Nikolay Valev

Part A

1) If our model performs great on the training data but generalizes poorly to new instances, that means that the model is overfitted. The two possible solutions are:

- Eliminating Noise, irrelevant features might cause overfitting, so they should be dropped.
- Cross-validate the model with different splits of the original data. The difference between the test-train splits would make the model better.

---

2) 5-Fold Cross-Validation error on 200 training examples

- N1=5
- Train data N2=160
- Test data N3=40

---

3)

A drug company has developed a classifier for detecting whether a vaccine developed is effective or not.

We evaluate the classifier on a test set. Here is the confusion matrix. (In the table, E means effective and NE not Effective.)

		Predicted	
		E	NE
Truth	E	970	25
	NE	10	15

All= 1020
FP = 25
FN = 10
TP = 970
TN = 15

A) Accuracy =(TP + TN) / All = (970+15) / 1020 =0.97

B) For the majority class E,
Accuracy = 970/1020=0.95
(the dataset is highly unbalanced, hence the accuracy is 0.95)

C) The classifier is more useful, because of its higher accuracy(0.97) compared to the accuracy of the majority class baseline(0.95)

D) Precision = (TP/(TP+FP)) =0.97

E) Recall = TP / (TP+FN) = 970 / 980 = 0.99

F) F1 score = TP/(TP+(FN+FP)/2) = 0.98