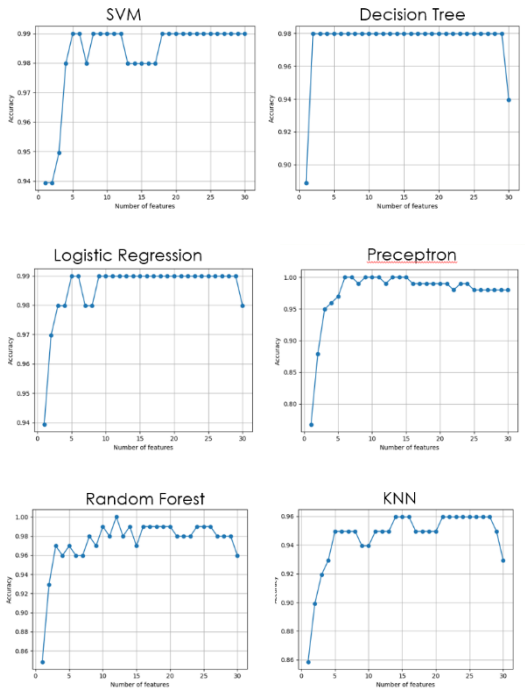
Breast Cancer Classification Using Machine Learning

Ⅰ.DATA PREPROCESSING

我所使用的是 UCI (Diagnostic) Breast Cancer Wisconsin 的數據集，數數據集的資訊說明這是有遺漏值的數據，卻在檢查數據發現並沒有遛值的產生，經一番驗證發現所有原來的遺漏值數據都被填上了 0，所以我利用一些工具把 0 變為遺漏值，再把遺漏值補上整個特徵的中位數。接著可把整個乳癌數據分成 Training set(70%) 和 Testing set(30%)，分別進行建模和測試。

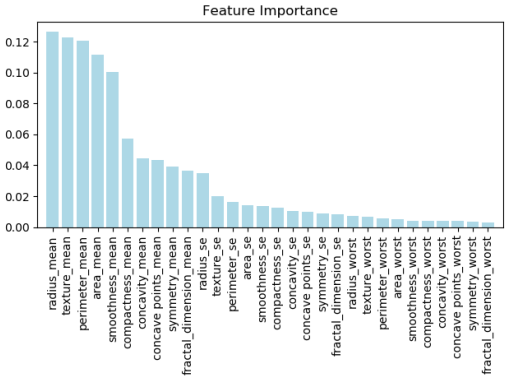
Ⅱ. Feature Selection

利用 Sequential Backward selection 來對不同模型做特徵選擇，看用多少的特徵就可以達到很好的準確度，以下:



相較之下，特別的是 Decision Tree，只要兩個特徵就能讓準確值達到最大，不經令我好奇到底是哪兩個特徵如此強大，發現是: perimeter_worst 和 smoothness _worst 導致。

接著，我使用 Random Forest 來計算出每一個特徵的重要性佔了所有數據的比例，如下:

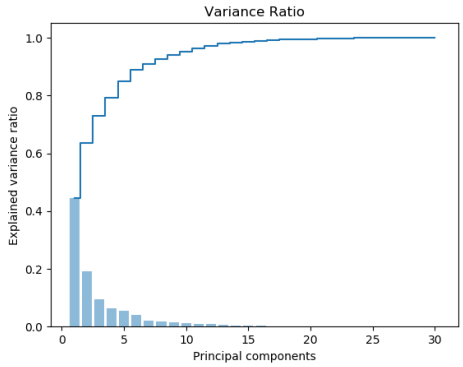


發現這五個佔了極大的重要性:

1. Radius_mean
2. Texture_mean
3. Perimeter_mean
4. Area_mean
5. Smoothness_mean

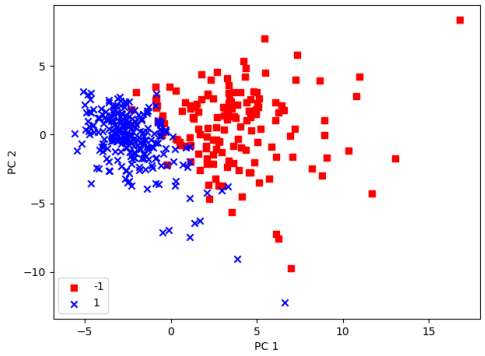
Ⅲ. Feature Extraction

此部分是利用降維來壓縮數據，用的方法是 Principal Component Analysis(PCA)，主要的原理是把數據投影到低維度使的變異數最大，此為最佳化問題，需解出 Covariance matrix 的特徵值和特徵向量，此特徵向量等同於投影軸，可以解使變異數所佔的比例，如下圖:



由此得知，前兩個主成分就幾乎可以解釋了所有數據的 60%的變異數了。

接著是投影到 PC 座標上的散佈圖:



最後用 Logistic Regression 對 Training set 和 Testing set 測試: