

---

# 02450 Project 1

Report

---

by Karol Dzitkowski  
Marco Becattini

Technical University of Denmark  
Department of Applied Mathematics and Computer Science  
Introduction to Machine Learning and Data Mining  
Tue Herlau  
25th September 2014



## **Abstract**

The objective of this report is to apply the methods we learned in the first section of the course on "Data: Feature extraction, and visualization" on letter recognition data set to get a basic understanding of the data prior to the further analysis. This report includes a description of the dataset, detailed explanation of the attributes of the data, data visualizations and a conclusion what we found out about the data set.



# Data set

## Description

**Title** Letter Image Recognition Data

**Characteristics** Multivariate

**Number of Instances** 20000

**Attribute Characteristics** Integer

**Number of Attributes** 17

**Missing Values?** No

The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts) which were then scaled to fit into a range of integer values from 0 through 15.

Data was obtained from a website:

<http://archive.ics.uci.edu/ml/datasets/Letter+Recognition>

and was a part of an article "Letter Recognition Using Holland-style Adaptive Classifier" available on website:

<http://link.springer.com/article/10.1007/BF00114162>

The research for this article investigated the ability of several variations of Holland-style adaptive classifier systems to learn to correctly guess the letter categories associated with vectors of 16 integer attributes extracted from raster scan images of the letters. The best accuracy obtained was a little over 80%. We typically train on the first 16000 items and then use the resulting model to predict the letter category for the remaining 4000. See the article cited above for more details.

## Analysis

- The aim of our future machine learning algorithm working on data set is classification, that is to assign a set of values (our data record) to a capital letter in the English alphabet. It will be probably implemented as some kind of neural network. Classification is a hard problem for a letter recognition, because there is no direct algorithm of doing that, what is the reason for implementing artificial intelligence based on neural network.
- Clustering groups similar characters together, what can be implemented on our data using some sort of N dimensional metric, for example Euclidean or another appropriate similarity measure for records describing letters.
- Association rule discovery will be considered here as a way of predicting a probable value of other attribute values basing on some which are known. We will try to find a set of rules which can tell us a range the other values will be in or the most likely value. For example:

$$\{x = 5, y = 6\} \Rightarrow z \in [-1, 1]$$

or

$$\{x = 1, y = 4, z = 5\} \Rightarrow letter = "a" or "e"$$

- Anomaly detection is used to identify records corresponding to a particular letter with a significant deviation in some or all attributes. It will be helpful to identify unusual patterns of some letters.

# Attributes

## Description

Each record in the data set contains a letter name and 16 attributes describing a letter from English alphabet. This attributes were scaled lineary to a range of integer values from 0 to 15.

No	Name	Description	Type
1.	lettr	capital letter	discrete/nominal (A to Z)
2.	x-box	horizontal position of box	discrete/ratio (integer)
3.	y-box	vertical position of box	discrete/ratio (integer)
4.	width	width of box	discrete/ratio (integer)
5.	high	height of box	discrete/ratio (integer)
6.	onpix	total # on pixels	discrete/ratio (integer)
7.	x-bar	mean x of on pixels in box	discrete/ratio (integer)
8.	y-bar	mean y of on pixels in box	discrete/ratio (integer)
9.	x2bar	mean x variance	discrete/ratio (integer)
10.	y2bar	mean y variance	discrete/ratio (integer)
11.	xybar	mean x y correlation	discrete/ratio (integer)
12.	x2ybr	mean of $x * x * y$	discrete/ratio (integer)
13.	xy2br	mean of $x * y * y$	discrete/ratio (integer)
14.	x-ege	mean edge count left to right	discrete/ratio (integer)
15.	xegvy	correlation of x-ege with y	discrete/ratio (integer)
16.	y-ege	mean edge count bottom to top	discrete/ratio (integer)
17.	yegvx	correlation of y-ege with x	discrete/ratio (integer)

Table 1.1: Attribute Information

## Analysis

789 A	766 B	736 C	805 D	768 E	775 F	773 G
734 H	755 I	747 J	739 K	761 L	792 M	783 N
753 O	803 P	783 Q	758 R	748 S	796 T	813 U
764 V	752 W	787 X	786 Y	734 Z		

Table 1.2: Class Distribution - Letters



# Visualization

In this section we provide an analysis of the data by means of statistical tools and graphical representation.

For each attribute, we compute basic summary statistics: mean, variance, median, range, standard deviation, as given in the table below:

	Mean	Variance	Median	Range	Std
x-box	4.0236	3.6604	4.0000	15.0000	1.9132
y-box	7.0355	10.9201	7.0000	15.0000	3.3046
width	5.1219	4.0585	5.0000	15.0000	2.0146
high	5.3724	5.1139	6.0000	15.0000	2.2614
onpix	3.5059	4.7981	3.0000	15.0000	2.1905
x-bar	6.8976	4.1048	7.0000	15.0000	2.0260
y-bar	7.5004	5.4073	7.0000	15.0000	2.3254
x2bar	4.6286	7.2898	4.0000	15.0000	2.7000
y2bar	5.1787	5.6683	5.0000	15.0000	2.3808
xybar	8.2820	6.1925	8.0000	15.0000	2.4885
x2ybr	6.4540	6.9225	6.0000	15.0000	2.6311
xy2br	7.9290	4.3290	8.0000	15.0000	2.0806
x-ege	3.0461	5.4407	3.0000	15.0000	2.3325
xegvy	8.3389	2.3924	8.0000	15.0000	1.5467
y-ege	3.6917	6.5899	3.0000	15.0000	2.5671
yegvx	7.8012	2.6162	8.0000	15.0000	1.6175

Table 1.3: Attribute Information

We plot a histogram for each attribute to investigate its distribution