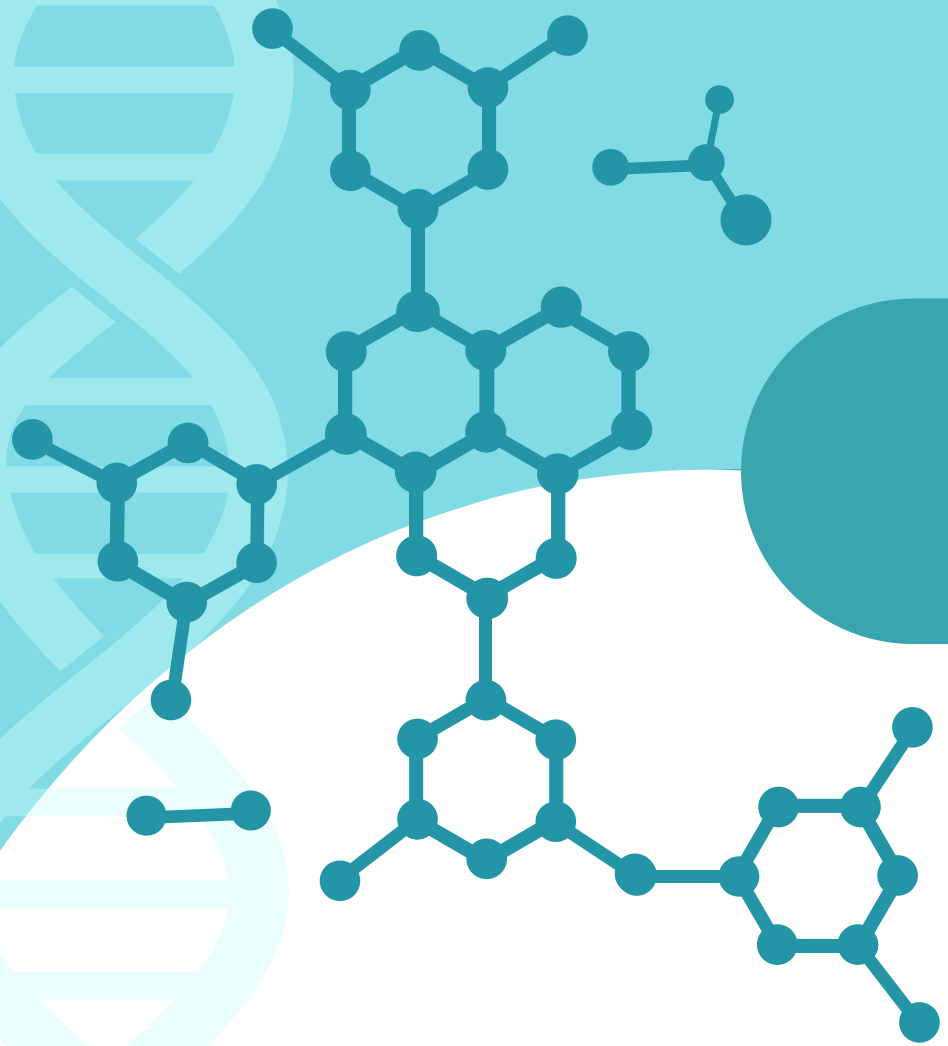
Bioinformatics LAB 5

ML basics

Prof.ssa Elisa Ficarra
Prof.ssa Santa Di Cataldo
Eng. Marta Lovino
Eng. Alessio Mascolini

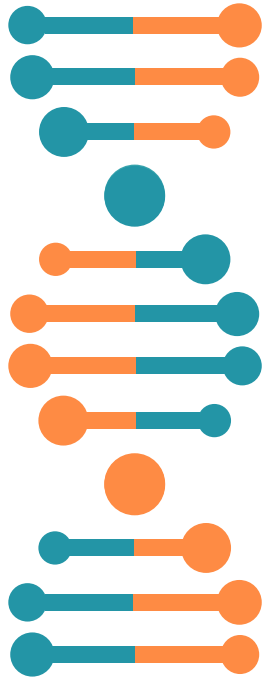
Politecnico di Torino
DAUIN
Dept. of Control and Computer Engineering

