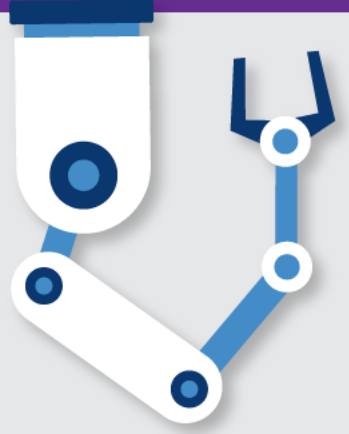
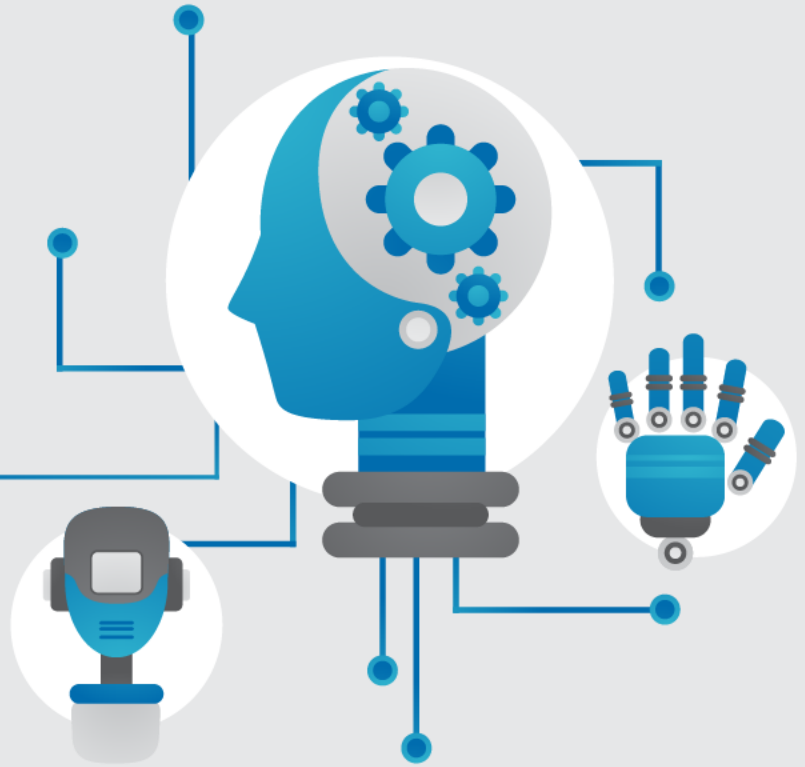


威斯康辛乳癌數據集介紹



威斯康辛乳癌數據集

- 威斯康辛乳癌數據集
  （Breast Cancer Wisconsin (Diagnostic) Data Set）
  是由威斯康辛大學醫院的Dr. William H. Wolberg建立

UCI Machine Learning Repository
Center for Machine Learning and Intelligent Systems

About Citation Policy Donate a Data Set Contact

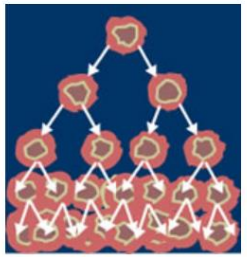
Search
Repository Web
Google

View ALL Data Sets

Breast Cancer Wisconsin (Diagnostic) Data Set

Download: Data Folder, Data Set Description

Abstract: Diagnostic Wisconsin Breast Cancer Database



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:**	1217248

Source:

Creators:

