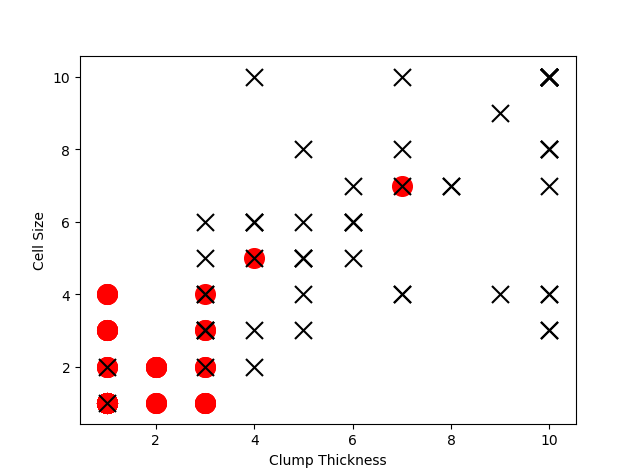
**实验一 良恶性乳腺癌肿瘤预测**

**一、实验目的**

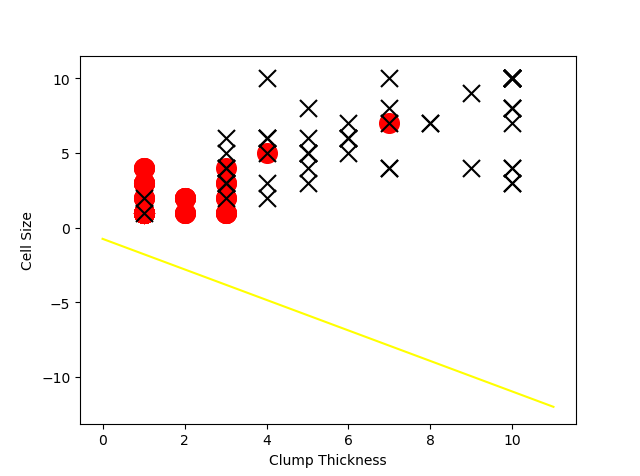
通过实验基本掌握Python编程要素，熟悉numpy、matplotlib、pandas、sklearn的使用，了解机器学习综合实践的方法和步骤。

**二、实验要求**

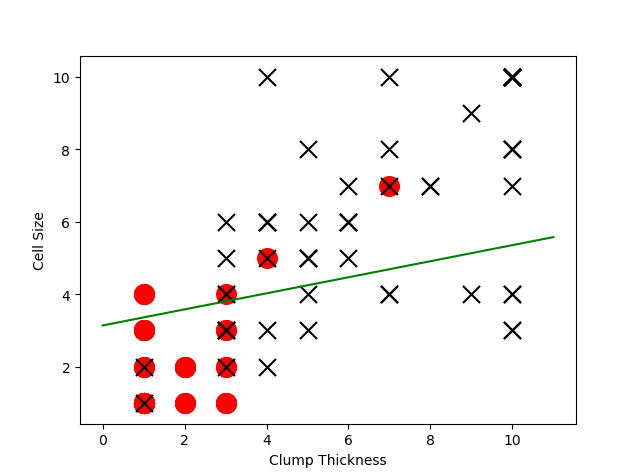
1、从本地读入数据，绘制良/恶性乳腺癌肿瘤测试集数据分布，良性肿瘤样本点标记为红色的Ο，恶性肿瘤样本点标记为黑色的×



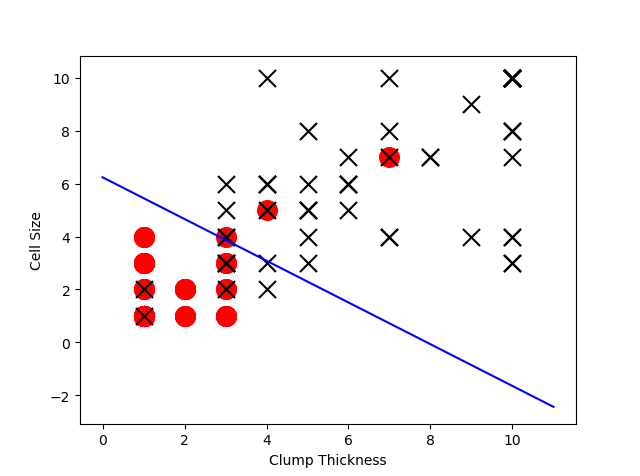
2、绘制随机参数下的二类分类器（用黄色直线表示），利用numpy中的random函数随机采样直线的截距和系数。



3、（选做）绘制利用10条训练样本得到的二类分类器（用绿色直线表示）。



4、（选做）绘制使用所有训练样本得到的二类分类器（用蓝色直线表示）。



**三、实验代码**

# coding: utf-8

# 1、从本地读入数据，绘制良/恶性乳腺癌肿瘤测试集数据分布，良性肿瘤样本点标记为红色的Ο，恶性肿瘤样本点标记为黑色的×

import pandas as pd

test = pd.read\_csv('breast-cancer-test.csv')

print(test.shape)

