Basics on Support Vector Machines

By Samuel S. Shepard

Classification

- Given data, like “height” and “weight” can you *classify* data into male or female?

- Height and weight are used as *input* datasets.

- Male and female are two *classes* or output data.

- The model or rules that help you make a decision are called a *classifier*.

Machine-learning

- Machine-learning tools find trends in training datasets and create an internal model to classify new data.
- Datasets may include tabular data like measured values, prediction scores, or statistics.
- Test datasets are used to *validate* the accuracy of the machine-learning classification process.

Support Vector Machines

- There exist *many* machine learning tools: neural networks, decision trees, Bayesian networks, SVM, *etc.*

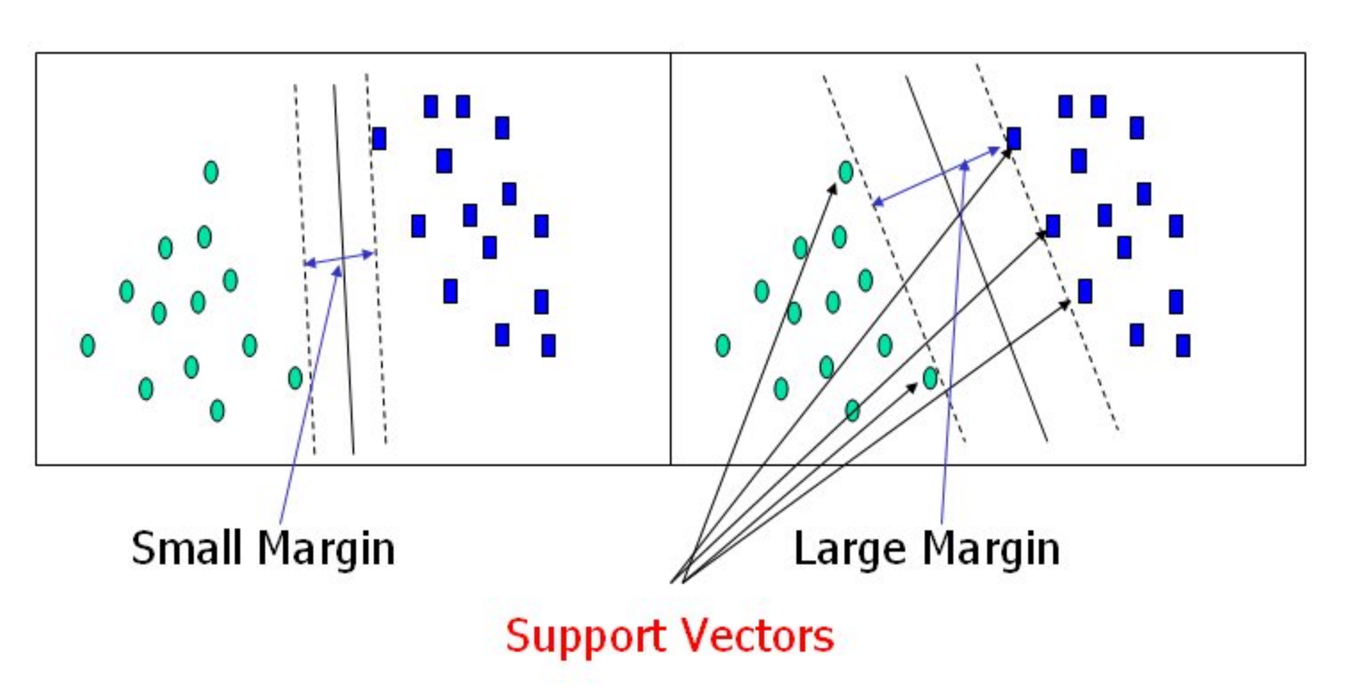
- The *Support Vector Machine* is one such algorithm (SVM).

Support Vector Machines

- Looks at data *spatially*.
- Finds a *vector* or line that divides the two classes of data the best.
- Can do so linearly or nonlinearly.
- May use fancy mathematical functions (kernels) to compute the boundary.

