Tutorial 9: Machine Learning
Due: Tuesday, April 7, 2020

Breast Cancer Dataset

Use the breast cancer dataset (*datasets.load_breast_cancer()*) in scikit learn to develop a machine learning model to diagnose breast cancer, and to answer the questions.

Features:

1. radius (mean of distances from center to points on the perimeter)
2. texture (standard deviation of gray-scale values)
3. perimeter
4. area
5. smoothness (local variation in radius lengths)
6. compactness (perimeter^2 / area - 1.0)
7. concavity (severity of concave portions of the contour)
8. concave points (number of concave portions of the contour)
9. symmetry
10. fractal dimension ("coastline approximation" - 1)

Targets:

- Malignant (0)
- Benign (1)

Questions

1. Is this a supervised or unsupervised machine learning problem? Why?

2. Is this a classification or regression machine learning problem? Why?

3. Train and test 3 different machine learning algorithms for this dataset. Report the accuracy of the prediction for each model.

4. Redo question 3 using principal component analysis (PCA) on the features prior to training the model. Remember to scale the features before using PCA.

5. Select your favourite model, and develop a confusion matrix for it. Comment on the effectiveness of your model. Do you think that it would be better to have too many false positives or too many false negatives for a cancer diagnosis?