LAB 5 - Objectives

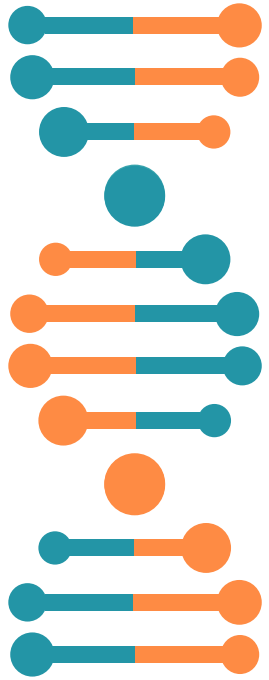


Objectives

- T test statistics
- Bonferroni adjustment
- ENSG notation and common gene name
- Machine learning basics



Install libraries and download dataset



Suggested libraries

`scipy`

`pandas`

`requests`

`scikit-learn`

Gene expression dataset comes from an RNA-seq experiment onto two breast cancer subtypes: Luminal A and Luminal B. Please, download the dataset (***dataset_LUMINAL_A_B.csv***) from the Teaching Portal.

T test statistics

In many cases, we analyze microarrays with “one gene at a time” approach. **That is, for each gene we would like to know if this gene is “different in the two classes”**. In our example the two classes are two breast cancer subtypes (Luminal A and Luminal B). A classic analytical method to answer to this question is to perform Independent Student’s t-test (so called Welch test).

1. **Define hypotheses and level of significance α:**

Null Hypothesis: Means of the two populations are equal, so the two groups are from the same populations. In our example it means that the gene we are investigating is not differentially expressed between Luminal A and Luminal B breast cancer subtypes and we cannot exploit its gene expression to distinguish between the two-cancer subtypes.

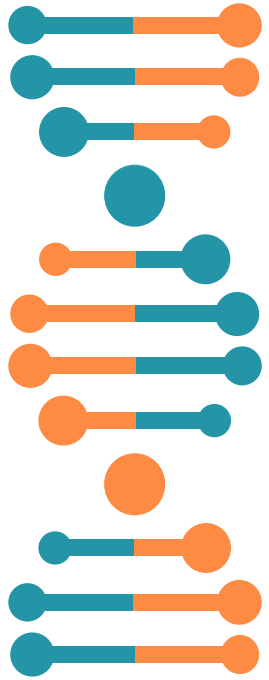
$$H_0 : \mu_1 = \mu_2$$

Alternative Hypothesis: The mean of the two populations are un-equal and the two groups are from different populations. In our example it means that the gene we are investigating is differentially expressed between Luminal A and Luminal B breast cancer subtypes we can exploit its gene expression to distinguish between the two-cancer subtypes.

$$H_1 : \mu_1 \; ! = \; \mu_2$$

Level of significance **α** is defined a priori and it is the risk we are prepared to take in rejecting H_0 when it is in fact true. For example, **α** = 0.05 means a possibility of 5% in rejecting H_0 when in reality H_0 is true.

T test statistics



2. For each gene perform Welch T-test

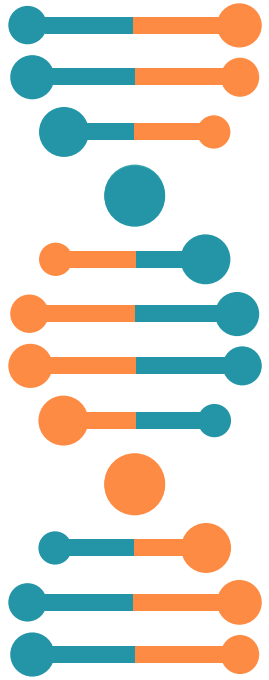
Welch T-test is used to investigate the significance of the difference between the means of two populations. Use scipy library to perform the test. E.g.

```
t_value, p_value = stats.ttest_ind(np.array(Lum_A), np.array(Lum_B), equal_var=False)
```

3. Reject null hypothesis if the p-value is lower or equal to α

If the p value is lower than α, we can reject the null hypothesis and say that the gene is differentially expressed between Luminal A and Luminal B cancer subtypes.

