **Numerical variables** data with value representable with numbers
- **Visualize All** shows all graphics at once
CLASSIFY THE DATA WITH WEKA LINEAR REGRESSION MODEL

- Classify Tab
CLASSIFY THE DATA WITH WEKA LINEAR REGRESSION MODEL

- Classify Tab
- Use training set
CLASSIFY THE DATA WITH WEKA
LINEAR REGRESSION MODEL

- Classify Tab;
- Use training set;
- Class = sellingPrice;
CLASSIFY THE DATA WITH WEKA LINEAR REGRESSION MODEL

- Classify Tab;
- Use training set;
- Class = sellingPrice;
- Start building the model
CLASSIFY THE DATA WITH WEKA LINEAR REGRESSION MODEL

- Classify Tab;
- Use training set;
- Class = sellingPrice;
- Start building the model

Prediction formulae:

\[-26.6882 \times \text{houseSize} + 7.0551 \times \text{lotSize} + 43166.0767 \times \text{bedrooms} + 42292.0991 \times \text{bathroom} -21661.1288\]
FINAL PREDICTION

\[
\text{sellingPrice} = -26.68 \times [\text{houseSize} = 3198] + 7.05 \times [\text{lotSize} = 9669] + 43.166.07 \times [\text{bedrooms} = 5] + 42.292.09 \times [\text{bathroom} = 1] - 21.661.12 = 219.328.25
\]
ANOTHER EXAMPLE OF CLASSIFICATION
CAR DEALERSHIP

The dealership is starting a promotional campaign, whereby it is trying to push a two-year extended warranty to its past customers.

The dealership has done this before and has gathered 4,500 data points from past sales of extended warranties.

The attributes in the data set are:

- Income bracket [0=$0-$30k, 1=$31k-$40k, 2=$41k-$60k, 3=$61k-$75k, 4=$76k-$100k, 5=$101k-$150k, 6=$151k-$500k, 7=$501k+]
- Year/month first car bought
- Year/month most recent car bought
- Whether they responded or not to the extended warranty offer in the past
PREPROCESS THE DATA WITH WEKA
LOAD THE DATASET FOR TRAINING

- Nominal variables – labelled data
PREPROCESS THE DATA WITH WEKA
LOAD THE DATASET FOR TRAINING

- Nominal variables – labelled data
- 1500 instances
CLASSIFY THE DATA WITH WEKA

- Classify Tab;
- Use training set;
- Class = responded;
- Start building the model
CLASSIFY THE DATA WITH WEKA

Decision Tree Model Testing

- Supplied test set;
CLASSIFY THE DATA WITH WEKA
DECISION TREE MODEL TESTING

- Supplied test set;
- Start testing the model;
CLASSIFY THE DATA WITH WEKA

- Supplied test set;
- Start testing the model;
- Compare models accuracy between train and test.
AN EXAMPLE OF CLUSTERING
CAR DEALERSHIP BEHAVIOUR ANALYSIS

The dealership has kept track of how people walk through the dealership and the showroom, what cars they look at, and how often they ultimately make purchases.

They are hoping to mine this data by finding patterns in the data and by using clusters to determine if certain behaviours in their customers emerge.
CLUSTERING THE DATA WITH WEKA
K-MEANS BEHAVIOUR ANALYSIS

- Cluster Tab;
- Use training set;
- No Class;
- Start
CLUSTERING THE DATA WITH WEKA
K-MEANS BEHAVIOUR ANALYSIS

- Cluster Tab;
- Use training set;
- No Class;
- Start;
- Evaluate patterns.
<table>
<thead>
<tr>
<th>Attribute</th>
<th>Full Data (100.0)</th>
<th>Cluster# 0 (26.0)</th>
<th>Cluster# 1 (27.0)</th>
<th>Cluster# 2 (5.0)</th>
<th>Cluster# 3 (14.0)</th>
<th>Cluster# 4 (28.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dealership</td>
<td>0.6</td>
<td>0.9615</td>
<td>0.6667</td>
<td>1</td>
<td>0.8571</td>
<td>0</td>
</tr>
<tr>
<td>Showroom</td>
<td>0.72</td>
<td>0.6923</td>
<td>0.6667</td>
<td>0</td>
<td>0.5714</td>
<td>1</td>
</tr>
<tr>
<td>ComputerSearch</td>
<td>0.43</td>
<td>0.6538</td>
<td>0</td>
<td>1</td>
<td>0.8571</td>
<td>0.3214</td>
</tr>
<tr>
<td>M5</td>
<td>0.53</td>
<td>0.4615</td>
<td>0.963</td>
<td>1</td>
<td>0.7143</td>
<td>0</td>
</tr>
<tr>
<td>3Series</td>
<td>0.55</td>
<td>0.3846</td>
<td>0.4444</td>
<td>0.8</td>
<td>0.0714</td>
<td>1</td>
</tr>
<tr>
<td>Z4</td>
<td>0.45</td>
<td>0.5385</td>
<td>0</td>
<td>0.8</td>
<td>0.5714</td>
<td>0.6786</td>
</tr>
<tr>
<td>Financing</td>
<td>0.61</td>
<td>0.4615</td>
<td>0.6296</td>
<td>0.8</td>
<td>1</td>
<td>0.5</td>
</tr>
<tr>
<td>Purchase</td>
<td>0.39</td>
<td>0</td>
<td>0.5185</td>
<td>0.4</td>
<td>1</td>
<td>0.3214</td>
</tr>
</tbody>
</table>

