Exercise 3: Multivariate Logistic Regression

You will now try to classify woman breast cancer tumours. This dataset breast\_cancer.csv contains 683 observations and has 9 features and (in column 10) binary labels of either benign (2) or malignant (4). A full description of the dataset is available is this note1 (note that the ID has been removed from the original data).

1 https://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer- wisconsin.names

1. Read data and shuffle the rows in the raw data matrix:

data = ... load breast\_cancer.csv  
np.random.shuffle(data) % Shuffle rows in Python

1. Replace the responses 2 and 4 with 0 and 1 and divide the dataset into a training set and a test set. How many observations did you allocated for testing, and why this number?
2. Normalize the training data and train a linear logistic regression model using gradient descent. Print the hyperparameters α and Niter and plot the cost function J(β) as a function over iterations.
3. What is the training error (number of non-correct classifications in the training data) and the training accuracy (percentage of correct classifications) for your model?
4. What is the number of test error and the test accuracy for your model?
5. Repeated runs will (due to the shuffling) give different results. Are they qualitatively the same? Do they depend on how many observations you put aside for testing? Is the difference between training and testing expected?