Lecture: SVM

Dr. Goutam Chakraborty

SAS® Professor of Marketing Analytics

Director of MS in Business Analytics and Data Science* (http://analytics.okstate.edu/mban/)
Director of Graduate Certificate in Business Data Mining (http://analytics.okstate.edu/certificate/grad-data-mining/)
Director of Graduate Certificate in Marketing Analytics (http://analytics.okstate.edu/certificate/grad-marketing-analytics/)

1

1

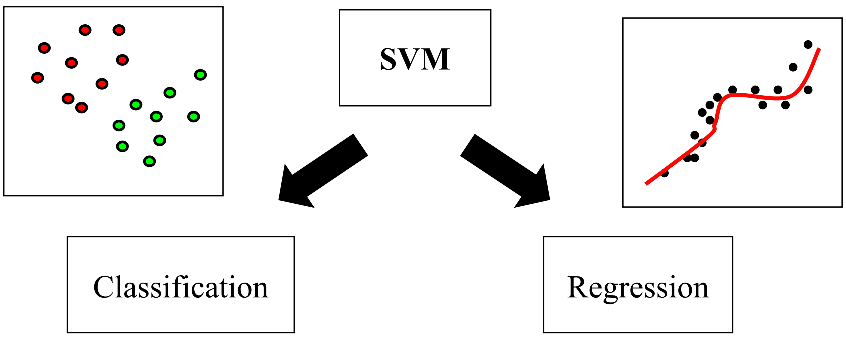
Outline

- What is SVM?
- Some ideas about how it works.

2

2

Tasks Handled by SVMs



3

3

History of SVM

- Theoretically well motivated algorithm: developed from Statistical Learning Theory (Vapnik & Chervonenkis) since the 60's.
- SVM applications started appearing in machine learning literature since 1990's and has become quite popular with data scientists.
- Empirically good performance (as predictive models): successful applications in many fields (bioinformatics, text, image recognition,…)
- A good source for resources is located at: http://www.kernel-machines.org/

4

4

Classification Model: Starting Point

- Training data set: Patients with known diagnoses
- Input variables: Data about patients

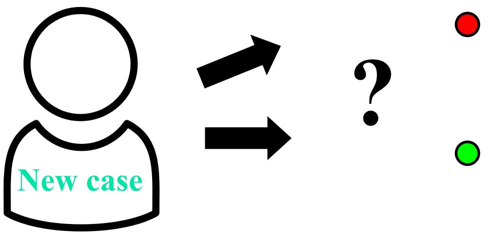
  $x_i \in \mathcal{R}^d$

- Response variable: two diseases

$$y_i \in \{+1, -1\}$$

Classification Function

- Classification function: $f: \quad \mathcal{R}^d \mapsto \{+1, -1\}$
- Diagnosis = *f*(new patient)



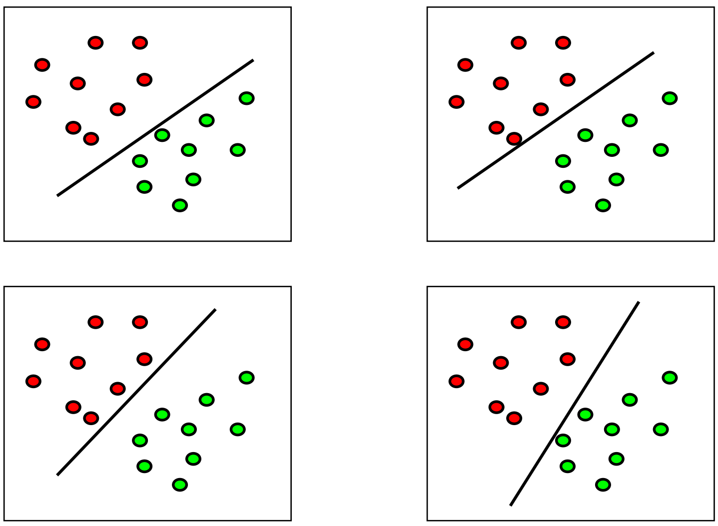
Classifying Data into 2 Classes

- In machine learning and statistics, *classification* is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.
- In SVM, a data point is viewed as a p-dimensional vector (or, p-input variables), and we want to know whether we can separate such points with a (p − 1) dimensional hyperplane.
  - This is called a linear classifier.
  - For 2 input variables, the hyperplane becomes a straight line

7

7

How to Classify Red versus Green?



