Multivariate prediction modelling with applications in precision medicine (HT2019) - examination

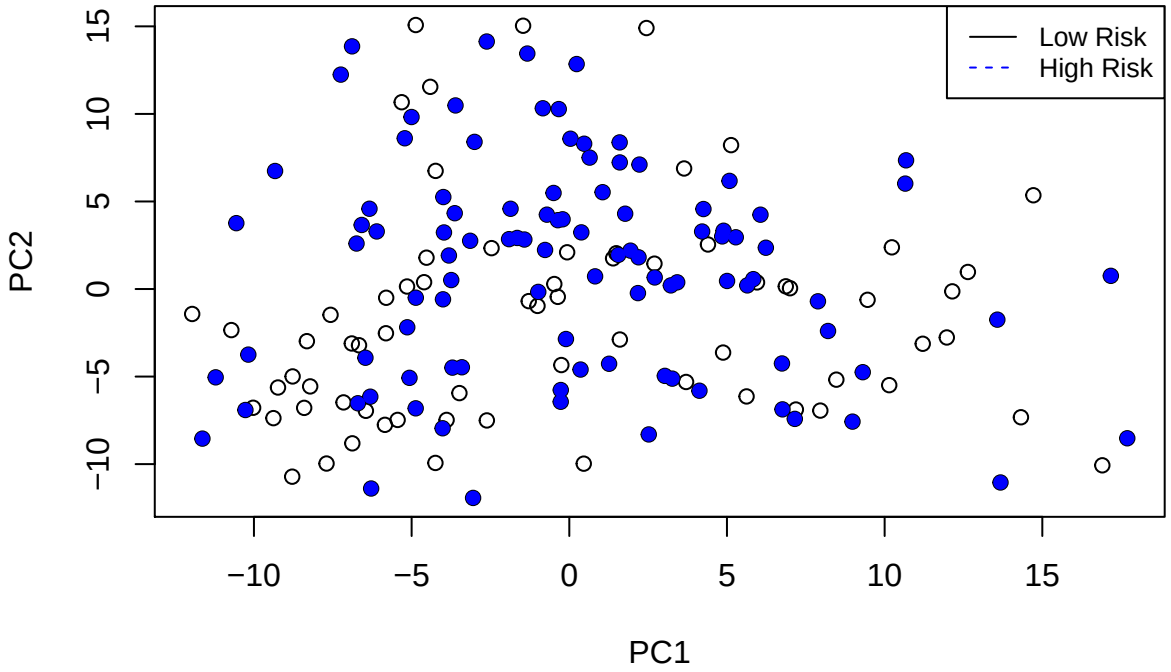
Taner Arslan (19880301-5778)

11/29/2019

1 Part 1 of exam: Data analysis task

1.2 Analysis tasks

1. Apply an unsupervised method of your choice for analysis of the predictor data matrix (xTraining). Interpret your results. Does the unsupervised method appear to capture class-related information based on visual inspection?



I have performed PCA analysis on training data. Patients were colored based on their risk group. Patients looked randomly distributed across 2D PCA plot based on their risk group annotation and they were not able to be separated by PCA analysis, although first two principal components explained around 30% variance of the training data.

