HW2 Problem 4:

Instructions
Use the provided Jupyter notebook to answer this problem and rename it accordingly. Make sure that it runs without any errors. Try to keep the code easy to read and use comments if you need to clarify what you are doing.

Upload to Blackboard:
- Your Jupyter notebook
- PDF printout (File > Download as > PDF via latex) of the code and output
- If you are having trouble generating a PDF file, you can upload an HTML version (File > Download as > HTML)

In this problem we will be using the breast cancer dataset "Breast Cancer Wisconsin (Diagnostic) Data Set" downloaded from Kaggle to try to predict malignancy based the radius. https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

1. **Logistic regression**
   a. Load the dataset "data.csv" into a Pandas dataframe
   b. Use the 'radius_mean' and 'diagnosis' columns to fit a logistic function to the data. You can use the functions from sklearn for fitting the model and predictions. Refer to chapter 4 of Geron's "*Hands-On Machine Learning with Scikit-Learn and TensorFlow*" to learn more about logistic regression with sklearn. What is the value of the radius at the 50% classification boundary?
   c. Generate a figure with the data, logistic fit, and 50% classification boundary. It should look similar to Figure 1 below (ref: Geron Fig 4.23).

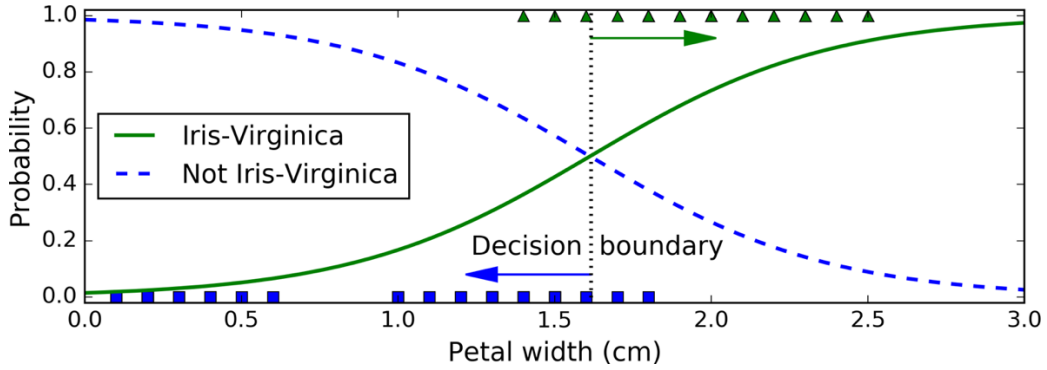

Figure 1: logistic regression with sklearn

2. **Contours of cost function**

We want to visualize the cost function used for the logistic regression above, by plotting the 3D surface and the 2D contours.
   a. Plot the contours of the cost function used in fitting the data from part 1. Identify and label the coefficients of the optimal fit.
   b. Generate a 3D surface plot of the cost function.

3. ROC and area under the curve (AUC)

We want to evaluate the performance of a classifier for malignancy using the mean radius as the only feature. By varying the decision boundary (radius in this case) for the classification, we are going to generate an ROC and find the area under the curve. The AUC is the integral of the ROC.

a. Generate an ROC by sweeping the decision boundary from 5 to 30 in steps of 0.5.
b. What is the AUC for the above ROC?
c. Plot the sensitivity and specificity as a function of the decision boundary.

4. Confusion matrix for optimal fit

a. Generate the confusion matrix for a classifier having the optimal logistic fit coefficients. A typical confusion matrix looks similar to the one below. You can plot or print the matrix in any format you want as long as it contains all the information.

		Prediction	
		Positive	Negative
True class	Positive	TP	FN
	Negative	FP	TN

b. What are the sensitivity and specificity for the optimal fit?