

# Naive Bayes Variations and SVM saved

## Two Classic Naïve Bayes Variants for Text

- **Multinomial Naïve Bayes**
  - Data follows a multinomial distribution
  - Each feature value is a count (word occurrence counts, TF-IDF weighting, ...)
- **Bernoulli Naïve Bayes**
  - Data follows a multivariate Bernoulli distribution
  - Each feature is binary (word is present / absent)

## Case study: Sentiment analysis



one word: wow.

★★★★★★★★

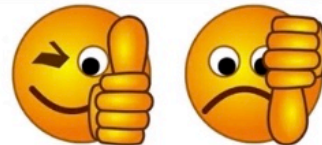
Author: [REDACTED]

10 January 2005

This is a truly great movie. It is one that I can watch over and over, yet can never seem to get enough of! Kate Winslet is gorgeous, Emma Thomson is inspiring, and Hugh Grant shines in this unforgettable film. I have always loved a good movie; one i can sink into and fall in love with the characters. I feel that Sense and Sensibility presents all these things to the audience. There is love, heartbreak, humour and great music. (my applause to Kate Winslet for her unforgettable versions of "Weep You No More Sad Fountains" and "The Dream" ...so beautiful.) If you have not seen this movie, please go RIGHT now to rent it...or even ADD it to your video library! I must say it is well worth it! And if you have seen it. ....you know what i mean.

Brava! Brava! Bravo!

- **Words that you might find in typical reviews**
  - wow, great, Bravo!
  - boring, lame, worst



# Classifier = Function on input data

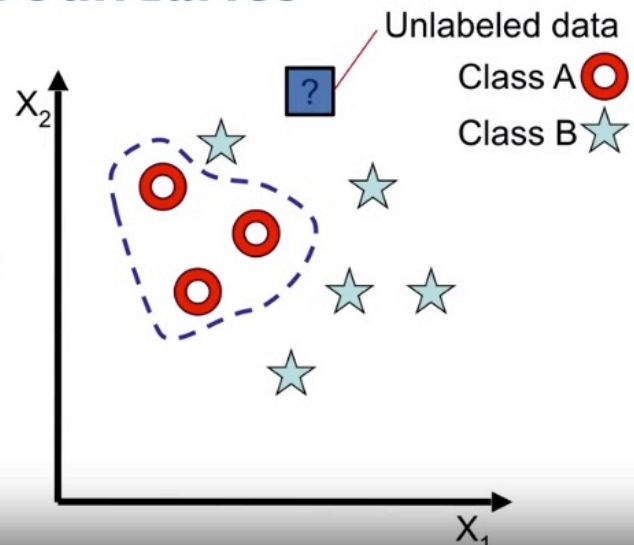
- $f(\text{tinea pedis, or athlete's foot, is a very common fungal skin infection of the foot. It often first appears between the toes. It can be a one-time occurrence or it can be chronic. The fungus, known as Trichophyton, thrives under warm, damp conditions as people whose feet sweat a great deal are more susceptible. It is easily transmitted in showers and pool walkways. Those people with immunosuppressive conditions, such as diabetes mellitus, are also more susceptible to athlete's foot.}) \rightarrow \{\text{kidney, brain, foot}\}$

- $g(\text{your overall score}) \rightarrow \{+1, -1\}$



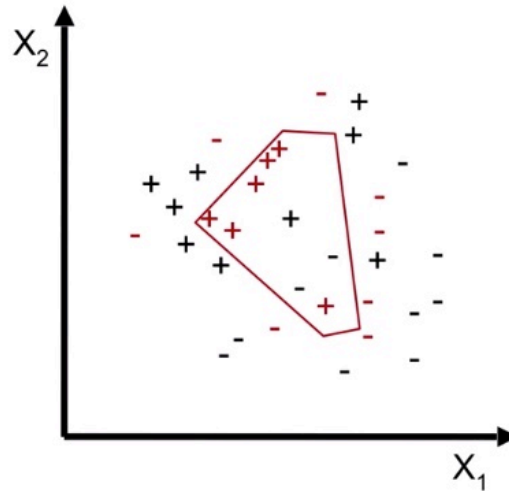
## Decision Boundaries

- Classification function is represented by decision surfaces
  - How do you find them?



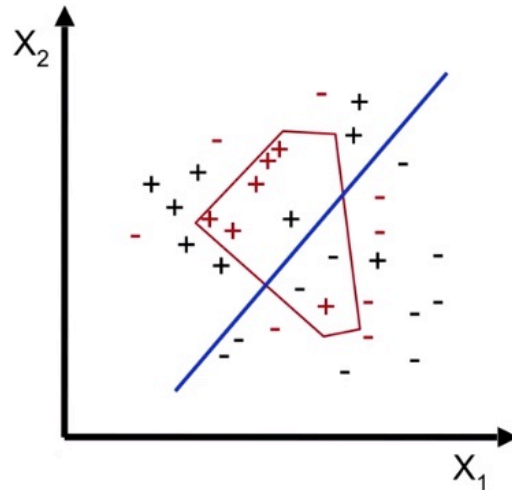
# Choosing a Decision Boundary

- Red +/- : Training data
- Black +/- : Test data
- **Data overfitting:** Decision boundary learned over training data doesn't generalize to test data



