MACHINE LEARNING FOR BIOMEDICAL AND HEALTHCARE APPLICATIONS
Ngee Ann Polytechnic

Assignment 03: Principal Component Analysis and Supervised Classification

Student ID	S10228323
Student Name	Julia Loh Jie Min

INSTRUCTIONS

Dataset: "Breast Cancer.csv"

In this assignment, we shall make use of the "Breast Cancer.csv" data set. The data set contains features of cell nuclei extracted from digitized images of breast tissues. Besides the "id" column which represents the record ID, and "diagnosis" column which indicates the diagnosis outcome ("M" – Malignant, "B" – Benign), the other columns are numeric and represent features extracted or calculated from the cell nuclei.

We shall first be implementing a number of machine learning algorithms (including logistic regression) to predict cancerous cells using the features available.

Following that, we shall perform principal component analysis on the dataset and then build a logistic regression model using a limited number of principal components.

Implement R codes to accomplish the following:

PART 1: Import and Analyse Data

(1) Import the data set to a data frame, **d**, and note the following:
   a) the number of columns and rows
   b) the ratio of malignant cases to benign cases

(2) Do the following:
   a) Remove the column "id" from the data set.
   b) Convert the data type of the column "diagnosis" to factor.
   c) Check the data set for missing data and remove records that contain missing data (if any).

PART 2: Build Machine Learning Models

(3) Partition the dataset such that **70%** is used for training and **30%** is used for testing. Assign the training samples to the data frame **trgSamples**, and the test samples to the data frame **tstSamples**.

(4) Using **seven-fold** cross-validation, train the following classification models:

   a) **model1** – Logistic Regression (method = "glm", family = "binomial")
   b) **model2** – Support Vector Machine (method = "svmRadial)
   c) **model3** – Random Forest (method = "rf")
   d) **model4** – Neural Net (method = "nnet")

(5) Write the following the function that you will use to do the prediction and return the model performance. You will need to complete the sections as indicated. (Note: Use malignant as the positive cases)

```
doPrediction <- function(m, nameOfModel, tst) {
    # Arguments
    # m - model
    # nameOfModel - character string describing the model
    # tst - test samples to run the prediction on
    # -------------------------------------------------

    # predict
    predicted.prob <- predict(m, newdata = tst, type = "prob")
    # convert  probabilities to outcomes using threshold of 0.5
    predicted.outcomes <- factor(…) # to be completed
    # generate confusion matrix
    cm <- confusionMatrix(…) # to be completed
    # compare performance indicators, put into data frame for ease of display
    modelPerformance <- data.frame(
        model = nameOfModel, # name of model
        Accuracy = …, # to be completed
        Sensitivity = …, # to be completed
        Specificity = …, # to be completed
        Precision = …, # to be completed
        F1 = …) # to be completed
    row.names(modelPerformance) <- NULL
    return(modelPerformance)
}
```

PART 3: Perform Prediction (Classification) and Compare Performance

(6) Using the above function, perform prediction using each of the models that you have built e.g.

```
mp1 <- doPrediction(model1, "Logistic Regression", tstSamples)
```

where **mp1** is the model performance result for **model1** etc.

(7) For a confusion matrix,
   a) What do the following terms mean? True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN)?
   b) Write down the expression for accuracy, recall (sensitivity), precision and F1, in terms of TP, TN, FP and FN.
   c) By analyzing the expressions for the above, discuss why in practice, accuracy is not always the best performance indicator.

(8) Compare the performances of the various models. For <u>each</u> of the following cases, explain and state which model you will recommend to a hospital:
   a) The hospital places higher priority on detecting malignant cases over false alarms
   b) The hospital would like to have a balance between recall and precision

PART 4: Perform Principal Component Analysis

(9) Perform Principal Component Analysis on the numeric variables of data frame, **d**. Assign the result to the object variable **p**.

(10) What is the minimum number of principal components required to account for at least 97% of the variance in the data set?

(11) Create a new data frame, named **d.pca**, containing only the column "diagnosis" from the original data set, **d** and principal components **PC1 to PCx,** where x is the minimum number of principal components you have found in (10)**.**

Hint: You can do the above using the following statement:

```
d.pca <- data.frame(p$x) %>% select(PC1:PCx) %>% data.frame(diagnosis=d$diagnosis,.)
```

PART 5: Partition d.pca, build and apply logistic regression model

(12) Using the same partitioning index that was generated in (3), partition **d.pca** such that 70% is used for training and 30% is used for testing. Assign the training samples to the data frame **pca.trgSamples**, and the test samples to the data frame **pca.tstSamples**.