Bonferroni adjustment



However, when we perform multiple statistical tests (in our case more than 1000, one for each gene), the overall probability in rejecting the null hypothesis when actually it is true is given by $\alpha \times G$, where G is the number of genes. This overall error is called **Family-Wise Error Rate (FWER).**

To achieve the global **FWER lower than α**, we can use Bonferroni adjustment that requires to select only genes for which **p value ≤ α/G.**

From the ENSEMBL gene notation to the common gene name

Use the GFT file to translate ENSEMBL gene ID with common gene names

```
1	havana	gene	11869	14409	.	+	.	gene_id "ENSG00000223972";
gene_version "5"; gene_name "DDX11L1"; gene_source "havana"; gene_biotype
"transcribed_unprocessed_pseudogene"; havana_gene "OTTHUMG00000000961"; havana_gene_version "2";
1	havana	transcript	11869	14409	.	+	.	gene_id "ENSG00000223972";
gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; gene_name "DDX11L1";
gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene"; havana_gene
"OTTHUMG00000000961"; havana_gene_version "2"; transcript_name "DDX11L1-002"; transcript_source
"havana"; transcript_biotype "processed_transcript"; havana_transcript "OTTHUMT00000362751";
havana_transcript_version "1"; tag "basic"; transcript_support_level "1";
1	havana	exon	11869	12227	.	+	.	gene_id "ENSG00000223972";
gene_version "5"; transcript_id "ENST00000456328"; transcript_version "2"; exon_number "1"; gene_name
"DDX11L1"; gene_source "havana"; gene_biotype "transcribed_unprocessed_pseudogene"; havana_gene
"OTTHUMG00000000961"; havana_gene_version "2"; transcript_name "DDX11L1-002"; transcript_source
"havana"; transcript_biotype "processed_transcript"; havana_transcript "OTTHUMT00000362751";
havana_transcript_version "1"; exon_id "ENSE00002234944"; exon_version "1"; tag "basic";
transcript_support_level "1";
"1";
```

Scikit-learn library

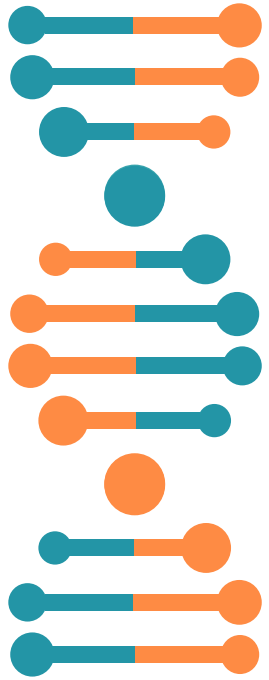
Scikit-learn is a useful library to perform machine learning with Python. In order to work with, you have to import the library in your program:

```
import sklearn
```

