



Department of Computer Science

UNIVERSITY OF COLORADO **BOULDER**



Deep Learning

Advanced Machine Learning for NLP

Jordan Boyd-Graber

BIG PICTURE

Administrivia

- Sign up (readings posted sooner)
- What Alvin is looking for
- Examples of five-minute outlines this week

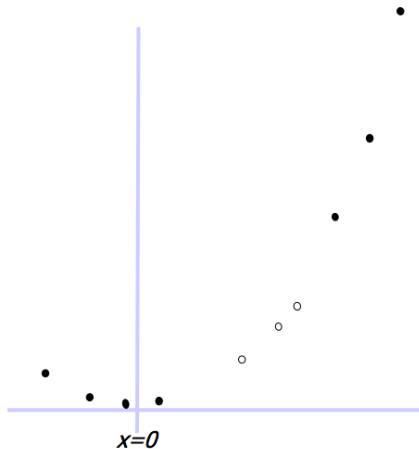
Roadmap

- Less languagey
- Intuitions about non-linear decision functions
- Deep learning
- How this connects to language next week

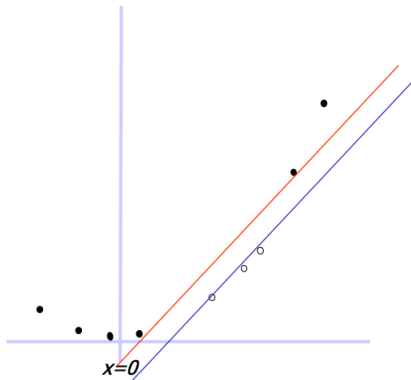
Can you solve this with linear separator?



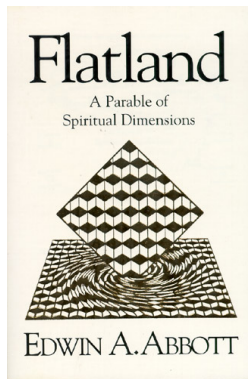
Can you solve this with linear separator?



Can you solve this with linear separator?



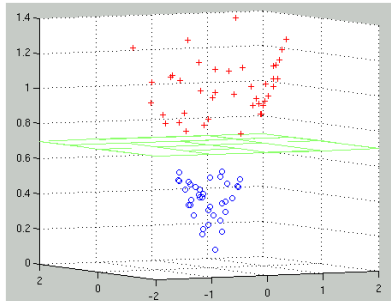
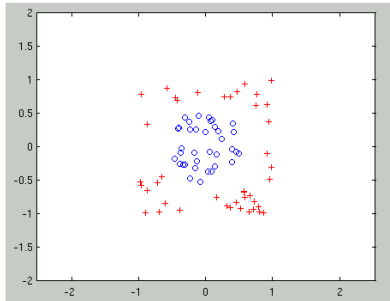
Adding another dimension



Behold yon miserable creature. That Point is a Being like ourselves, but confined to the non-dimensional Gulf. He is himself his own World, his own Universe; of any other than himself he can form no conception; he knows not Length, nor Breadth, nor Height, for he has had no experience of them; he has no cognizance even of the number Two; nor has he a thought of Plurality, for he is himself his One and All, being really Nothing. Yet mark his perfect self-contentment, and hence learn this lesson, that to be self-contented is to be vile and ignorant, and that to aspire is better than to be blindly and impotently happy.

Problems get easier in higher dimensions

$$(x_1, x_2) \Rightarrow (x_1, x_2, \sqrt{x_1^2 + x_2^2})$$



SVM Kernels

Statistics Professors HATE Him!



Doctor's discovery revealed the secret to learning any problem with just 10 training samples. Watch this shocking video and learn how rapidly you can find a solution to your learning problems using this one sneaky kernel trick! Free from overfitting!
<http://www.oneweirdkerneltrick.com>

- Can make training more expensive
- Always increase dimensionality (bad for language)
- Combine well with feature engineering

SVM Kernels

Statistics Professors HATE Him!

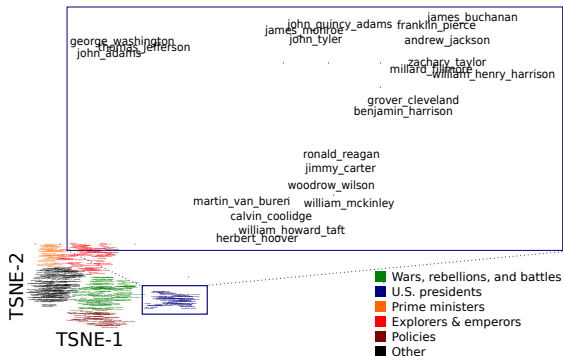


Doctor's discovery revealed the secret to learning any problem with just 10 training samples. Watch this shocking video and learn how rapidly you can find a solution to your learning problems using this one sneaky kernel trick! Free from overfitting!
<http://www.oneweirdkerneltrick.com>

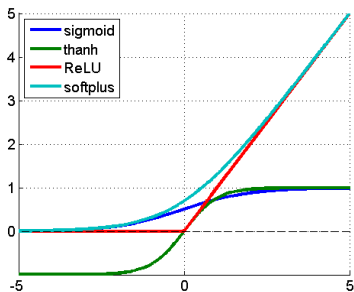
- Can make training more expensive
- Always increase dimensionality (bad for language)
- Combine well with feature engineering
- Still requires feature engineering

Reasons to use fewer dimensions

- Faster
- More interpretable
- Leads to “closeness”
- Need non-linearity (example later)



Non-linearity



- Logistic / Sigmoid

$$f(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

- tanh

$$f(x) = \tanh(x) = \frac{2}{1 + e^{-2x}} - 1 \quad (2)$$

- ReLU

$$f(x) = \begin{cases} 0 & \text{for } x < 0 \\ x & \text{for } x \geq 0 \end{cases} \quad (3)$$

- SoftPlus: $f(x) = \ln(1 + e^x)$