Time taken to build model (full training data): 0.01 seconds

=== Model and evaluation on training set ===

Clustered Instances

<table>
<thead>
<tr>
<th>0</th>
<th>26 (26%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>27 (27%)</td>
</tr>
<tr>
<td>2</td>
<td>5 (5%)</td>
</tr>
<tr>
<td>3</td>
<td>14 (14%)</td>
</tr>
<tr>
<td>4</td>
<td>28 (28%)</td>
</tr>
</tbody>
</table>
EXAMPLE CONCLUSION
K-MEANS BEHAVIOUR ANALYSIS

‣ **Cluster 0**— This group we can call the “Dreamers,” as they appear to wander around the dealership, looking at cars parked outside on the lots, but trail off when it comes to coming into the dealership, and worst of all, they don’t purchase anything.

‣ **Cluster 1**— We’ll call this group the “M5 Lovers” because they tend to walk straight to the M5s, ignoring the 3-series cars and the Z4. However, they don’t have a high purchase rate – only 52 percent. This is a potential problem and could be a focus for improvement for the dealership, perhaps by sending more salespeople to the M5 section.

‣ **Cluster 2**— This group is so small we can call them the “Throw-Aways” because they aren’t statistically relevent, and we can’t draw any good conclusions from their behaviour. (This happens sometimes with clusters and may indicate that you should reduce the number of clusters you’ve created).

‣ **Cluster 3**— This group we’ll call the “BMW Babies” because they always end up purchasing a car and always end up financing it. Here’s where the data shows us some interesting things: It appears they walk around the lot looking at cars, then turn to the computer search available at the dealership. Ultimately, they tend to buy M5s or Z4s (but never 3-series). This cluster tells the dealership that it should consider making its search computers more prominent around the lots (outdoor search computers?), and perhaps making the M5 or Z4 much more prominent in the search results. Once the customer has made up his mind to purchase the vehicle, he always qualifies for financing and completes the purchase.

‣ **Cluster 4**— This group we’ll call the “Starting Out With BMW” because they always look at the 3-series and never look at the much more expensive M5. They walk right into the showroom, choosing not to walk around the lot and tend to ignore the computer search terminals. While 50 percent get to the financing stage, only 32 percent ultimately finish the transaction. The dealership could draw the conclusion that these customers looking to buy their first BMWs know exactly what kind of car they want (the 3-series entry-level model) and are hoping to qualify for financing to be able to afford it. The dealership could possibly increase sales to this group by relaxing their financing standards or by reducing the 3-series prices.
NEURAL NETWORKS VS. K-NEAREST NEIGHBOUR BREAST CANCER CLASSIFICATION

Let's look at the differences between adopting two classification techniques, one will be the neural network multilayer perceptron classification, compared with k-nearest neighbour model.

- Dataset: breast-cancer.arff (286 instances)
- Nominal Class: no-recurrence-events | recurrence-events
- Classification with: Multilayer Perceptron | KNN (IBk in Weka)
- Cross Validation Test Options: takes 10% of the dataset for testing in 10 folds, computing the mean.
- Pay attention to the False Positive rate, which are the differences?
Apply K-nearest neighbour Model (IBk in the choose button) to the problem of car dealership:

- Start parameter \( KNN = 5 \);
- Try decreasing KNN.
REFERENCES

- Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. [PDF]

- The WEKA software. [LINK]

- Wikipedia – Multivariate Linear Regression Model. [LINK]

- Wikipedia – Decision Tree. [LINK]

- Wikipedia – K-means. [LINK]


- Wikipedia – K-nearest neighbour. [LINK]

- Web version of the tutorial by Michael Abernethy, IBM. – [LINK]