**Question 3.** *Evaluating the model on the same data that we use to train the model is considered to be a major mistake in Machine Learning. Implement a hold–out (hint: check out numpy.shuffle()) or cross–validation evaluation strategy. How does your estimate of Accuracy change, compared to testing on the training data? Explain why. (The result might surprise you!)*

This question was answered in the context of the breast cancer data set and using the supervised classifier. An 80-20 holdout strategy was implemented. Using the holdout strategy, the average accuracy over five runs was, to two decimal places, 71.93%. When training and testing against the entire data set, the accuracy was, to two decimal places, 74.83%.

It can be seen that when using the training set as the testing set, the prediction accuracy is significantly higher. However, this result is misleading. When testing using the training set, the classifier is classifying a data set for which it has already been shown the solutions, so a higher accuracy is expected. The accuracy is misleading as it is not classifying anything new, it is simply classifying what it has already seen. Conversely, the lower accuracy for the holdout strategy can be expected as the classifier in this case has not seen the testing data.