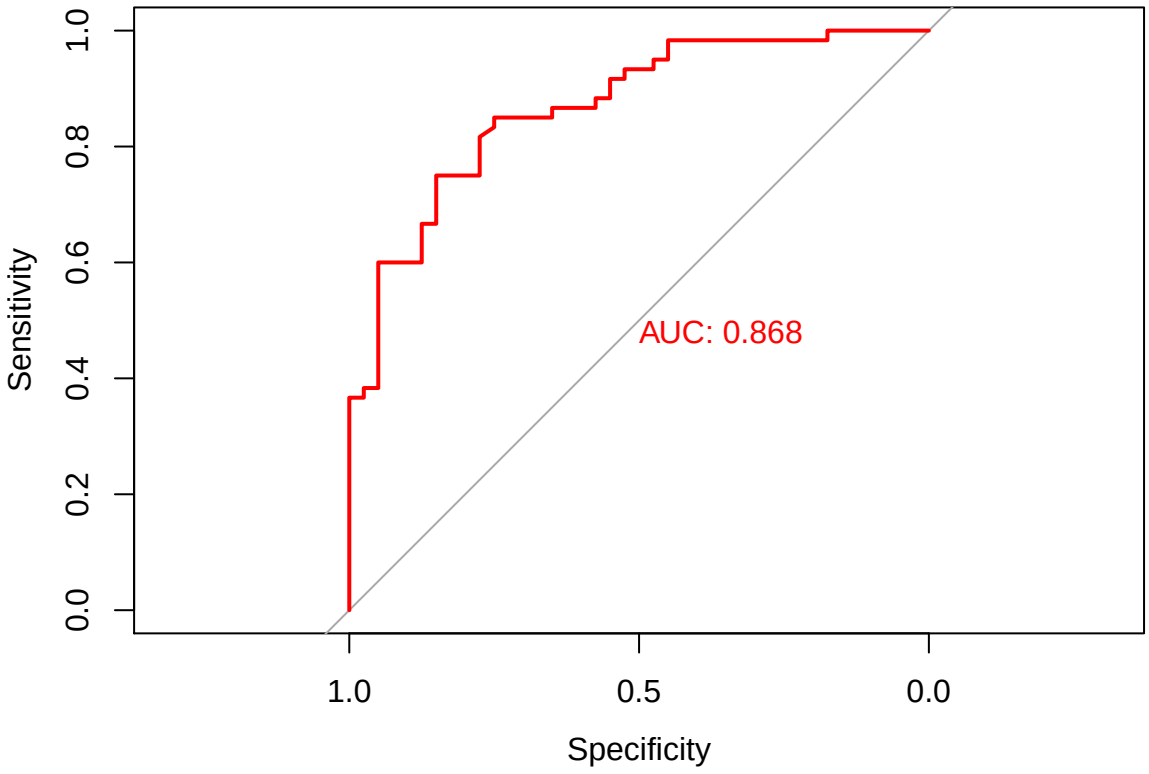
2. Train a supervised classification model of your choice using the train- ing data (xTraining and yTraining) and report classification performance estimates (of your choice) from cross-validation and from the test set (xTest,yTest).

I used SVM model with radial basis function kernel in order to capture the non-linearity in the data. SVM focuses closest points from different classes while the remaining points are already easy to classify. It tries to find a separation plane (or hyperplane) whose distance is maximized in order to be more generalizable model. To do that first, I have done 10-fold cross-validation for hyper-paramater optimization. The best model was achieved with using gamma = 0.01 and cost = 10.

Next, the SVM model trained with 10-fold cross-validation with the optimized hyper-parameters. The accuray for the training data was **100%**, whereas the accuracy dropped to **84%** in the independent test data. Despite the decreasing in the test accuray, the fact that the model performed well in test data.

3. Train a supervised classification model (of your choice) that output a continuous prediction. Generate a ROC curve based on the test dataset (xTest,Ytest) to visualize the trade-off between the sensitivity and specificity. Provide an interpretation of the results. You should also indicate what the specificity is at sensitivity = 0.9 (based on the ROC curve).

