Homework 1 report:
by He Qu
netid: hequ2

3.1:

Part A:

Build a simple naive Bayes classifier to classify this data set. You should hold out 20% of the data for evaluation, and use the other 80% for training. You should use a normal distribution to model each of the class-conditional distributions.

Output:

```
[1] The scores from 20 classifier trial:
0.7581699 0.7516340 0.7385621 0.7450980 0.7908497 0.7254902 0.7385621 0.7385621 0.7450980
0.7516340 0.7385621 0.8169935 0.7385621 0.6928105 0.7581699 0.7712418 0.6928105 0.7450980
0.7189542 0.7516340

[2] The average score from 20 classifier trial:
0.7454248

[3] The standard deviation of scores from 20 classifier trial:
0.0282498
```

Part B:

Now adjust your code so that, for attribute 3 (Diastolic blood pressure), attribute 4 (Triceps skin fold thickness), attribute 6 (Body mass index), and attribute 8 (Age), it regards a value of 0 as a missing value when estimating the class-conditional distributions, and the posterior.

Output:

```
[1] The scores from 20 classifier trial:
0.7058824 0.7189542 0.7581699 0.7189542 0.7254902 0.7058824 0.7450980 0.7450980 0.6993464
0.7320261 0.7058824 0.7385621 0.7320261 0.7058824 0.7516340 0.7647059 0.7712418 0.7320261
0.7189542 0.7058824


[2] The average score from 20 classifier trial:
0.729085

[3] The standard deviation of scores from 20 classifier trial:
0.02167469

I ran the new classifier several time, but the average scores did not improve at all and even
dropped a little bit. I believe the reason is that if 0 was considered, we would have a more
complete data set to train therefore the test score would improve. And if we just dropped all 0
s in our data set, we would lose a lot of information and the test scores would drop.
```

Part C:

Now use the caret and klaR packages to build a naive bayes classifier for this data, assuming that no attribute has a missing value. The caret package does cross-validation (look at train) and can be used to hold out data.

Output: From one trial:

Accuracy	0.7647
95% CI	(0.6894, 0.8294)
No Information Rate	0.6536
P-Value [Acc > NIR]	0.001988
Kappa	0.4938
Mcnemar's Test P-Value	0.404657
Sensitivity	0.7900
Specificity	0.7170
Pos Pred Value	0.8404
Neg Pred Value	0.6441
Prevalence	0.6536
Detection Rate	0.5163
Detection Prevalence	0.6144
Balanced Accuracy	0.7535

```
[1] The scores from 5 classifier trial:
0.7712  0.7642  0.7323  0.7647  0.7974

[2] The average score from 5 classifier trial:
0.76596

[3] The standard deviation of scores from 5 classifier trial:
0.02075385265
```

Part D:

Now install SVMLight, which you can find at http://svmlight.joachims. org, via the interface in klaR (look for svmlight in the manual) to train and evaluate an SVM to classify this data.

Output:

```
[1] The scores from 20 classifier trial:
0.7581699 0.7450980 0.7516340 0.7712418 0.7647059 0.7385621 0.7908497 0.7385621 0.7516340
0.6993464 0.7712418 0.7777778 0.7058824 0.7908497 0.7712418 0.7385621 0.7124183 0.7908497
0.7320261 0.7647059 0.7124183 0.7908497


[2] The average score from 20 classifier trial:
0.753268

[3] The standard deviation of scores from 20 classifier trial:
0.02722853
```

3.3:

Part A:
Take the disease attribute, and quantize this into two classes, num = 0 and num > 0. Build and evaluate a naive bayes classifier that predicts the class from all other attributes Estimate accuracy by cross-validation. You should use at least 10 folds, excluding 15% of the data at random to serve as test data, and average the accuracy over those folds. Report the mean and standard deviation of the accuracy over the folds.

Output:

```
[1] The scores from 20 classifier trial:
0.8000000 0.8222222 0.7333333 0.8444444 0.8666667 0.8000000 0.8222222 0.8222222 0.7333333
0.8222222 0.8000000 0.8888889 0.8000000 0.8222222 0.9111111 0.7777778 0.8222222 0.7111111
0.8444444 0.8444444


[2] The average score from 20 classifier trial:
0.8144444

[3] The standard deviation of scores from 20 classifier trial:
0.04957257
```

Part B:
Now revise your classifier to predict each of the possible values of the disease attribute (0-4 as I recall). Estimate accuracy by cross-validation. You should use at least 10 folds, excluding 15% of the data at random to serve as test data, and average the accuracy over those folds. Report the mean and standard deviation of the accuracy over the folds.

Output from one trial:

Accuracy	0.6379
95% CI	(0.5012, 0.7601)
No Information Rate	0.5517
P-Value [Acc > NIR]	0.1169
Kappa	0.3741
Mcnemar's Test P-Value	NA

	Class: 0	Class: 1	Class: 2	Class: 3	Class: 4
Sensitivity	1.0000	0.0000	0.28571	0.42857	0.0000
Specificity	0.6923	0.91667	0.92157	0.90196	1.00000
Pos Pred Value	0.8000	0.00000	0.33333	0.37500	NaN
Neg Pred Value	1.0000	0.81481	0.90385	0.92000	0.96552
Prevalence	0.5517	0.17241	0.12069	0.12069	0.03448
Detection Rate	0.5517	0.00000	0.03448	0.05172	0.00000

Detection Prevalence	0.6897	0.06897	0.10345	0.13793	0.00000
Balanced Accuracy	0.8462	0.45833	0.60364	0.66527	0.50000

Output from 20 trials:

```
[1] The scores from 20 classifier trial:
0.6034483 0.5689655 0.6724138 0.5862069 0.6034483 0.5689655 0.6379310 0.6379310 0.6034483
0.6034483 0.5862069 0.4827586 0.6551724 0.5862069 0.5862069 0.5689655 0.6206897 0.6379310
0.5689655 0.5689655


[2] The average score from 20 classifier trial:
0.5974138

[3] The standard deviation of scores from 20 classifier trial:
0.04121071
```