(13) In (12), why is it necessary to use the same partitioning index and training:test ratio as in (3)?

(14) Using **seven-fold** cross-validation, build a Logistic Regression model, **model5**.

(15) Perform prediction using **model5** i.e.

```
mp5 <- doPrediction(model5, "Logistic Regression (PCA)", pca.tstSamples)
```

(16) Compare the performance of **model5** with that of **model1**. In your opinion, is the performance of **model5** superior, inferior or about the same as that of **model1**?

(17) Describe the benefits of principal component analysis in Machine Learning.

ANSWERS TO QUESTIONS
In addition to submitting your R script (.R), please provide answers to the following questions asked earlier.
Please note that your answers have to be consistent with the outcomes that you produce using your code.

---

(1) (a) Rows: 569, Columns: 12, (b) Ratio of Malignant to Benign cases is 37:63

(7) (a)

	YES (Predicted)	NO (Predicted)
YES (Actual)	**TP** (True Positive)	**FN**(Fast Negative)
NO (Observed)	**FP** (False Positive)	**TN** (True Negative)

True Positive (TP): When predicted result is 'Yes' and observed result is 'Yes'
True Negative (TN): When predicted result is 'No' and observed result is 'No'
False Positive (FP): When predicted result is 'Yes' but observed result is 'No'
False Negative (FN): When predicted result is 'No' but observed result is 'Yes'

(b)
Accuracy= (TP+TN)/ (TP+FN+FP+TN)
Recall (Sensitivity)= TP/(TP+FN)
Precision= TP/(TP+FP)
Specificity=TN/ (TN+FP)
F1= 2* Recall* Precision/Recall+Precision
=2 [TP/(TP+FN)] [TP/(TP+FP)] / [TP/(TP+FN)] + [TP/(TP+FP)]

(c) Based on the equation shown in (b), a high accuracy would mean a high true positive or high true negative, but ignores possibility of false positives and false negatives that may come about whilst trying to achieve high accuracy in machine learning. For example, a model that has can predict accurately no recurrence of breast cancer can result in high false negatives -women with incorrectly thinking their breast cancer was not going to reoccur. Furthermore, in an imbalanced data set where if 95 out of 100 cases are positive, and the model is

used to predict positive cases, the accuracy will be 95% but if this model were to be applied to other data, the model will fail as it is only suited for one type of data.

Compare the performances of the various models. For each of the following cases, explain and state which model you will recommend to a hospital:

a) The hospital places higher priority on detecting malignant cases over false alarms
b) The hospital would like to have a balance between recall and precision

8a)The hospital places higher priority on detecting malignant cases over false alarms- high true positive, low false positives meaning high precision.

	model	def.Accuracy	def.Sensitivity	def.Specificity	def.Precision	def.F1
1	Logistic Regression	0.9411765	0.9365079	0.9439252	0.9076923	0.9218750
2	SVM (Radial)	0.9529412	0.9365079	0.9626168	0.9365079	0.9365079
3	Random Forest	0.9294118	0.9206349	0.9345794	0.8923077	0.9062500
4	nnet	0.9294118	0.9365079	0.9252336	0.8805970	0.9076923

Based on the results, SVM(Radial) will be the recommended model as it has the highest precision at 93.65%

8b)The hospital would like to have a balance between recall and precision-
When a good balance between recall and precision is desired, F1, which is
F1=2∗ Recall∗ Precision/Recall+Precision is used as performance indicator. Therefore, SVM(Radial) will be the recommended model as it has the highest precision at 93.65%

10) 5 prinicpal components. Cumulative proportion value of PC5 is 97.5%.

13) The idea is to compare the logistic regression results between model1 without PCA and model5 with PCA(dimensionality reduction using the same data set. Thus to make a fair comparison it is necessary to use the same partitioning index and training:test ratio as in (3).

16) The performance of **model5** is the same as that of **model1.** Accuracy, specificity, precision, sensitivity and F1 results are the same

17) Benefits include
-Dimensionality reduction- removes correlated features (reduce feature space) and thereby improves model performance and reduces overfitting
-Improves visualization of data through dimensionality reduction ie. use for multispectral sensing- able to differentiate images much better
- saves computing time by removing correlated features
-Reduction of noise since the maximum variation basis is chosen and so the small variations in the background are ignored automatically.

Please submit your answers (in PDF) and your R script (.R) separately on PolyMall

---