Machine Learning for Medicine TP 4

Feature Selection
Model Selection

The goal of the TME is to learn various techniques of feature selection.

Data (both data sets are provided)

- Molecular classification of leukemia data set of *Golub et al. 1999* contains gene expressions of 72 patients and 3562 genes.
- Breast cancer data set

You will need to load at least the following packages:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import ElasticNet
from sklearn.svm import LinearSVC
from sklearn.feature_selection import SelectFromModel
from sklearn import linear_model
```

Analysis

Repeat the same analyses for the two data sets.

To read the data:

- For the *Golub et al. 1999* data

```
X = pd.read_csv('data/Golub_X',sep=' ') # Observations
y = pd.read_csv('data/Golub_y',sep=' ') # Classes
```

- For the Breast cancer data

```
X = pd.read_csv('data/Breast.txt',sep=' ')
y = X.as_matrix()[:,30] # Classes
X = X.as_matrix()[:,0:29] # Observations
```

We will use the `sklearn` Python library only.

1. A simple heuristic approach is to delete features whose variance is less then a threshold. Try it (with two different arbitrary thresholds) but do not expect this method to return an optimal performance (although it can be quite efficient on some data sets).

   `http://scikit-learn.org/stable/modules/feature_selection.html`

2. Univariate feature selection with statistical tests to get rid of features which are not statistically significant with respect to the vector of class. Try the `SelectFdr` function that computes p-values for an estimated false discovery rate.

   `http://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SelectFdr.html#sklearn.feature_selection.SelectFdr`

3. L_1-based feature selection is designed to find an optimal solution. The sparsity parameter is important (since it controls the number of non-zero parameters: it too many parameters are kept, no really feature selection; if too few parameters are chosen, it is possible that the accuracy is very poor).

   (a) Logistic regression penalized by the L_1 penalty term

   ```
   linear_model.Lasso(alpha=alpha)
   ```

   (b) A support vector machine penalized by the L_1 penalty term

   ```
   LinearSVC(C=C, penalty="l1", dual=False)
   ```

   (c) Explore the Elastic Net which is a compromise between the L_1 and L_2 penalty terms.

   ```
   ElasticNet(alpha=alpha, l1_ratio=0.7)
   ```

4. How many features do you keep using these different methods? It is quite normal that each method selects a different number of features.

5. What method leads to the best performance (on the given data sets) ?

References

- *The original Lasso paper:*

  Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal. Statist. Soc B., Vol. 58, No. 1, pages 267–288

  `http://statweb.stanford.edu/~tibs/lasso/lasso.pdf`

- T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity. The Lasso and Generalizations.* (a good book)

  `https://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS_corrected_1.4.16.pdf`