# print(test.loc[20:23, 'Clump Thickness': 'Cell Size'])

test\_0 = test.loc[test['Type'] == 0][['Clump Thickness', 'Cell Size']]

print(test\_0.shape)

test\_1 = test.loc[test['Type'] == 1][['Clump Thickness', 'Cell Size']]

print(test\_1.shape)

import matplotlib.pyplot as plt

from matplotlib import font\_manager

font\_set = font\_manager.FontProperties(fname='c:/Windows/Fonts/simkai.ttf')

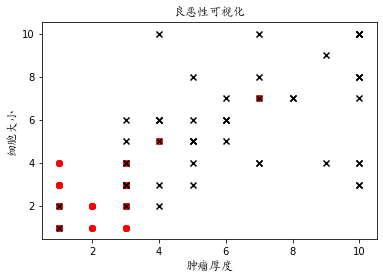
plt.scatter(test\_0['Clump Thickness'], test\_0['Cell Size'], marker='o', c='r')

plt.scatter(test\_1['Clump Thickness'], test\_1['Cell Size'], marker='x', c='k')

plt.title('良恶性可视化', fontproperties=font\_set, fontsize=12)

plt.xlabel('肿瘤厚度', fontproperties=font\_set, fontsize=12)

plt.ylabel('细胞大小', fontproperties=font\_set, fontsize=12)

plt.show()

# 2、绘制随机参数下的二类分类器（用黄色直线表示），利用numpy中的random函数随机采样直线的截距和系数。

import numpy as np

a = np.random.random([2])

print(a)

b = np.random.random([1])

print(b)

x = np.arange(0, 12)

y = (-a[0]\*x-b)/a[1]

plt.plot(x, y, 'y-')

plt.scatter(test\_0['Clump Thickness'], test\_0['Cell Size'], marker='o', c='r')

plt.scatter(test\_1['Clump Thickness'], test\_1['Cell Size'], marker='x', c='k')

plt.title('良恶性可视化', fontproperties=font\_set, fontsize=12)

plt.xlabel('肿瘤厚度', fontproperties=font\_set, fontsize=12)

plt.ylabel('细胞大小', fontproperties=font\_set, fontsize=12)

plt.show()图片包含 屏幕截图

描述已自动生成

train = pd.read\_csv('breast-cancer-train.csv')

print(train.shape)

# 3、（选做）绘制利用10条训练样本得到的二类分类器（用绿色直线表示）。

from sklearn.linear\_model import LogisticRegression

lr = LogisticRegression(solver='liblinear')

lr.fit(train.head(10)[['Clump Thickness', 'Cell Size']], train.head(10)['Type'])

print(lr.coef\_[0, :])

print(lr.intercept\_)

coef = lr.coef\_[0, :]

intercept = lr.intercept\_

x = np.arange(0, 12)

y = (-coef[0] \* x - intercept) / coef[1]

plt.plot(x, y, 'g-')

plt.scatter(test\_0['Clump Thickness'], test\_0['Cell Size'], marker='o', c='r')

plt.scatter(test\_1['Clump Thickness'], test\_1['Cell Size'], marker='x', c='k')

plt.title('良恶性可视化', fontproperties=font\_set, fontsize=12)

plt.xlabel('肿瘤厚度', fontproperties=font\_set, fontsize=12)

plt.ylabel('细胞大小', fontproperties=font\_set, fontsize=12)

plt.show()图片包含 天空

描述已自动生成

# 4、（选做）绘制使用所有训练样本得到的二类分类器（用蓝色直线表示）。

lr = LogisticRegression(solver='liblinear')

lr.fit(train[['Clump Thickness', 'Cell Size']], train['Type'])

print(lr.coef\_[0, :])

print(lr.intercept\_)

coef = lr.coef\_[0, :]

intercept = lr.intercept\_

x = np.arange(0, 12)

y = (-coef[0] \* x - intercept) / coef[1]

plt.plot(x, y, 'b-')

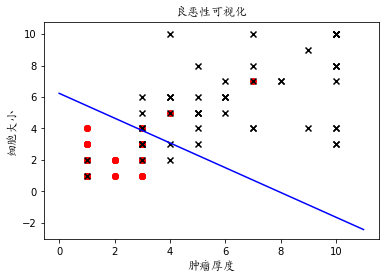
plt.scatter(test\_0['Clump Thickness'], test\_0['Cell Size'], marker='o', c='r')

plt.scatter(test\_1['Clump Thickness'], test\_1['Cell Size'], marker='x', c='k')

plt.title('良恶性可视化', fontproperties=font\_set, fontsize=12)

plt.xlabel('肿瘤厚度', fontproperties=font\_set, fontsize=12)

plt.ylabel('细胞大小', fontproperties=font\_set, fontsize=12)

plt.show()

**四、实验体会**