Check for useful functions at this link: https://scikit-learn.org/stable/

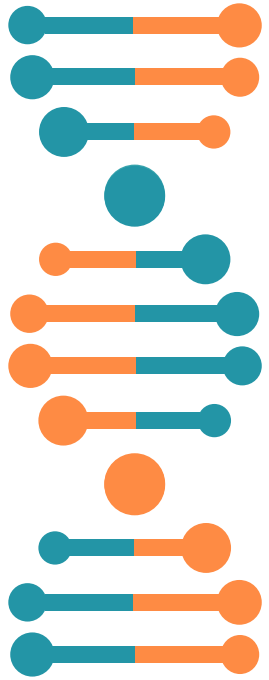
Some useful scikit-learn classes

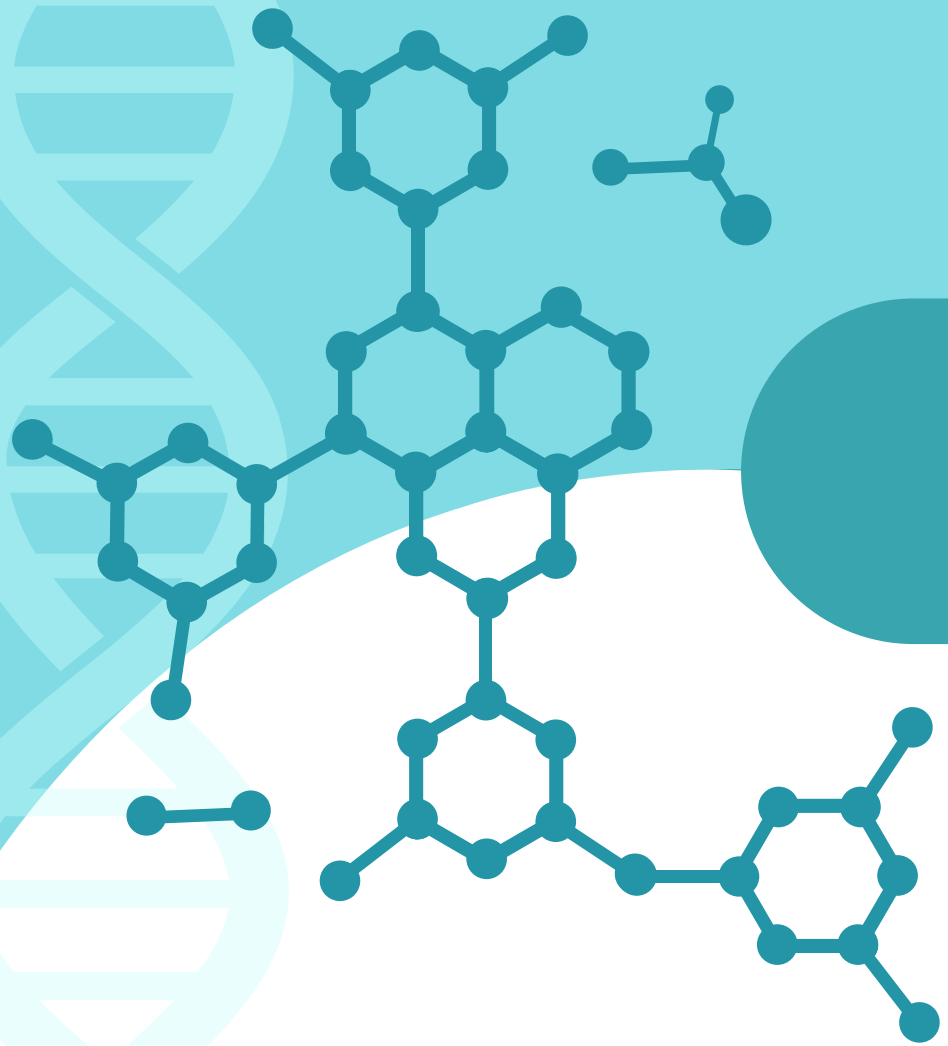
```
sklearn.preprocessing.StandardScaler
sklearn.decomposition.PCA
sklearn.model_selection.train_test_split
sklearn.metrics.accuracy_score
sklearn.metrics.precision_recall_fscore_support
sklearn.neighbors.KNeighborsClassifier
sklearn.svm.SVC
sklearn.ensemble.RandomForestClassifier
sklearn.naive_bayes.GaussianNB
```



PCA + KNN example

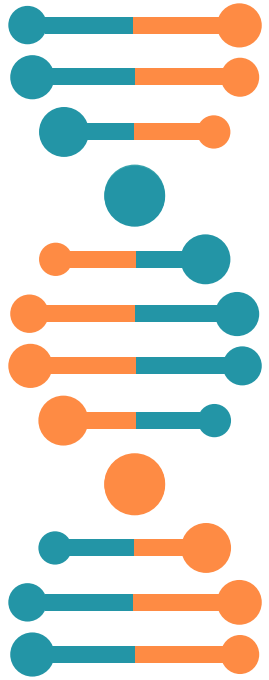
```
# PCA and number of components
pca = PCA(n_components=80) # creates pca object
# fit pca object onto train set features and transform train set
X_train_pca = pca.fit_transform(X_train)
# transform test set using the features fitted onto train_set
X_test_pca = pca.transform(X_test)

# KNN on dataset with PCA
neigh = KNeighborsClassifier(n_neighbors=3) # creates knn object
neigh.fit(X_train_pca, y_train) # fit knn object onto training set
# use knn to predict the outcome onto training set
y_test_predicted = neigh.predict(X_test_pca)
```





LAB 5 - Assignments

Assignment 1: T-test to perform differential gene expression analysis



Download ***dataset_LUMINAL_A_B.csv*** from the Teaching Portal and perform a differential gene expression analysis in order to find which genes are able to distinguish between breast cancer Luminal A subtype and breast cancer Luminal B subtype.

