SUPPORT VECTOR MACHINES WITH JRUBY

Matt Kirk
SocialVolt

SUPPORT VECTOR MACHINES SAFARI

Matt Kirk
SocialVolt

PROBLEM

There is so much data out in the world.

What the hell should people do with it?

Ruby doesn’t have many machine learning or data mining packages out there.

Writing C-extensions to pull in existing open source work is a pain!

SOLUTION

Given a data set we can utilize a supervised learning model.

While many supervised learning models exist support vector machines are great for classification and are well supported in packages.

Since ruby's svm gems are somewhat outdated the best option is to tie into the java package that is under active development.

WHY JRUBY?!!

Ruby's power along with Java's tons of packages.

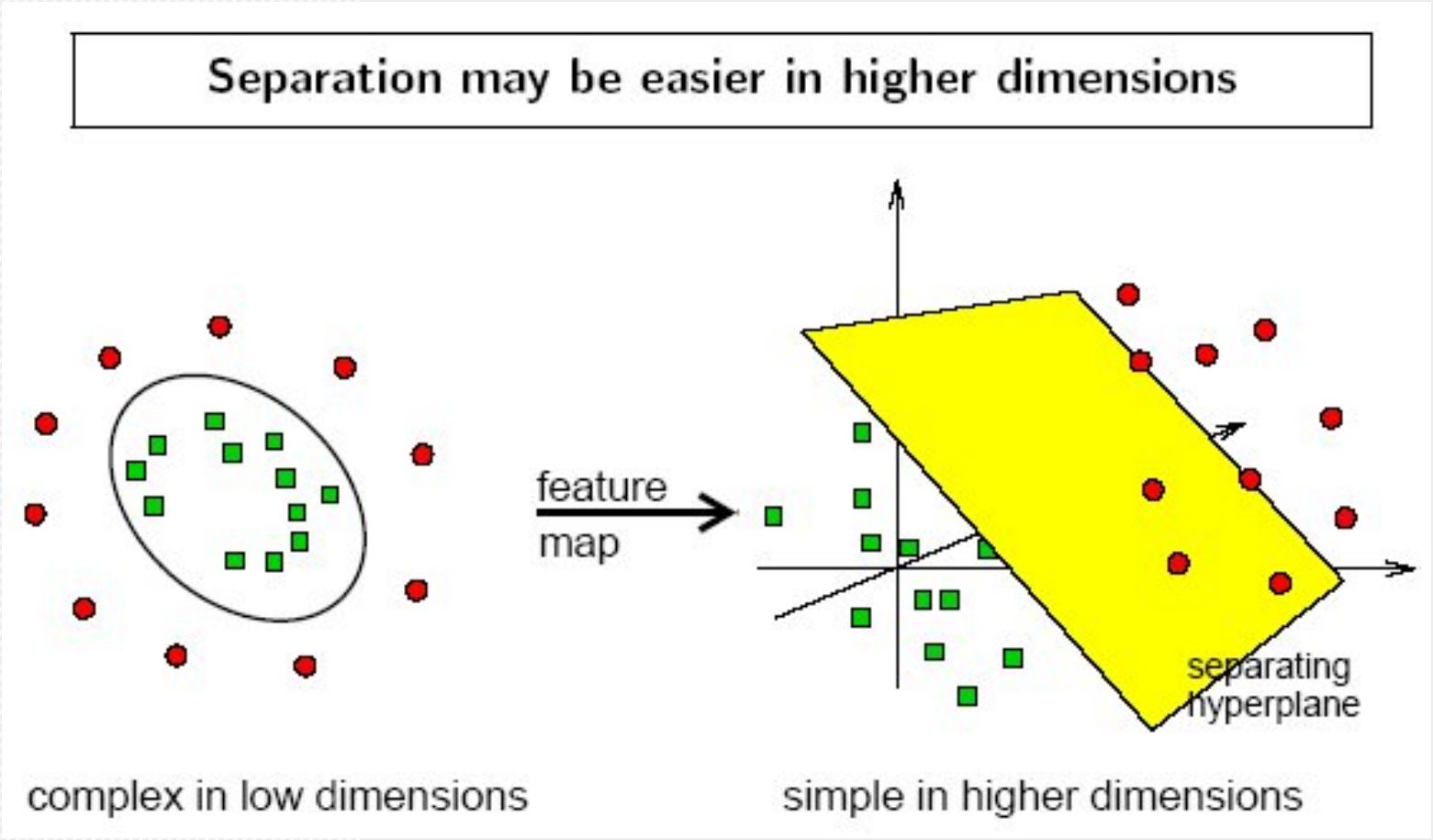
Don't think I like Java :-P

MODEL

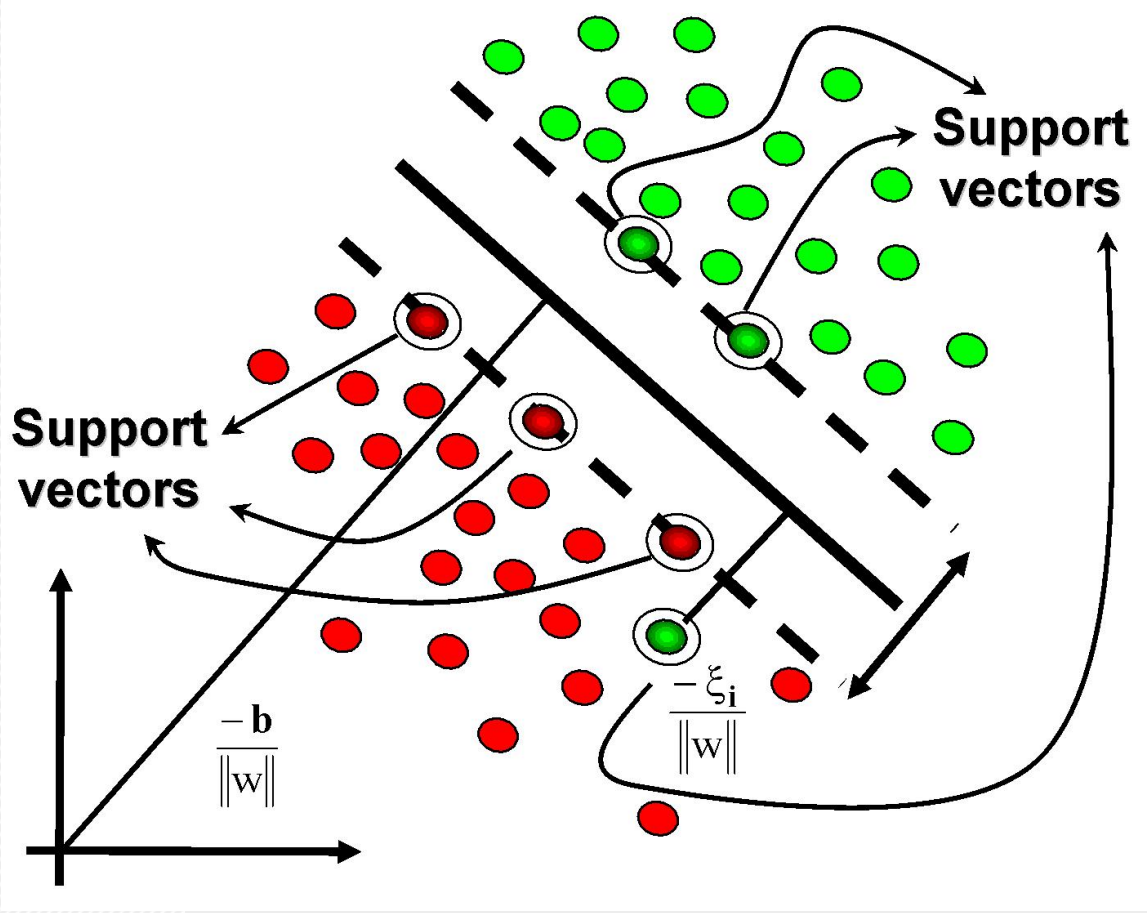
Three steps to support vector machines:

1. Collect data
2. Map data onto high dimensional planes
3. Use hyper planes to determine classification of new data
4. Rinse and repeat

