Unsupervised Learning Reveals Structure in Breast Cancer Tumour Data

Gillian Macdonald, Lenka Okasova, and Bob Rice
School of Physics and Astronomy, University of Nottingham, UK

Unsupervised learning methods offer a systematic approach to uncover latent structure in high-dimensional biomedical data without reliance on diagnostic labels. We analyse the Breast Cancer Wisconsin dataset using principal component analysis (PCA) followed by k-means clustering to probe whether malignant and benign tumour classes emerge naturally from feature statistics alone. Correlation analysis reveals a strongly coupled cluster of morphological and nuclear features associated with malignancy, motivating dimensionality reduction. PCA shows that the first two components capture 74.2% of the total variance (80.2% with three components), with clear separation between tumour classes in the reduced space. Applying k-means clustering in this low-dimensional representation yields robust unsupervised class recovery, achieving and adjusted Rand index of 0.84 relative to known tumour classes. These results demonstrate that correlated cytological abnormalities organise the data into a low-dimensional structure that cleanly separates malignant and benign tumours, highlighting the utility of simple unsupervised approaches for structure recovery in data.