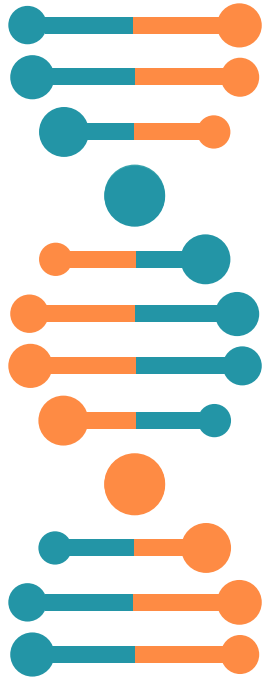
Calculate t-value and p-value for each gene and select only genes for which p-value < adjusted Bonferroni p-value.

Finally, convert differentially expressed genes from ENSEMBL notation to the common name (e.g. ENSG00000268889 is AC008750).

Which genes are differentially expressed between the two populations?

Create now a reduced dataset (from now on referred to as ***reduced_dataset.csv***) made up of all samples with only differentially expressed genes (use common name notation).

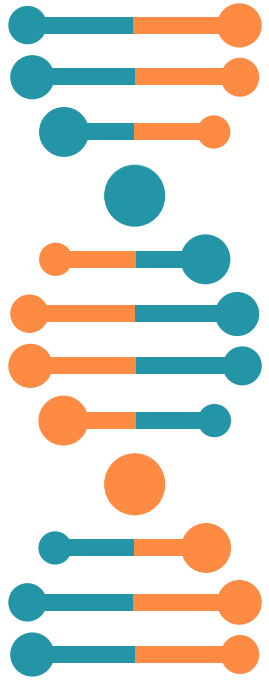
Assignment 2: Use gene expression data to create a classifier for Luminal A / Luminal B breast cancer subtypes



In order to create a Luminal A / Luminal B breast cancer classifier, consider two dataset: ***dataset_LUMINAL_A_B.csv***(the one we provided you) and ***reduced_dataset.csv*** (the one you created in Assignment 1).

1. Divide both dataset_LUMINAL_A_B.csv and reduced_dateset.csv into train set and test set
2. Standardize features of both datasets by removing the mean and scaling to unit variance
3. Perform the dimensionality reduction onto dataset.csv with PCA (principal component Analysis) using 80 features.
4. Train a KNN classifier onto the train set of ***dataset_LUMINAL_A_B.csv*** with PCA
5. Test the classifier obtained at the previous step onto the test set of ***dataset_LUMINAL_A_B.csv*** (remember to apply PCA transformation onto test set)
6. Implement from scratch the following performance metrics: accuracy, precision and recall, F1 score. Compare your results with performance metrics provided by sklearn.metrics

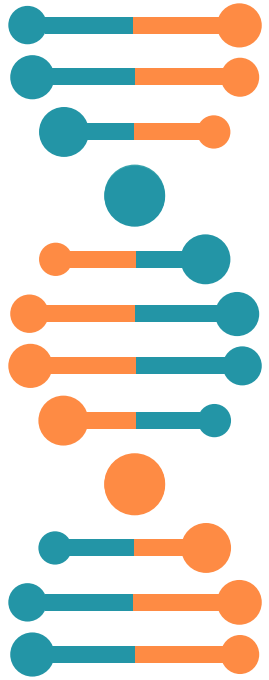
Assignment 2: Use gene expression data to create a classifier for Luminal A / Luminal B breast cancer subtypes



7. Train and test a KNN classifier onto **reduced_dataset.csv**
8. Which are the differences between the performances obtained onto dataset_LUMINAL_A_B.csv with PCA and these obtained onto reduced_dataset.csv?
9. Use reduced_dataset.csv to train and test (with default parameters) SVM, Random Forest and Naïve Bayes classifiers
10. Which classifier provides you the best result on reduced_dataset.csv?

LAB5 – Take home message

- Select the appropriate statistical test according to the specific properties of your dataset. T-test is the simplest one!!
- Key differences between feature selection and dimensionality reduction
- Choose the best classifier according to the characteristics of your dataset.



Questions?

Remember:
no question is
stupid