mlp

July 14, 2018

1 Loading dataset

```
In [1]: %plot inline
```

```
In [2]: !cat ../data/winequality.names
```

```
Citation Request:
  This dataset is public available for research. The details are described in [Cortez et al., 2
  Please include this citation if you plan to use this database:

  P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
  Modeling wine preferences by data mining from physicochemical properties.
  In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

  Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016
                [Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf
                [bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

1. Title: Wine Quality

2. Sources
   Created by: Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and

3. Past Usage:

  P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.
  Modeling wine preferences by data mining from physicochemical properties.
  In Decision Support Systems, Elsevier, 47(4):547-553. ISSN: 0167-9236.

  In the above reference, two datasets were created, using red and white wine samples.
  The inputs include objective tests (e.g. PH values) and the output is based on sensory data
  (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality
  between 0 (very bad) and 10 (very excellent). Several data mining methods were applied to mod
  these datasets under a regression approach. The support vector machine model achieved the
  best results. Several metrics were computed: MAD, confusion matrix for a fixed error toleranc
  etc. Also, we plot the relative importances of the input variables (as measured by a sensitiv
  analysis procedure).
```

```
4. Relevant Information:

   The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine
   For more details, consult: http://www.vinhoverde.pt/en/ or the reference [Cortez et al., 200
   Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) v
   are available (e.g. there is no data about grape types, wine brand, wine selling price, etc

   These datasets can be viewed as classification or regression tasks.
   The classes are ordered and not balanced (e.g. there are munch more normal wines than
   excellent or poor ones). Outlier detection algorithms could be used to detect the few excell
   or poor wines. Also, we are not sure if all input variables are relevant. So
   it could be interesting to test feature selection methods.

5. Number of Instances: red wine - 1599; white wine - 4898.

6. Number of Attributes: 11 + output attribute

   Note: several of the attributes may be correlated, thus it makes sense to apply some sort o
   feature selection.

7. Attribute information:

   For more information, read [Cortez et al., 2009].

   Input variables (based on physicochemical tests):
   1 - fixed acidity
   2 - volatile acidity
   3 - citric acid
   4 - residual sugar
   5 - chlorides
   6 - free sulfur dioxide
   7 - total sulfur dioxide
   8 - density
   9 - pH
   10 - sulphates
   11 - alcohol
   Output variable (based on sensory data):
   12 - quality (score between 0 and 10)

8. Missing Attribute Values: None
```

In [3]:
```
!head ../data/winequality-red.csv
```

```
"fixed acidity";"volatile acidity";"citric acid";"residual sugar";"chlorides";"free sulfur dio
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.8;0.88;0;2.6;0.098;25;67;0.9968;3.2;0.68;9.8;5
```

```
7.8;0.76;0.04;2.3;0.092;15;54;0.997;3.26;0.65;9.8;5
11.2;0.28;0.56;1.9;0.075;17;60;0.998;3.16;0.58;9.8;6
7.4;0.7;0;1.9;0.076;11;34;0.9978;3.51;0.56;9.4;5
7.4;0.66;0;1.8;0.075;13;40;0.9978;3.51;0.56;9.4;5
7.9;0.6;0.06;1.6;0.069;15;59;0.9964;3.3;0.46;9.4;5
7.3;0.65;0;1.2;0.065;15;21;0.9946;3.39;0.47;10;7
7.8;0.58;0.02;2;0.073;9;18;0.9968;3.36;0.57;9.5;7
```

In [4]:
```
data = dlmread("../data/winequality-red.csv", ";" ,1, 0); % skip feature names
size(data)
```

```
ans =

   1599     12
```

2 Preparing data

In [5]:
```
% randomize order of data for excluding biases in the dataset
n = rand(length(data),1);
[_ index] = sort(n);
data_rand = data(index, :);
```

In [6]:
```
data_norm = [];
mu = [];
sigma = [];

% normalizing data for optimum use of algorithms
for j = 1:size(data_rand,2),
    mu = [mu; mean(data_rand(:,j))];
    sigma = [sigma; std(data_rand(:,j))];
    data_norm = [data_norm, (data_rand(:,j)- mu(j)*ones(size(data,1),1))/sigma(j)];
end
```

In [7]:
```
% correlation matrix
cor = corr(data_norm)

imagesc(cor);

set(gca, 'XTick', 1:size(cor,2)); % center x-axis ticks on bins
set(gca, 'YTick', 1:size(cor,2)); % center y-axis ticks on bins
set(gca, 'YTickLabel', ["fixed acidity - 1";"volatile acidity - 2";"citric acid - 3";"
                        "chlorides - 5";"free sulfur dioxide - 6";"total sulfur dioxid
                        "pH - 9";"sulphates - 10";"alcohol - 11";"quality - 12"]); % s

title('Correlation Matrix', 'FontSize', 14); % set title
```

```
caxis([-1, 1]);  % colorbar ranging from -1 to 1
colormap('jet'); % set the colorscheme
colorbar;        % enable colorbar
```

```
cor =

 Columns 1 through 6:

   1.0000000  -0.2561309   0.6717034   0.1147767   0.0937052  -0.1537942
  -0.2561309   1.0000000  -0.5524957   0.0019179   0.0612978  -0.0105038
   0.6717034  -0.5524957   1.0000000   0.1435772   0.2038229  -0.0609781
   0.1147767   0.0019179   0.1435772   1.0000000   0.0556095   0.1870490
   0.0937052   0.0612978   0.2038229   0.0556095   1.0000000   0.0055621
  -0.1537942  -0.0105038  -0.0609781   0.1870490   0.0055621   1.0000000
  -0.1131814   0.0764700   0.0355330   0.2030279   0.0474005   0.6676665
   0.6680473   0.0220262   0.3649472   0.3552834   0.2006323  -0.0219458
  -0.6829782   0.2349373  -0.5419041  -0.0856524  -0.2650261   0.0703775
   0.1830057  -0.2609867   0.3127700   0.0055271   0.3712605   0.0516576
  -0.0616683  -0.2022880   0.1099032   0.0420754  -0.2211405  -0.0694084
   0.1240516  -0.3905578   0.2263725   0.0137316  -0.1289066  -0.0506561

 Columns 7 through 12:

  -0.1131814   0.6680473  -0.6829782   0.1830057  -0.0616683   0.1240516
   0.0764700   0.0220262   0.2349373  -0.2609867  -0.2022880  -0.3905578
   0.0355330   0.3649472  -0.5419041   0.3127700   0.1099032   0.2263725
   0.2030279   0.3552834  -0.0856524   0.0055271   0.0420754   0.0137316
   0.0474005   0.2006323  -0.2650261   0.3712605  -0.2211405  -0.1289066
   0.6676665  -0.0219458   0.0703775   0.0516576  -0.0694084  -0.0506561
   1.0000000   0.0712695  -0.0664946   0.0429468  -0.2056539  -0.1851003
   0.0712695   1.0000000  -0.3416993   0.1485064  -0.4961798  -0.1749192
  -0.0664946  -0.3416993   1.0000000  -0.1966476   0.2056325  -0.0577314
   0.0429468   0.1485064  -0.1966476   1.0000000   0.0935948   0.2513971
  -0.2056539  -0.4961798   0.2056325   0.0935948   1.0000000   0.4761663
  -0.1851003  -0.1749192  -0.0577314   0.2513971   0.4761663   1.0000000
```