1. Dr. William H. Wolberg, General Surgery Dept.

威斯康辛乳癌數據集

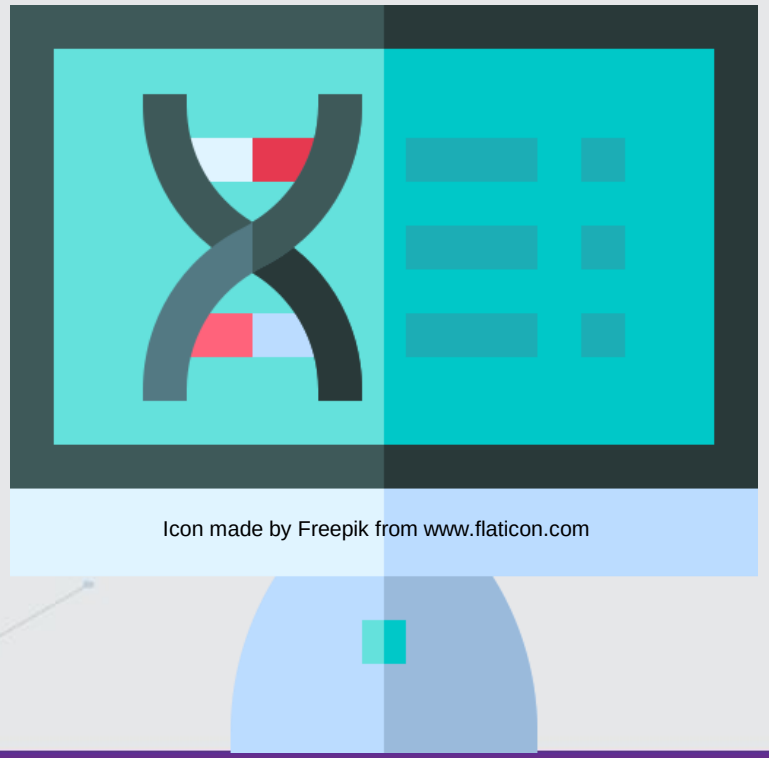
- **數據集總共包含569個惡性或良性腫瘤細胞樣本，並以10種不同屬性呈現**，例如：
  平均半徑（mean radius）、平均面積（mean area）等。
  每個特徵都有三種不同數據：平均值、標準差、最差值，
  共有30個特徵。

- **分為2個類別**
  良性（benign）與惡性（malignant）。

- **類別資料不平均**
  良性有357個樣本，惡性有212個樣本。

威斯康辛乳癌數據集

- 威斯康辛乳癌數據集有以下10種屬性
  1. radius（腫瘤半徑）
  2. texture（細胞核紋路）
  3. perimeter（腫瘤周長）
  4. area（腫瘤面積）
  5. smoothness（平滑程度）
  6. compactness（緊湊度）
  7. concavity（邊緣凹凸度）
  8. concave points（凹凸點）
  9. symmetry（對稱性）
  10. fractal dimension（輪廓估計）



威斯康辛乳癌數據集下載

http://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)

Parent Directory		-
Index	1996-12-03 04:07	326
breast-cancer-wiscon..>	1992-07-16 10:15	19K
breast-cancer-wiscon..>	1992-07-16 14:13	5.5K
unformatted-data	1992-07-16 06:17	21K
wdbc.data	1996-02-05 11:04	121K
wdbc.names	1996-02-05 11:04	4.6K
wpbc.data	1996-02-01 16:00	43K
wpbc.names	1996-02-01 16:00	5.5K

- sklearn套件內建資料集

```
from sklearn.datasets import load_breast_cancer
bunch = load_breast_cancer()
```

威斯康辛乳癌數據集下載

- Kaggle 網頁
  - https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

威斯康辛乳癌數據集處理

- **從sklearn.datasets載入數據資料**
  函式回傳一個bunch物件，bunch是一個類似dictionary的物件。

```
from sklearn.datasets import load_breast_cancer
bunch = load_breast_cancer()
print(bunch)
```

威斯康辛乳癌數據集處理

- 從print出來的Bunch資料中，找到'data', 'target'，就是我們需要的資料與類別。

```
{'data': array([[1.799e+01, 1.038e+01, 1.228e+02, ..., 2.654e-01, 4.601e-01,
        1.189e-01],
       [2.057e+01, 1.777e+01, 1.329e+02, ..., 1.860e-01, 2.750e-01,
        8.902e-02],
       [1.969e+01, 2.125e+01, 1.300e+02, ..., 2.430e-01, 3.613e-01,
        8.758e-02],
       ...,
       [1.660e+01, 2.808e+01, 1.083e+02, ..., 1.418e-01, 2.218e-01,
        7.820e-02],
       [2.060e+01, 2.933e+01, 1.401e+02, ..., 2.650e-01, 4.087e-01,
        1.240e-01],
       [7.760e+00, 2.454e+01, 4.792e+01, ..., 0.000e+00, 2.871e-01,
        7.039e-02]]), 'target': array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 0, 1, 0, 0,
```

威斯康辛乳癌數據集處理

› Bunch物件可以透過以下屬性將儲存的資料取出

- data： 用來訓練的資料
- feature_names：特徵的名稱
- target：分類的標籤
- target_names：標籤的名稱
- DESCR：數據集的詳細介紹
- filename：乳癌數據檔案所在的位置

