Greg Simpson SupportVectorMachine-CancerClassification

GALVANIZE DATA SCIENCE FELLOW

IBM DATA SCIENTIST (SOON)

Motivation and Background

- Long career in software
  - Many Big name companies; several small companies
  - Lots of different roles
  - Plenty of success
    - GPS-OCX ground station contract win
  - Too many “less than successes”
- Looking for something new to keep me interested

After Galvanize

TRUVEN
HEALTH ANALYTICS™
an IBM Company

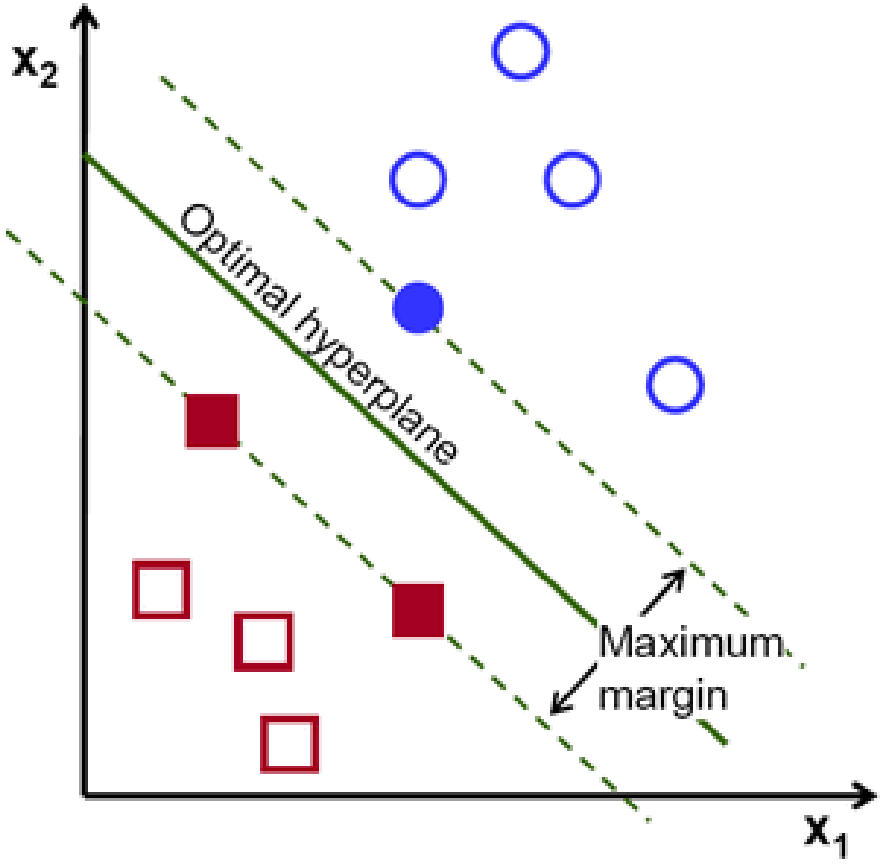
3 GOALS for Capstone

- Combine
  - HealthCare Domain
  - Data Science
  - IBM Watson Tools

HealthCare Domain

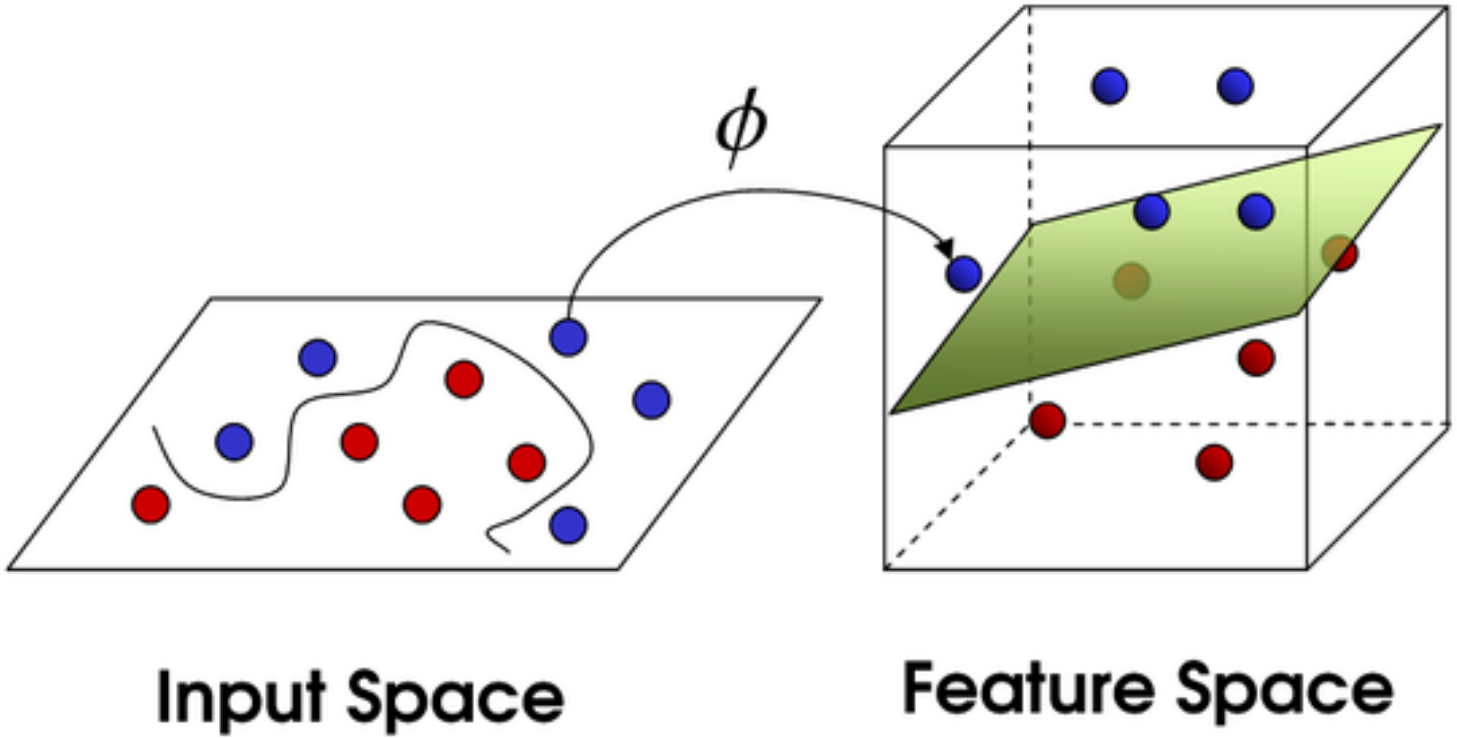
- Breast Cancer Study from the University of Wisconsin
  - Publicly available data
  - 700 records
  - 10 features – cell characteristic measurements
  - 1 actual result
    - 2 – benign
    - 4 – malignant
- Goal is to compare the results of your classifier against the actual result

