ML Lab #1:
Breast Cancer Classification

Sunglok Choi, Assistant Professor, Ph.D.
Computer Science and Engineering Department, SeoulTech
sunglok@seoultech.ac.kr | https://mint-lab.github.io/

Overview

- **Prerequisite**
  - Anacodna (Individual Edition)

- **Practice: Breast Cancer Classification**
  - The given data
  - Expected results
  - Practice with the skeleton code
    - Step #1) Visualize all features and its classification results
    - Step #2) Use another classifier

Pinkwashing

- **Assignment**
  - Complete the following three missions
    1. Load the data from the raw file
    2. Try at least two different classifiers
    3. Calculate balanced accuracy

Practice: Breast Cancer Classification

- The given data: [Breast Cancer Wisconsin (Diagnostic) Data Set]
  - Classes (#: **2**): *Malignant* (M; 악성종양 in Korean), *Benign* (B; 양성종양)
  - Attributes: **30** real numbers (except ID and target class)
    - Radius
    - Texture
    - Perimeter
    - Area
    - …
  - The number of data: **569** (M: 212, B: 357)
  - cf. Load the dataset using scikit-learn [API]

```python
from sklearn import datasets
wdbc = datasets.load_breast_cancer()
```

UCI Machine Learning Repository

https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

About Citation Policy Donate a Data Set Contact

Search

Repository Web

UCI

Machine Learning Repository

Center for Machine Learning and Intelligent Systems

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:**	1604079

Source:

Creators:

1. Dr. William H. Wolberg, General Surgery Dept.
University of Wisconsin, Clinical Sciences Center
Madison, WI 53792
wolberg '@' eagle.surgery.wisc.edu

2. W. Nick Street, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
street '@' cs.wisc.edu 608-262-6619

3. Olvi L. Mangasarian, Computer Sciences Dept.
University of Wisconsin, 1210 West Dayton St., Madison, WI 53706
olvi '@' cs.wisc.edu

Donor:

Nick Street

Data Set Information:

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. A few of the images can be found at [Web Link]

Separating plane described above was obtained using Multisurface Method-Tree (MSM-T) [K. P. Bennett, "Decision Tree Construction Via Linear Programming." Proceedings of the 4th Midwest Artificial Intelligence and Cognitive Science Society, pp. 97-101, 1992], a classification method which uses linear programming to construct a decision tree. Relevant features were selected using an exhaustive search in the space of 1-4 features and 1-3 separating planes.

The actual linear program used to obtain the separating plane in the 3-dimensional space is that described in: [K. P. Bennett and O. L. Mangasarian: "Robust Linear Programming Discrimination of Two Linearly Inseparable Sets", Optimization Methods and Software 1, 1992, 23-34].

This database is also available through the UW CS ftp server:
ftp ftp.cs.wisc.edu
cd math-prog/cpo-dataset/machine-learn/WDBC/

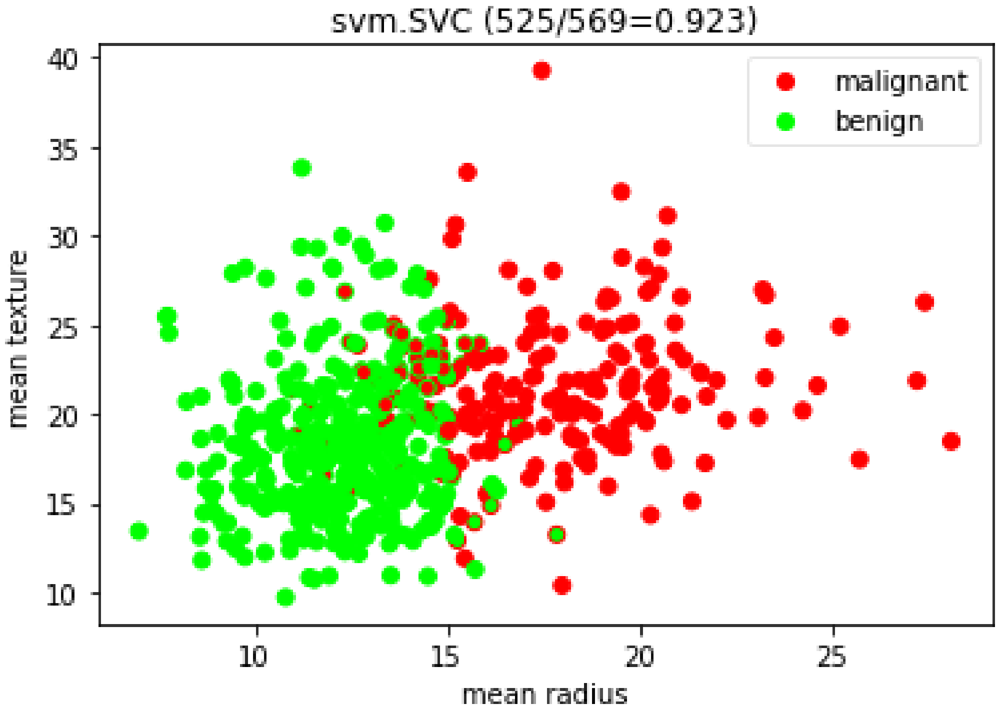
Practice: Breast Cancer Classification

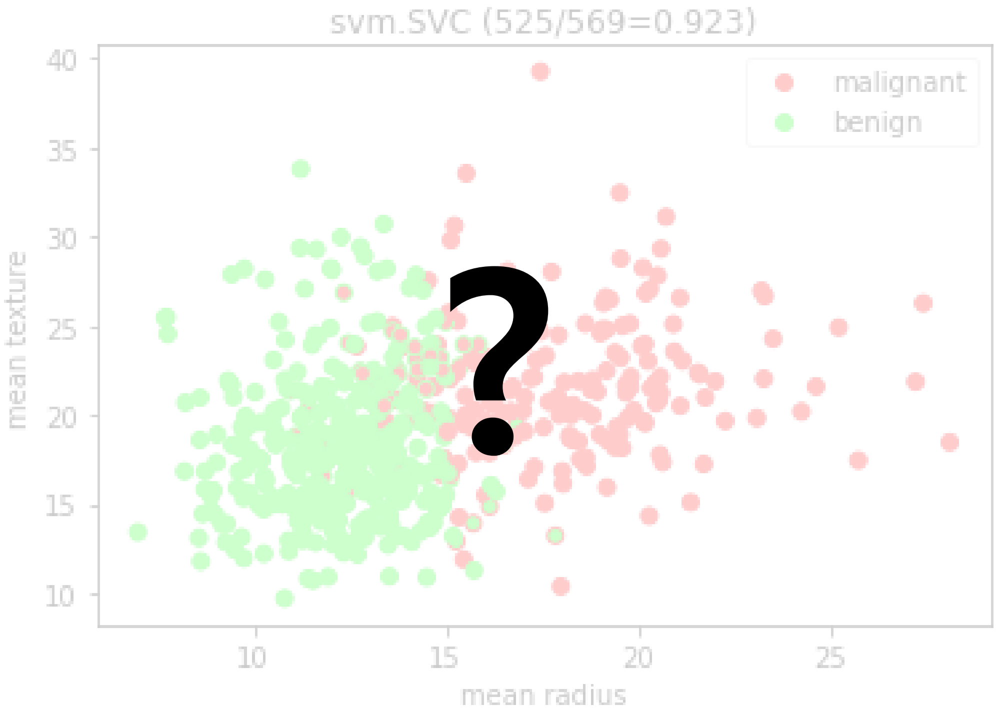
- The given data (file: `data/wdbc.data`)
  - File format: CSV (comma-separated values)
    - `ID,target class (M or F),radius,texture,perimeter,area, ...`
  - Example

```
842302,M,17.99,10.38,122.8,1001,0.1184,0.2776,0.3001,0.1471,0.2419,0.07871,1.095,0.9053,8.589,15
3.4,0.006399,0.04904,0.05373,0.01587,0.03003,0.006193,25.38,17.33,184.6,2019,0.1622,0.6656,0.711
9,0.2654,0.4601,0.1189
...
```

Practice: Breast Cancer Classification

- Expected results
  - Our default classifier: SVM (`svm.SVC`)





