Kholby Lawson
CS544
AP3

Data Set:

#	Attribute	Domain
1.	Sample code number	id number
2.	Clump Thickness	1 - 10
3.	Uniformity of Cell Size	1 - 10
4.	Uniformity of Cell Shape	1 - 10
5.	Marginal Adhesion	1 - 10
6.	Single Epithelial Cell Size	1 - 10
7.	Bare Nuclei	1 - 10
8.	Bland Chromatin	1 - 10
9.	Normal Nucleoli	1 - 10
10.	Mitoses	1 - 10
11.	Class:	(2 for benign, 4 for malignant)

There are 683 usable data points in the data set. The sample code number and class are both removed to create the training features. Since all the features are in the same domain, they do not need to be normalized. When calculating the scores, the classes are changed from 2 and 4 to 0 and 1, respectively.

SVM:

My solution uses the default value for C and a gamma value of .00001. The SVM is fit using 100 different train/test splits generated using sklearn.cross_validation.train_test_split. The precision, accuracy, recall, and coefficients returned are those found by the training set that gave the highest accuracy on its corresponding test set.

Precision: 0.981481481481
Accuracy: 0.988304093567
Recall: 0.981481481481

Coefficients:

Clump Thickness	0.26833184
Uniformity of Size	-0.04341114
Uniformity of Shape	0.09300743
Marginal Adhesion	0.20285911
Single Epithelial Size	0.13569462
Bare Nuclei	0.20036573
Bland Chromatin	0.20988885
Normal Nucleoli	0.08698340
Mitoses	0.27323624

Based on these coefficients, Mitoses seems to be the most important feature in predicting malignancy. Clump Thickness, Marginal Adhesion, Epithelial Size, Bare Nuclei, and Bland Chromatin all seem to be “good” predictors, while it seems that Uniformity of Size, Uniformity of Shape, and Normal Nucleoli are not as important in the prediction.