**Machine learning project**

**서울대학교 4차 산업혁명 아카데미**

**인공지능 에이전트 2학기 기계학습**

**Project description**

The objective of this project is to apply learnt machine learning techniques to the Wisconsin Diagnostic Breast Cancer (WDBC) data. The WDBC data is class labeled, hence it will be a classification problem. The data has two classes (B=Benign, M=Malignant) and 32 attributes, or features.

Using the following machine learning techniques, you are asked to build classifiers to predict the classes of the data. For the machine learning techniques, use 1) Linear regression, 2) Random Forest, 3) Neural network and 4) K-means algorithm.

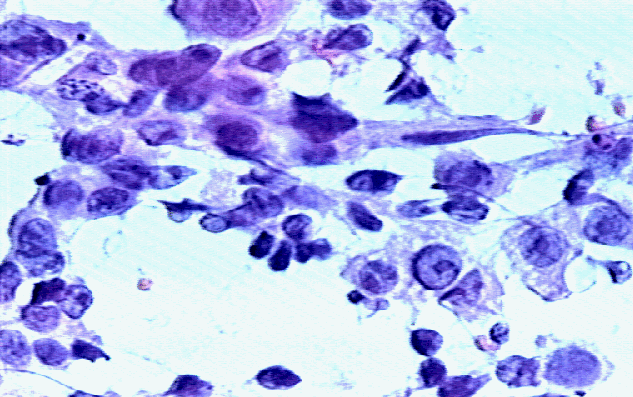
Despite the data has label information, the unsupervised K-means clustering shall be applied to the data to compare it with the classification results (i.e., does the clustering algorithm well differentiate the Benign data from the Malignant data without using the label data?).

**Data Description**

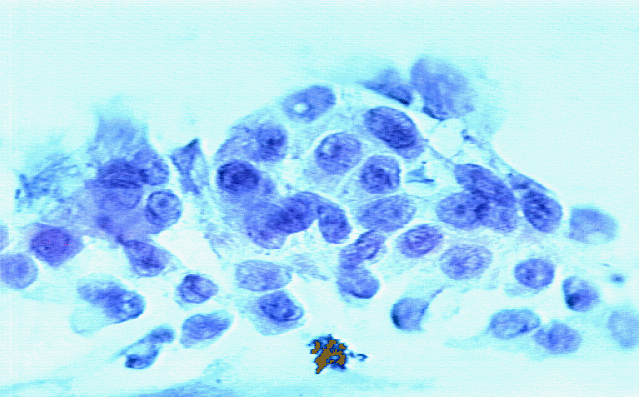
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Data Set Characteristics:** | Multivariate | **Number of Instances:** | 569 | **Area:** | Life |
| **Attribute Characteristics:** | Real | **Number of Attributes:** | 32 | **Date Donated** | 1995-11-01 |
| **Associated Tasks:** | Classification | **Missing Values?** | No | **Number of Web Hits:** | 563646 |

Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. An example figure is shown below.

<Image of Benign tumor>



<Image of Malignant tumor>



**Data file:** “wdbc.data”

**Class distribution:** 357 benign, 212 malignant (total 569 data)

**Attribute Information:**

The attribute or features are measured from each image as shown above. Below is the description of each feature column in the “wdbc.data” file.

1) ID number

2) Diagnosis (M = malignant, B = benign) 🡪 this is the class label

3-32)

Ten real-valued features are computed for each cell nucleus:

a) radius (mean of distances from center to points on the perimeter)

b) texture (standard deviation of gray-scale values)

c) perimeter

d) area

e) smoothness (local variation in radius lengths)

f) compactness (perimeter^2 / area - 1.0)

g) concavity (severity of concave portions of the contour)

h) concave points (number of concave portions of the contour)

i) symmetry

j) fractal dimension ("coastline approximation" - 1)

\* The **mean**, **standard error**, and **"worst" or largest** (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius. All feature values are recoded with four significant digits.

**Missing attribute values:** none

**Project tasks**

1. Build classifiers using logistic Regression, Random Forest and Neural Networks
2. Perform cross validation to measure each methods classification accuracy
3. Discuss the result of each method
   1. Logistic Regression: What are the coefficients of the Linear Regression
   2. Random Forest: What 5 features are the most informative and what are their information strength (i.e., information gain)?
   3. Neural Network: How does the classification accuracy change with different number of hidden layers?
4. Perform K-means clustering
   1. Measure the rand index, and silhouette scores
   2. How does the score change with K={2,3,…, 10}?
   3. Discuss the result of the clustering by comparing it with the classification results

The report shall include the following contents and shall be less than 5 pages:

* Introduction (One paragraph)
* Method
* Results
* Discussion
* Conclusion