LINEAR: U'*V

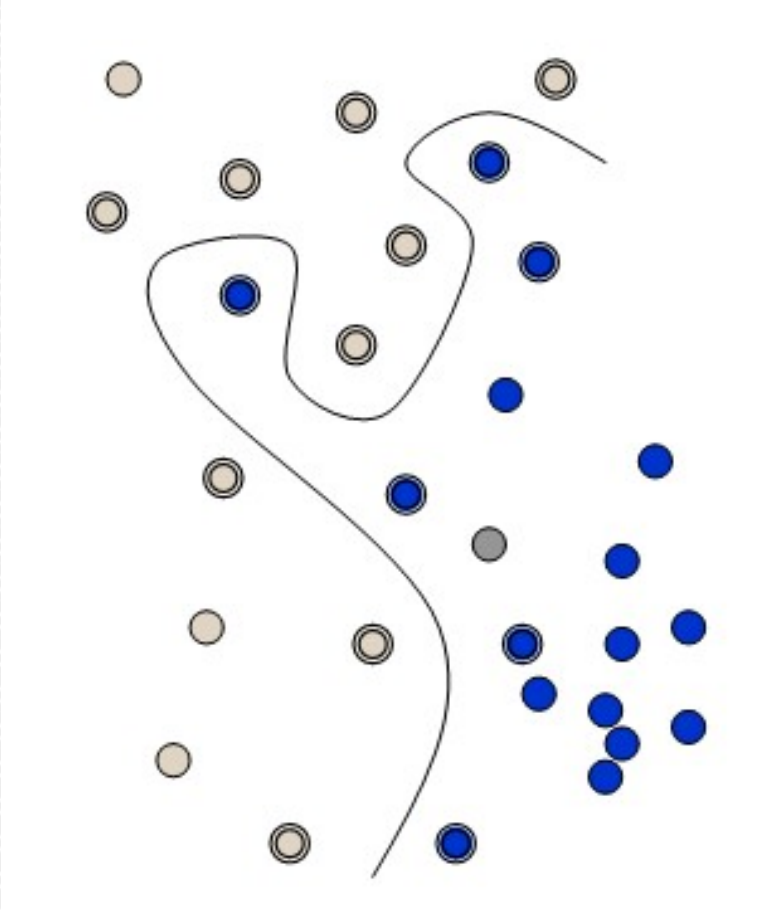

Separation may be easier in higher dimensions
feature
map
separating
hyperplane
complex in low dimensions
simple in higher dimensions

SUPPORT MARGINS



“KERNEL TRICK”

$$e^{-\gamma|u-v|^2}$$



UNDERLYING TECHNOLOGY

LibSVM is the post popular svm library out there written originally in C

But also simultaneously ported to Java.

To utilize it's bindings it's easier to load it all up in a Jar and mess around with it in IRB

BUT THERE IS A RUBY VERSION!!!

Unfortunately the ruby version isn't very good and runs exceptionally slow.

One you can't really find easily, and the other one using Swig has some interesting implementation problems such as overcomplicating svm_node creation.

STOP! EXAMPLE TIME

SOME DATA

Ghost Fever (Campy, 1987): 5 Stars

Attack of the Killer Tomatoes (“Man eating monster”, 1978): 3 Stars

The Blob (Teenagers, 1988): 2 stars

Turns into:

		1	2	3	4	5
Subtitle	Stars	Campy	MEM	Teens	Horror	Year
GF	5	1	nil	nil	nil	1987
AKT	3	nil	1	nil	1	1978
TB	2	nil	nil	1	nil	1988

SOURCE CODE ALL LOCATED ON GITHUB

COMPETITION

For classifying generic attributes support vector machines shine but there are many other supervised and unsupervised learning algoritms

- Linear regression
- Logistic regression
- Naïve Bayesian ranks
- K-nearest neighbor algorithm
- Etc etc

PLEASE ☺

GIVE ME FEEDBACK

Twitter: @mjkirk

E-Mail: matt@matthewkirk.com

Github: http://github.com/hexgnu/rmw-svm

SpeakerRate: http://spkr8.com/t/8842