Practice: Breast Cancer Classification

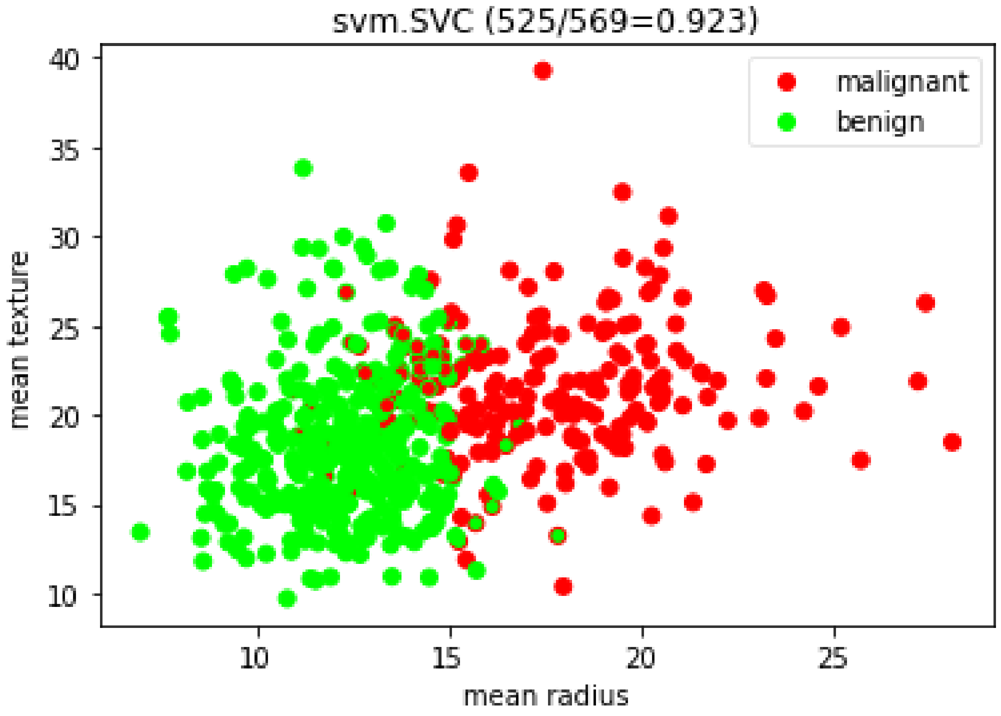
- The given skeleton code (`wdbc_classification_skeleton.py`)

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import (datasets, svm) # Mission #2 and #3) You need to import some modules if necessary
from matplotlib.lines import Line2D # For the custom legend

def load_wdbc_data(filename):
    # TODO

# Load a dataset
wdbc = datasets.load_breast_cancer()
#wdbc = load_wdbc_data('data/wdbc.data') # Mission #1) Implement 'load_wdbc_data()'

# Train a model
model = svm.SVC()                      # Mission #2) Try at least two different classifiers
model.fit(wdbc.data, wdbc.target)

# Test the model
predict = model.predict(wdbc.data)
n_correct = sum(predict == wdbc.target)
accuracy = n_correct / len(wdbc.data)   # Mission #3) Calculate balanced accuracy
```

Practice: Breast Cancer Classification

- The given skeleton code (wdbc_classification_skeleton.py)

```
# Visualize testing results
cmap = np.array([(1, 0, 0), (0, 1, 0)])
clabel = [Line2D([0], [0], marker='o', lw=0, label=wdbc.target_names[i], color=cmap[i]) for i in range(len(cmap))]
for (x, y) in [(0, 1)]: # Not mandatory, but try [(i, i+1) for i in range(0, 30, 2)]
    plt.title(f'svm.SVC ({n_correct}/{len(wdbc.data)}={accuracy:.3f})')
    plt.scatter(wdbc.data[:,x], wdbc.data[:,y], c=cmap[wdbc.target], edgecolors=cmap[predict])
    plt.xlabel(wdbc.feature_names[x])
    plt.ylabel(wdbc.feature_names[y])
    plt.legend(handles=clabel, framealpha=0.5)
    plt.show()
```



- Practice
  - Step #1) Visualize all features and its classification results
  - Step #2) Use another classifier

Assignment

- Mission
  - Complete the following three missions using the given skeleton code (`wdbc_classification_skeleton.py`)
    1. Load the data from the raw file (10 points)
    2. Try at least two different classifiers (5 points)
    3. Calculate balanced accuracy (5 points)
  - Submit your code (`wdbc_classification.py`) and its two result images (`wdbc_classification_???.png`)
- Condition
  - Please follow the above filename convention.
  - You can start from scratch (without using the given skeleton code).
    - However, you should use the given data.
  - You can freely change the given skeleton code if necessary.
- Submission
  - Deadline: **November 24, 2021 23:59** (firm deadline; no extension)
  - Where: e-Class > Assignments
  - Score: Max 20 points