Applying Clustering Algorithms

Yuxuan Zhang

Westcliff University

AIT 508

Instructor: Professor Aliah Algassmi

Date: Feb 1, 2026

Classification Model

The code and plots for this assignment are submitted separately in a Jupyter notebook. It can also be viewed in the assignment github repo https://github.com/yuki172/ait508-assignments/tree/main/week%201

In this assignment, I build an end-to-end classification model using scikit-learn for the Breast Cancer dataset available through scikit-learn, which is a dataset commonly used for benchmarking for classification (Pedregosa et al., 2011). In this report, I will discuss how each step of the assignment is performed and why, including feature scaling, dimensionality reduction, model training and evaluation. Finally, I will also discuss the results and interpret them from an application point of view.

Feature scaling is applied to the original data. Specifically, I used Z score scaling to bring the data features into comparable scales (Han et al., 2012). This is a good choice for some of the models that were experimented with like logistic regression because they are sensitive to feature scales. In general, a model can be sensitive to feature scales if they rely on quantities such as Euclidean distance and the dot product, or if they use a gradient based optimization method. This is because for these types of models, the gradient information for features on a larger scale will dominate the optimization process.

The original dataset has 30 features and it is evident from computing the pairwise correlations and the VIF factor that some are correlated. I use PCA to reduce the feature set to a smaller set of uncorrelated components (Hotelling, 1933). The criterion used to determine how many features to keep is the percentage of the variance preserved. More precisely, I take the first k directions where k is the first number at which the ratio of the sum of variances of the first k

directions over the sum of variances of all original features is greater than or equal to 95%. The number of directions selected is 10.

The train test split is 90 to 10. I trained a logistic regression model and evaluated its performance on the test set. Logistic regression is chosen for simplicity and interpretability. For such a simple task, logistic regression can have good performance and enable one to interpret the results in terms of which features have more impact and the type of impact. It is crucial that we obtain the coefficients of the original features for interpretation, which can be done using the coefficients of the features from PCA.

The model has good performance as shown below

Accuracy: 0.9298

Precision: 0.9706

Recall: 0.9167

F1-Score: 0.9429

We see from the relatively low recall that accuracy alone does not give a good indication of model performance. This is true for this application since recall is more important than precision. Indeed, false negatives are much more harmful than false positives. For logistic regression, this means that we should lower the probability threshold for predicting the positive class. With a lower threshold value, the model demonstrates good prediction performance for this application and can be used as a support tool for medical professionals.

References

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology, 24*(6), 417–441. https://doi.org/10.1037/h0071325

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research, 12*, 2825–2830.