Assignment 4: Notes

Points that are perhaps interesting, but that were not an expected part of the answer, have been enclosed in square brackets.

Question 1:
Test set error is better when test data are available such that the difference between the training and test data more or less accurately reflects the difference between the source and target population.

rpart's relative error should be ignored because, if the model is over-fitted, it will be optimistic. It inevitably continues to decrease as the number of splits increases.

Question 2b:
The one SE rule is used because it leads to a model that is less affected by statistical sampling variation. A loss (hopefully, small) in expected accuracy is traded for an increase in model stability.

[Observe that, around the minimum, the change in CP with a one SE change in predictive accuracy, is a maximum. (The curve is flat at that point.) Moving back slightly from the minimum gives a tangent with a modest angle to the horizontal, and the the change in CP with a one SE change in predictive accuracy is much reduced. Actually, this sort of argument suggests that the SE may not be the right quantity; instead, one should be looking at the slope in the plot of cross-validated error against CP. This issue requires further research!]

The SE for the average error over 3 runs might be taken as `mean(SE)/sqrt(3)`. It might also be reasonable to take, eg, the minimum of the three SEs. [What is more relevant; the SE for the minimum on a curve from a single run, or the SE for the minimum on the average of three runs? As indicated above, maybe neither of these.] More sophisticated and theoretically more satisfactory estimates of the SE of the average are of course possible, but the tools for finding them have not been covered in this course.

[As the 3 estimates of error are not independent, the above estimate for the SE of the mean of the three errors is a fudge. However the exact amount by which the model is under-fitted is not too crucial; the point is to move back from the flat part of the curve.]

Question 2c:
As an example, take the confusion matrix, after replacing numbers by proportions in each row, to be:

	Predicted Type	
Acrual Type	No	Yes
No	.886	.114
Yes	.470	.530

Then if a is the proportion in the target population that are "Yes", the expected error rate is $(1\text{-}a)\times 0.114 + a \times 0.470$

[Note that this uses a model that was optimized for population proportions of 66% No and 34% Yes. If we refitted the model (maybe by specifying suitable weights) to give optimal predictive accuracy for each different value of a, we'd see a different curve.]

Question 3:
The comparison with rpart may be unfair because the svm model has not been tuned. [There's not much scope to tune rpart models.]

Question 4:
In 4a and 4b, for "cross-validation", read "OOB".

[It is interesting that in 4c, repetition of the comparison a large number of times, and plotting the OOB accuracy against the test set accuracy, gives a graph with a small negative correlation. On average the two accuracies agree; randomForest does not, on average, overfit. Whenever the OOB accuracy is a bit high, this indicates slight overfitting, and the test set accuracy is degraded. Most people who did this correctly found, even with just ten repeats, a small and statistically real negative correlation.]

Question 5b:
This question proved a too subtle for a number of students. There was however quite a lot of mileage to be gained by mechanically following instructions. You were to take the ordinates from `cmdscale` in 5a and use those as constructed variables for the discriminant analysis in 5b. (That part was a mechanical following of the strict letter of the question). Why do this? If the OOB error rate more or less agrees with the OOB error rate in 5a, then the two-dimensional representation is somewhat accurately reflecting the clarity of separation of classes that the 5a randomForest result indicates.

[A further step, which arguably stretches the use of these ideas beyond reasonable limits, is to use the proximities from the 5b analysis to repeat the use of `cmdscale` and plot another graph. The configuration of the points changes dramatically, indicating that the plot may have limits as a representation of the configuration of the points in space.]

Extraction of the OOB error rate from the model object:
If the model object is `diabetes.rf`, then you can calculate the OOB error rate from the final element in the first column of the object `diabetes.rf,$err.rate`. Assuming 500 trees, this is `diabetes.rf$err.rate[500,1]`