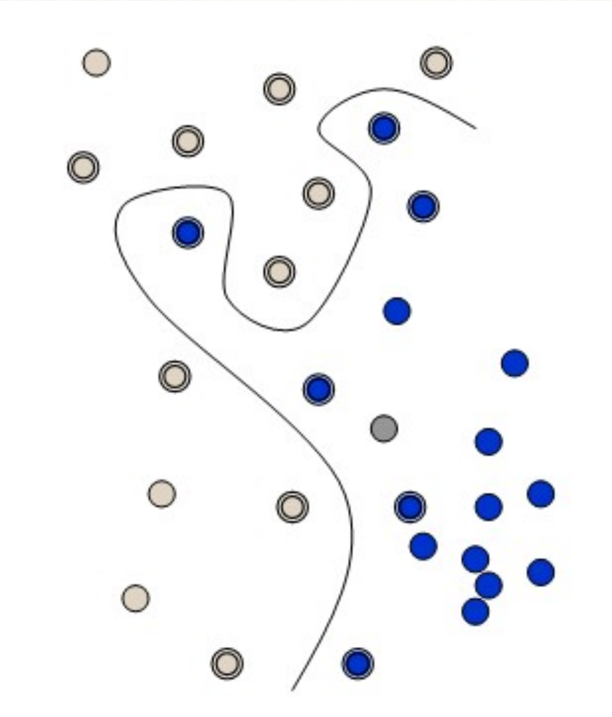
Support Vector Machines

Linear data separation (2D).



From *www.dtreg.com/svm.htm*

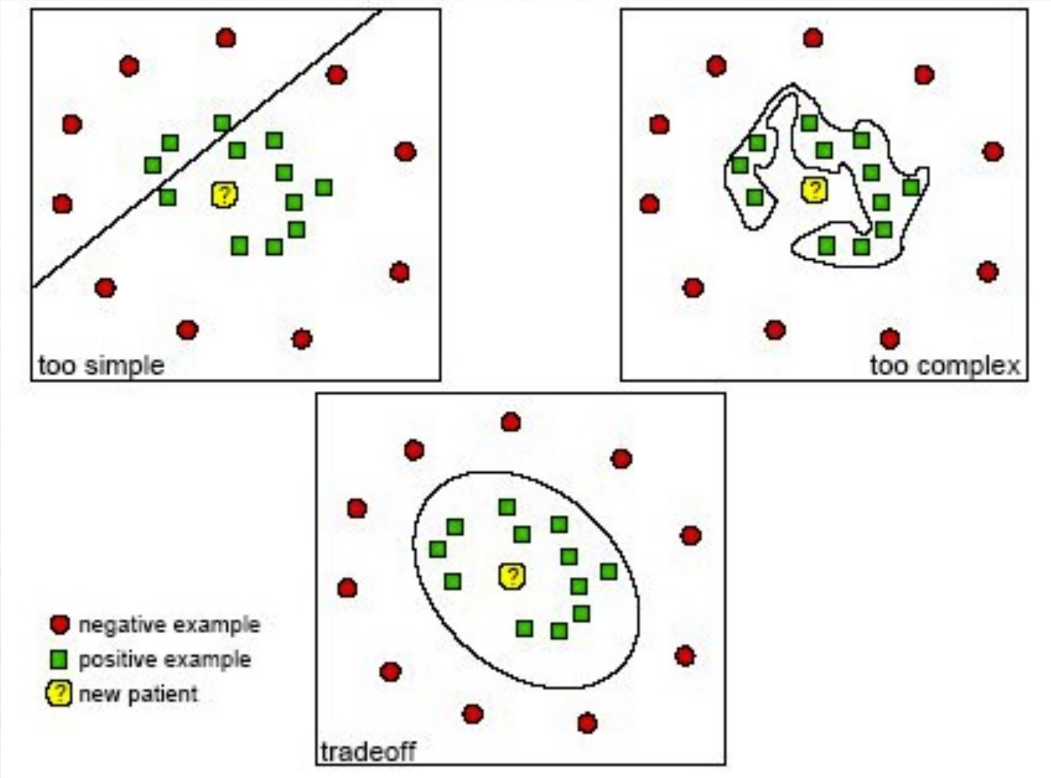
Support Vector Machines



Non-linear division boundary.

From *www.dtreg.com/svm.htm*

Over & under fitting.



- Drawing a good boundary ensures new data will be correctly classified.

From *www.dtreg.com/svm.htm*

SVM Technology

- In order to find the nonlinear decision boundary, the data is transformed into a higher dimensional space.
- Kernel functions are equivalent to the dot product in these higher dimensional spaces.
  - Ex: Gaussian, sigmoid, polynomial, *etc.*

Our Data

- For each data point there are multiple attributes or values (like height and weight for one person).
- We will examine multiple prediction scores (from different algorithms) applied to the same exons and introns.

Our Classes

- We do binary classification (2 classes).

- Exon class = +1
- Intron class = -1

- Prediction score > 0 => Exon
- Prediction score < 0 => Intron

Our Attributes

- Each exon or intron data point will have **10** different prediction scores (attributes) associated with them.
- Using these scores we will train and test the SVM to do a *global* prediction of the data points.

Your Task

- Familiarize yourself with the installation process for Octave-Shogun.

- Understand how to use Octave-Shogun to do prediction using the Gaussian, Sigmoid and Polynomial kernel functions.

Let’s Get Started

- *http://bpg.utoledo.edu/svmlab/*