威斯康辛乳癌數據集處理

- 訓練的資料

```
data = bunch.data
print(data)
print(type(data))
print(data.shape)
```

```
[[1.799e+01 1.038e+01 1.228e+02 ... 2.654e-01 4.601e-01 1.189e-01]
 [2.057e+01 1.777e+01 1.329e+02 ... 1.860e-01 2.750e-01 8.902e-02]
 [1.969e+01 2.125e+01 1.300e+02 ... 2.430e-01 3.613e-01 8.758e-02]
 ...
 [1.660e+01 2.808e+01 1.083e+02 ... 1.418e-01 2.218e-01 7.820e-02]
 [2.060e+01 2.933e+01 1.401e+02 ... 2.650e-01 4.087e-01 1.240e-01]
 [7.760e+00 2.454e+01 4.792e+01 ... 0.000e+00 2.871e-01 7.039e-02]]
<class 'numpy.ndarray'>
(569, 30)
```

威斯康辛乳癌數據集處理

- 特徵的名稱

```
feature_names = bunch.feature_names
print(feature_names)
print(type(feature_names))
print(feature_names.shape)
```

```
['mean radius' 'mean texture' 'mean perimeter' 'mean area'
 'mean smoothness' 'mean compactness' 'mean concavity'
 'mean concave points' 'mean symmetry' 'mean fractal dimension'
 'radius error' 'texture error' 'perimeter error' 'area error'
 'smoothness error' 'compactness error' 'concavity error'
 'concave points error' 'symmetry error' 'fractal dimension error'
 'worst radius' 'worst texture' 'worst perimeter' 'worst area'
 'worst smoothness' 'worst compactness' 'worst concavity'
 'worst concave points' 'worst symmetry' 'worst fractal dimension']
<class 'numpy.ndarray'>
(30,)
```

威斯康辛乳癌數據集處理

- 分類的標籤

```
labels = bunch.target
print(labels)
print(type(labels))
print(labels.shape)
```

```
[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 1 0 0 0 0 0 0 0 0 1 0 1 1 1 1 1 0 0 1 0 0 1 1 1 1 0 1 0 0 1 1 1 1 0 1 0 0
 1 0 1 0 0 1 1 1 0 0 1 0 0 0 1 1 1 0 1 1 0 0 1 1 1 0 0 1 1 1 1 0 1 1 0 1 1
 1 1 1 1 1 1 0 0 0 1 0 0 1 1 1 0 0 1 0 1 0 0 1 0 0 1 1 0 1 1 0 1 1 1 1 0 1
 1 1 1 1 1 1 1 1 0 1 1 1 1 0 0 1 0 1 1 0 0 1 1 0 0 1 1 1 1 0 1 1 0 0 0 1 0
 1 0 1 1 1 0 1 1 0 0 1 0 0 0 0 1 0 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 0 0 1 1
 1 0 1 1 1 1 1 0 0 1 1 0 1 1 0 0 1 0 1 1 1 1 0 1 1 1 1 1 0 1 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 1 1 1 1 1 1 0 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1
 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 0 0 0 1 1
 1 1 0 1 0 1 0 1 1 1 0 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0
 0 1 0 0 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 0 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1
 1 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 0 1 1 1 1 1 0 1 1
 0 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1
 1 1 1 1 1 1 0 1 0 1 1 0 1 1 1 1 1 1 0 0 1 0 1 0 1 1 1 1 1 0 1 1 0 1 0 1 0 0
 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 0 0 0 0 0 0 1]
<class 'numpy.ndarray'>
(569,)
```

威斯康辛乳癌數據集處理

- 標籤的名稱

```
label_names = bunch.target_names
print(label_names)
print(type(label_names))
print(label_names.shape)
```

```
['malignant' 'benign']
<class 'numpy.ndarray'>
(2,)
```

威斯康辛乳癌數據集處理

數據集的詳細介紹

```
descriptions = bunch.DESCR
print(descriptions)
```

```
.. _breast_cancer_dataset:

Breast cancer wisconsin (diagnostic) dataset
--------------------------------------------

**Data Set Characteristics:**

    :Number of Instances: 569

    :Number of Attributes: 30 numeric, predictive attributes and the class

    :Attribute Information:
        - radius (mean of distances from center to points on the perimeter)
        - texture (standard deviation of gray-scale values)
        - perimeter
        - area
        - smoothness (local variation in radius lengths)
        - compactness (perimeter^2 / area - 1.0)
        - concavity (severity of concave portions of the contour)
        - concave points (number of concave portions of the contour)
        - symmetry
        - 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.

        - class:
                - WDBC-Malignant
                - WDBC-Benign
```

威斯康辛乳癌數據集處理

- 乳癌數據檔案所在的位置

```
filename = bunch.filename
print(filename)
```

```
c:\users\admin\appdata\local\programs\python\python37\lib\site
-packages\sklearn\datasets\data\breast_cancer.csv
```