Support Vector Machine (SVM)

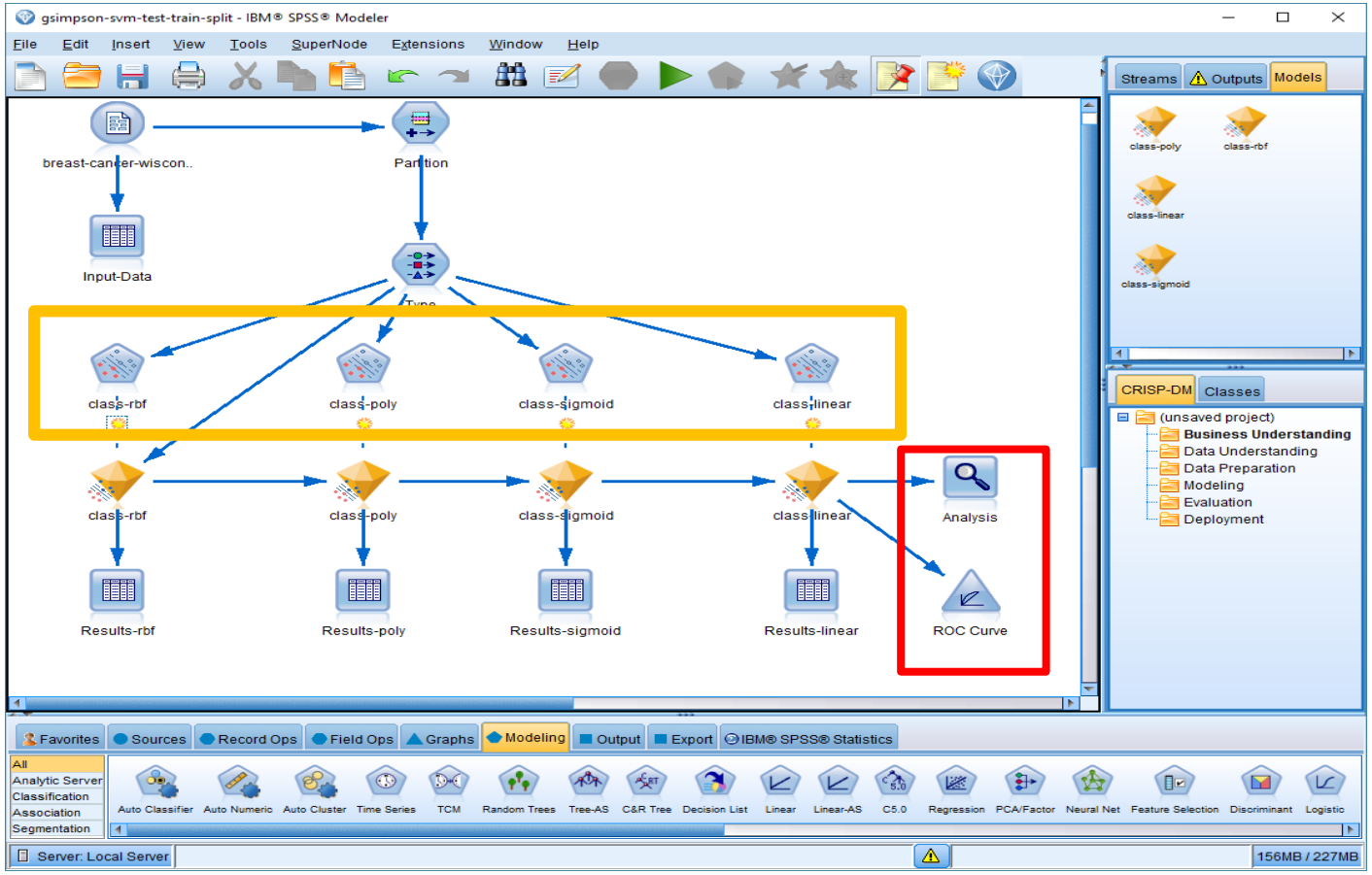


- Hyperplane
  - best separates two classes of points with the maximum margin.
    - https://www.quora.com/What-does-support-vector-machine-SVM-mean-in-laymans-terms

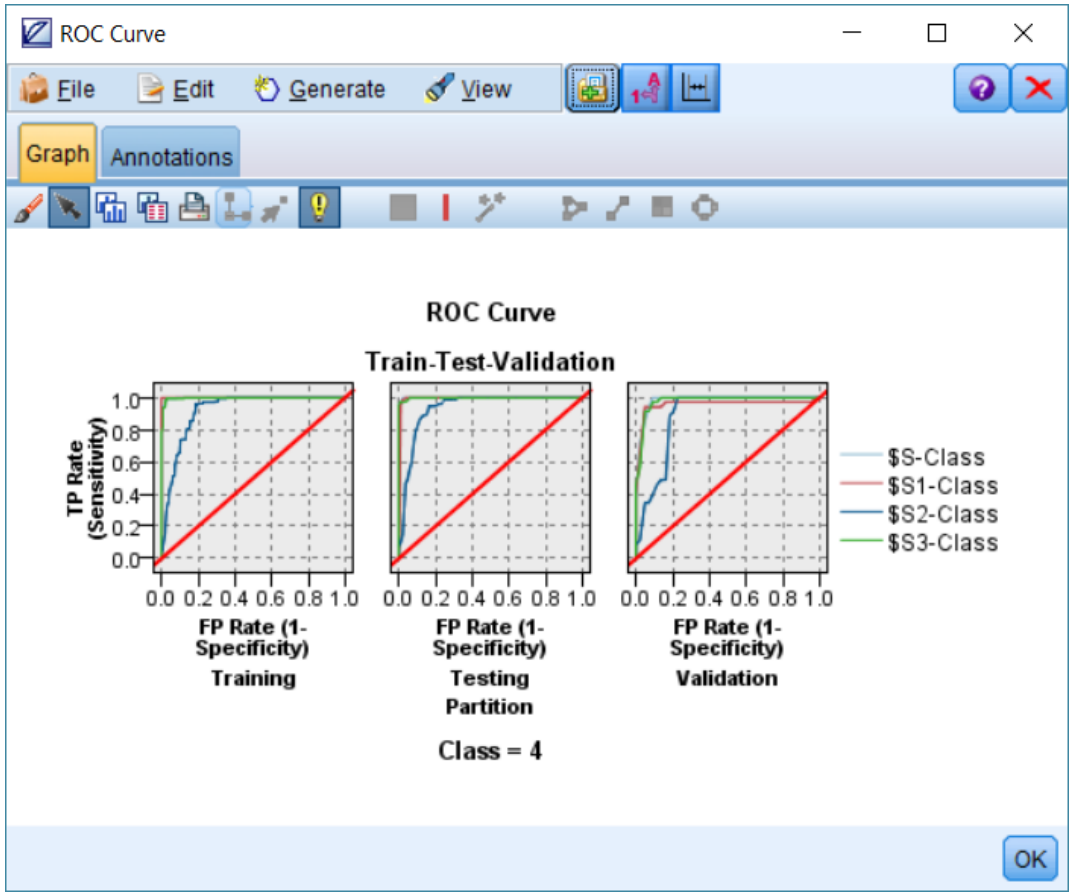
Kernels Project Data to Increase Separation



SPSS Stream Screen Shots



SPSS Stream Screen Shots



Model Results

- Train (correct / incorrect)
  - Polynomial : 100% / 0%
- Test (correct / incorrect)
  - Rbf (Radial basis function) : 96.89% / 3.11%
- Validation (correct / incorrect)
  - Linear : 92.41% / 7.59%

Which one to choose

- Test (correct / incorrect)
  - Rbf (Radial basis function) : 96.89% / 3.11%

- The prediction accuracy obtained from the unknown set more precisely reflects the performance on classifying an independent data set. An improved version of this procedure is known as cross-validation.
  - http://www.csie.ntu.edu.tw/%7Ecjlin/papers/guide/guide.pdf

Questions and Contact

- Greg Simpson
  - gsimpson@pobox.com
  - 303.907.2233
  - linkedin.com/in/gregoryjsimpson
  - github.com/GregSimpson
    - SupportVectorMachine-CancerClassification