8

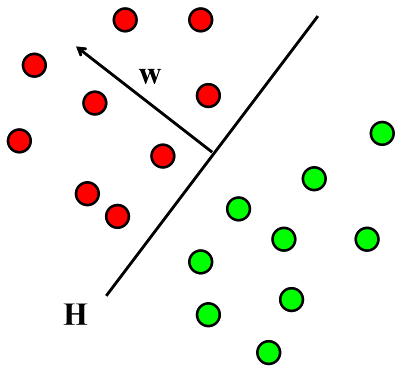
8

Linear Separation of the Training Data

- A separating hyperplane H is given by
  - the normal vector ***w***, (the direction of positive class)
  - an additional parameter, b, called bias.

$$H = \{x \mid \underbrace{\langle w, x \rangle} + b = 0\}$$

Dot product

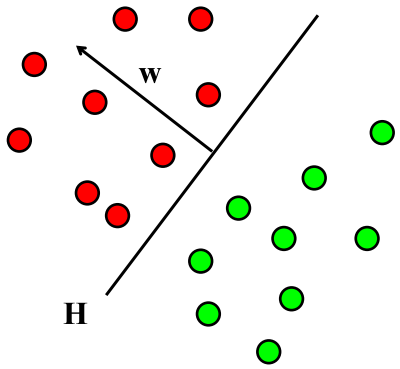


Training versus Prediction

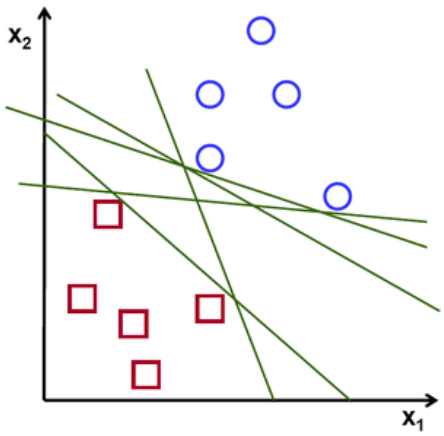
- Training:

Select ***w*** and ***b*** in such a way that the hyperplane separates the training data—that is, construction of a hyperplane.

- Prediction of the class for a new patient: On which side of the hyperplane is the new data point located?



Linear Classifier



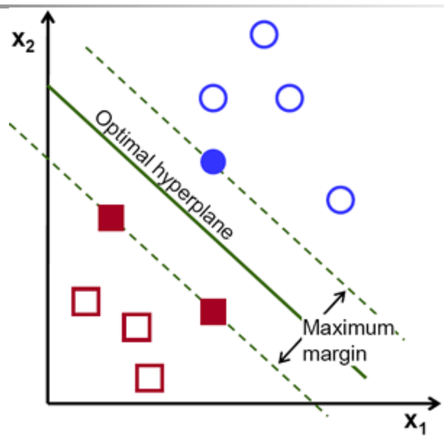
Which hyperplane (different straight lines in the diagram above) is the best that separates the circles from the squares?

Question: What happens when we build a decision tree to handle this problem?

11

11

SVM Approach



The boundary lines are the support vectors. The optimal hyperplane maximizes the decision surface to the boundaries of two classes (or, it maximizes the margin)