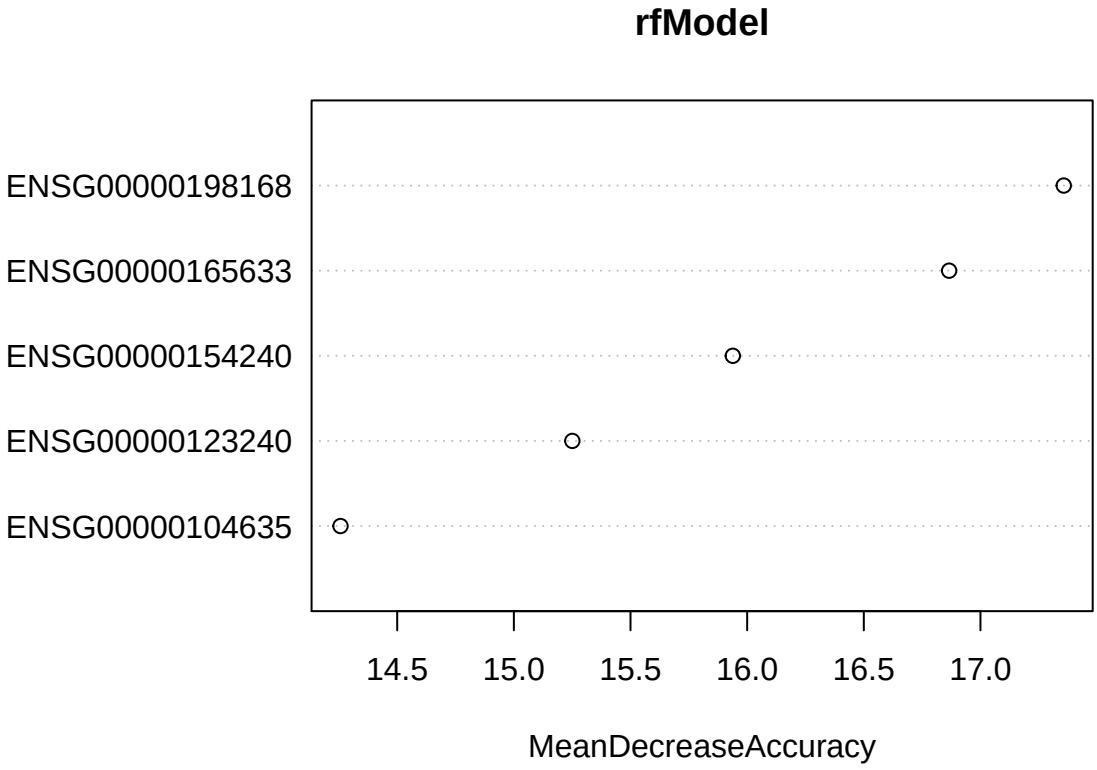
Random forest model was trained on training data in order to generate probability output instead of hard prediction. Random forest is an ensemler learner method. It builds many decision trees. Each tree outputs a class label and most voted class from random forest becomes the model prediction.



Given the AUC which is **0.866** at test data, the random forest model performs nicely. The specificity is **0.55** when sensitivity is **0.9**. When the model reaches 90% sensitivity (recall), its specificity (precision) is marginal outperforms random guessing. The model can be improved for more efficient specificity

4. Apply a classification model (or variable selection method) to determine which variables were the most important (or which were selected).Describe the approach you used and the principle that it is based upon. Report how many variables were selected and/or the 5 top-ranked variables.

I, again, applied random forest model in order to determine the most important variables. Random forest ranks all variables based on how much information they contribute. Very basicaly, they determine by using *Shannon information* and *Gini impurity*. Those algorithms select the variable whose information is the most in terms of separation of patients.



2 Part 2 of exam: Exam questions

Question 1 Describe what bias-variance trade-off is in the context of super- vised learning and its interpretation.

Variance refers to how much the model changes, when we estimate the model using a different training data set. Ideally the model should not vary much for better generalizability. However, if a model has high variance with small changes in the training data most likely results in large changes in model. In general, more flexible models have higher variance. Bias is simply, the error. Observation – prediction.
As a general rule, as we use more flexible methods, the variance will increase and the bias will decrease for both training and testing data. As we increase the flexibility of the model even more, training error continues to decline, however test error significantly starts increasing. So the model should have a balance between bias and variance.

Question 2 Cross-validation is often applied to optimize tuning parameters in supervised prediction models. Describe a situation when you should use nested cross-validation and why.

The main reason of appliying cross-validation is to prevent over-fitting. Since, ideally, we should use test data only once, we have to be very sure about the model that we build. It should be validated before testing phase.

In medicine or biology researches, sometimes (maybe most of the times), due to natural reasons, enough data is not genereted for machine learning or AI applications. Therefore, the information from each ivdiviuals become very important and we do not have luxury to put aside some data for only testing purpose. For thoses cases, cross-validation is a preferable choice, for example, outer loop splits the data for validation and training while inner loop optimizes possible hyperparamters or selects features etc.

Question 3 Briefy describe problems that can occur when transferring the use of a model from one population (where it was developed) to another.

The model might be optimized well for one population, but it may fail when we transfer the model to another population due to difference of the distribution of the ethnicity or genetic background. If that is the case, we likely encounter severe issues with sensitivity and specificity.

Question 4 Penalization of a loss function (e.g. residual sum of squares) is a technique that is often used in supervised learning methods. What is the primary purpose of the penalization?

The primary use of penalization is avoiding over-fitting. Then models tends to overfit if there are large numbers of variables due to curse of dimensioanlty. When we apply penalization to our model we force some variables become zero or very close to zero (depends on L1 or L2 regularization) which results lower numbers of variable in the model.
Penalization can be used as a feature selection as well. We drop-out some variables since their weights become zero.