DS420 Assignment #4

The objectives of this assignment is to practice on machine learning algorithms and performance metrics.

1. **Data**
   The dataset can be downloaded at http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data

2. **Description of Data**

   1) Number of instances: 569

   2) Number of attributes: 32 (ID, diagnosis, 30 real-valued input features)

   3) Attribute information

      a) ID number
      b) Diagnosis (M = malignant, B = benign)
         (1) 3-32) Numerical features, based on the following 10 real-valued features.

   4) 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. Fields 3 – 12 are the Mean of these 10 features. Fields 13 – 22 are the SE of these 10 features. Fields 23 – 32 are the worst or largest of these features. For instance, field 3 is Mean Radius, field 13 is Radius SE, field 23 is Worst Radius.

3. **Tasks**

   You are asked to build classification models to classify whether the diagnosis is M (positive), or B (negative).

   1) Build a logistic regression model with LASSO for variable selection. With lambda = 0.01 and lambda = 0.5, what are the features selected separately? Which lambda gives you

simpler model? Which of these two models perform better on the testing set you hold out from the training process? What performance metrics are you using?

2) In the model you trained in Task 1) which has better performance, what are the accuracy, recall, and precision? What probability threshold you are using when you calculate these performance metrics?
3) If you were told that the cost of misclassifying a malignant patient as benign is 100,000 USD, and the cost of misclassifying a benign patient as malignant is 10, 000 USD, and there is no cost of subjects accurately classified by the model, what are the costs when you set the probability threshold as 0.3, 0.5, 0.7? Which probability threshold gives you the lowest cost?