12

12

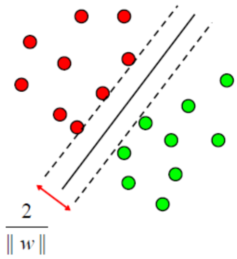
Optimization Problem

- To ensure correct separation, use the constraints:

$$\langle w, x_i \rangle + b \geq 1 \quad \text{if } y_i = 1$$

$$y_i \cdot (\langle w, x_i \rangle + b) \geq 1 \quad \text{for i=1, ..., n}$$

$$\langle w, x_i \rangle + b \leq -1 \quad \text{if } y_i = -1$$



Maximize the **separating distance**

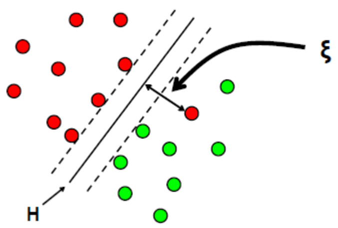
Minimize $\| w \|^2$

13

13

When Data are Not Separable

- Use a penalty : C (distance to hyperplane)



ξ allows for errors.

Optimization problem becomes: **Minimize** $\| w \|^2 + C \cdot \sum_i \xi_i$

under the condition

$$y_i \cdot (\langle w, x_i \rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

14

14

Lagrange Approach

- Lagrange function:

$$L(w,b,\alpha,\xi)=\frac{1}{2}\|w\|^2+C\cdot\sum_{i=1}^{n}\xi_i-\sum_{i=1}^{n}\alpha_i\left(\xi_i+y_i\left(\langle w,x_i\rangle+b\right)-1\right)$$

- Minimize $L(w,b,\alpha,\xi)$ for w, b, ξ.
- Maximize $L(w,b,\alpha,\xi)$ for α_i.

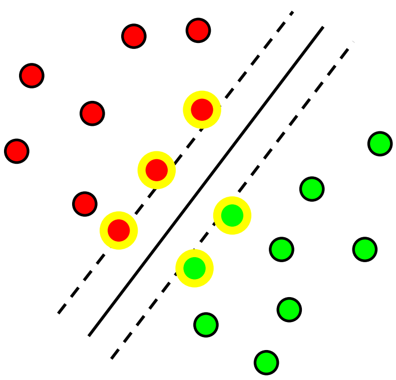
15

15

What Are the Support Vectors?

- “Carrying vectors”
- The points, located closest to the hyperplane
- Determining the location of the hyperplane
- All other data points have $\alpha_i = 0$.

$$w=\sum_{i=1}^{\#sv}\alpha_i y_i x_i^{sv}$$



16

16

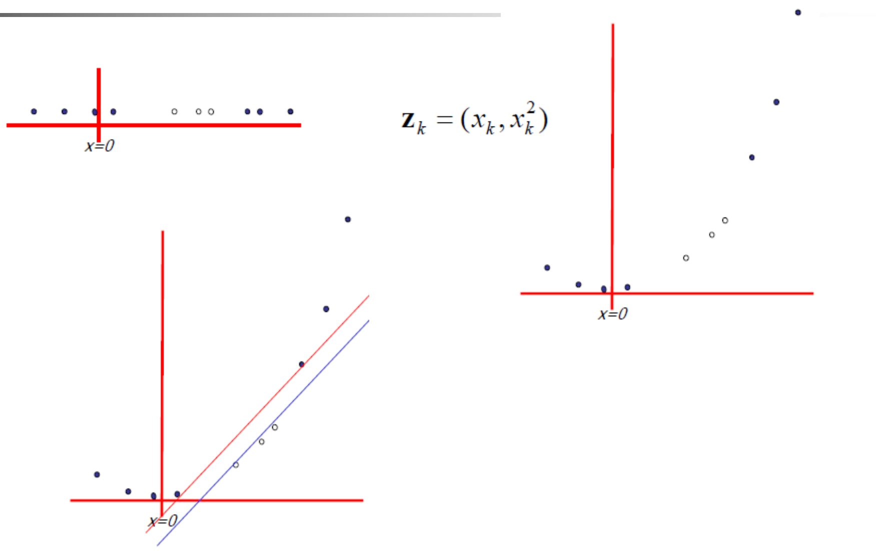
What happens when data are not separable nicely?