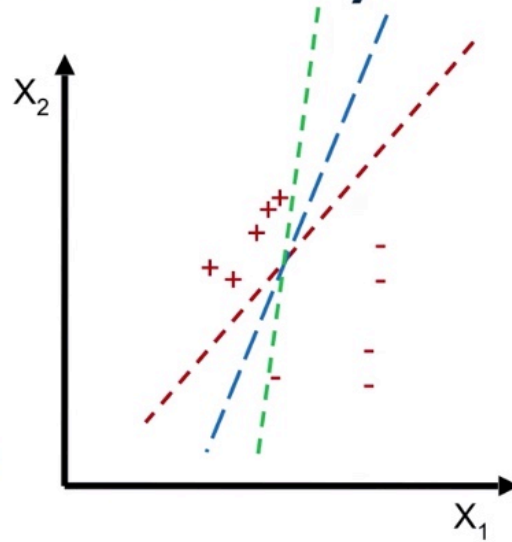
## Linear Boundaries

- Easy to find
- Easy to evaluate
- More generalizable: "Occam's razor"



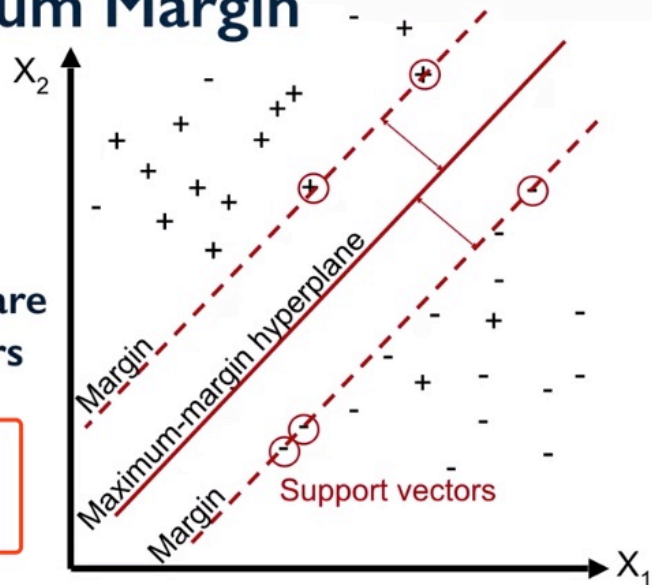
# Finding a Linear Boundary

- Find the linear boundary = Find  $w$
- Many methods
  - Perceptron
  - Linear Discriminative Analysis
  - Linear least squares
  - ...
- Problem: If linearly separable, then infinite number of linear boundaries!



## Maximum Margin

- What is a reasonable boundary?
  - **Support Vector Machines** are maximum-margin classifiers
- How do you find it?









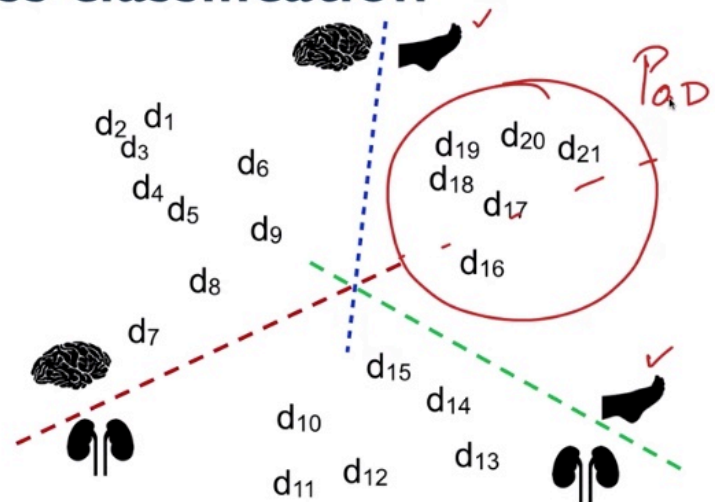
# Support Vector Machines (SVM)

- SVMs are **linear classifiers** that find a hyperplane to separate **two classes** of data: positive and negative
- Given training data  $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots$ ; where  $\mathbf{x}_i = (x_1, x_2, \dots, x_n)$  is instance vector and  $y_i$  is one of  $\{-1, +1\}$ 
  - SVM finds a linear function  $w$  (weight vector)
$$f(\mathbf{x}_i) = \langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b$$
if  $f(\mathbf{x}_i) \geq 0, y_i = +1$ ; else  $y_i = -1$

## SVM: Multi-class classification

- One vs One

-  vs. 
-  vs. 
-  vs. 



## SVM Parameters (I): Parameter C

- Regularization: How much importance should you give individual data points as compared to better generalized model
- Regularization parameter  $c$ 
  - Larger values of  $c$  = less regularization
    - Fit training data as well as possible, every data point important
  - Smaller values of  $c$  = more regularization
    - More tolerant to errors on individual data points



## SVM Parameters (2): Other params

- Linear kernels usually work best for text data
  - Other kernels include rbf, polynomial
- multi\_class: ovr (one-vs-rest)
- class\_weight: Different classes can get different weights

## Take Home Messages

- Support Vector Machines tend to be the most accurate classifiers, especially in high-dimensional data
- Strong theoretical foundation
- Handles only numeric features
  - Convert categorical features to numeric features
  - Normalization
- Hyperplane hard to interpret