- Think of the observed data in the 2-dimenisonal plane as a projection from a higher-dimensional plane (say 3-dimensions).
- **Assumption**: in the higher dimensional plane the data are *nicely separable*
- If we can transform (via a function) the observed 2-dimensional representation to that higher dimensions, then we can find the hyperplane
- This special transformation function is called *kernel function*.
  - The challenge is finding the right kernel function and its parameters.
  - This is an optimization task in finding right weights for a neural network

17

17

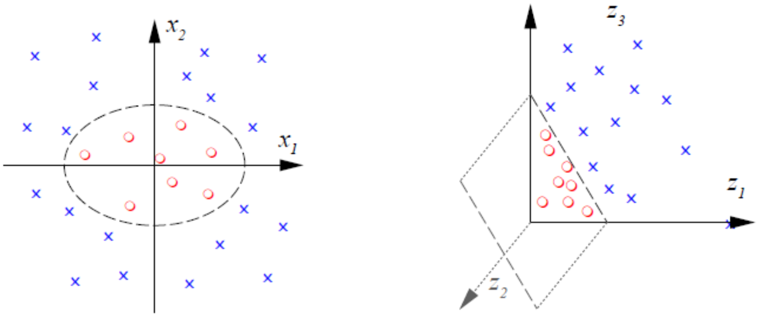
Mapping (From 1 to 2 dimensions)



18

18

Mapping (From 2 to 3 dimensions)



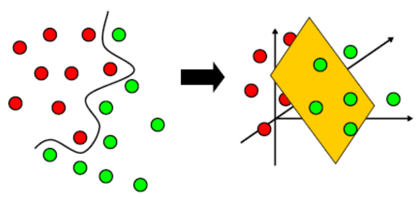
$$\Phi : R^2 \rightarrow R^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{(2)}x_1x_2, x_2^2)$$

19

19

The Kernel Trick

- We don't have to know exactly what the feature space (the higher dimensional representation) looks like.
- It is enough to specify the kernel functions (such as linear, polynomial, RBF, etc.).
  - But, we have the geometric interpretation in the form of a separating hyperplane, i.e. more transparencv.



20

20

Mathetmatical Challenge

- Dual optimization problem:

$$W(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$

- Classification function

$$f(x_{new}) = sign\left(\sum_{i=1}^{n} \alpha_i y_i \langle x_i, x_{new} \rangle + b \right)$$

Dot product

Feature Space $\mathcal{F}$

- The data points are transformed with a function Φ:

$$\Phi: \quad \mathcal{R}^d \mapsto \mathcal{F}$$

$$x \mapsto \Phi(x)$$

- Then we separate the data points Φ(x) in the feature space F.

Solution: The Kernel Trick

- We want to construct the separating hyperplane in the feature space.
- **Problem:**
  Dot products of the form

  $$\langle \Phi(x_i), \Phi(x_j) \rangle$$

  are difficult to calculate.

Solution: The Kernel Trick

- We use a kernel function, living in $\mathcal{R}^d$, but behaving as a dot product in the feature space:

$$\mathcal{K}(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle$$

- Trick: We do not have to know Φ(x) explicitly!

Examples of Kernel Functions

- Linear

$$K(x_i, x_j) = \langle x_i, x_j \rangle$$

- Polynomial

$$K(x_i, x_j) = \left(\gamma \langle x_i, x_j \rangle + k\right)^d$$

- Radial-Basis-Function

$$K(x_i, x_j) = \exp\left(-\gamma \| x_i - x_j \|^2\right)$$

- Sigmoid

$$K(x_i, x_j) = \tanh\left(\gamma \langle x_i, x_j \rangle - b\right)$$

25

25

Summary of SVM

- An *SVM* is a hyperplane with a maximum-margin in a feature space, constructed by use of a kernel function in the input space.
- Parmeters for SVMs for Classification are:
  - The penalty C (regularization term)
  - The kernel function and its parameters

26

26

Advantages of SVMs

- Finds a global, unique minimum
- The kernel trick
- A simple geometric interpretation
- Strong ability to generalize
- The complexity of the calculations does not depend on the dimension of the input space.
  - This avoids the *curse of dimensionality*.

27

27

Disadvantages of SVMs

- Which kernel function to use?
- How to select the parameters of the kernel function?

28

28