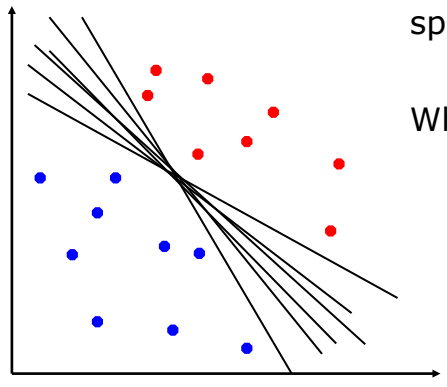
Fundamentals of Artificial Intelligence

Support Vector Machines

Shyamanta M Hazarika
Mechanical Engineering
Indian Institute of Technology Guwahati
s.m.hazarika@iitg.ac.in

http://www.iitg.ac.in/s.m.hazarika/

Linear separation

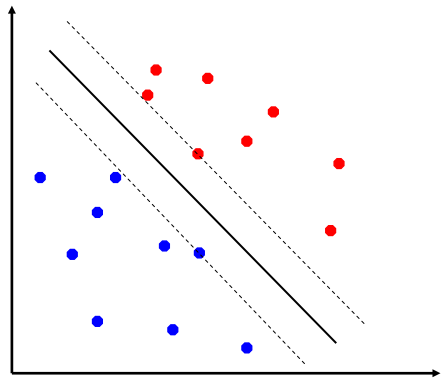


Binary classification can be viewed as the task of separating classes in feature space:

Which of the linear separators is optimal?

Classification Margin

Have some negative examples; positive examples.
Let the RED be the +ve and BLUE be the –ve.



How do you divide the positive examples from the negative examples?

Line with a view toward putting in the **widest street** separating the positive from the negative samples.

Two parallel hyperplanes that separate the two classes of data, so that the distance between them is as large as possible

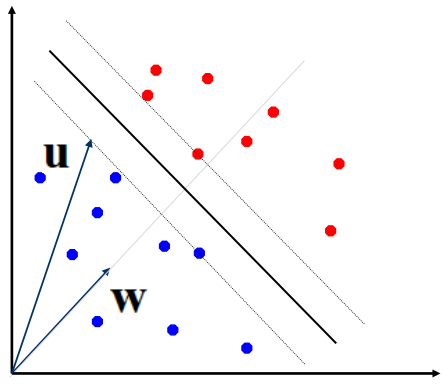
Examples closest to the hyperplane are **support vectors**.

Margin of the separator is the distance between support vectors.

Classification Margin

w – Vector constrained to be perpendicular to the `median'.

u – Vector that points to an unknown.



Whether the unknown is on the +ve side or on the –ve side?

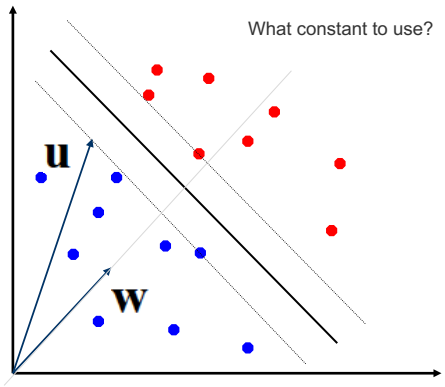
Project **u** on **w** to that is perpendicular to the separator. Reflects distance in this direction!

Further out we go, the closer we'll get to the positive samples.

$$(\mathbf{w}.\mathbf{u}) \geq c$$

Classification Margin

Lot of w's that are perpendicular to the median line; because it could be of any length.



Classify unknown **u** as plus if

Decision Rule.

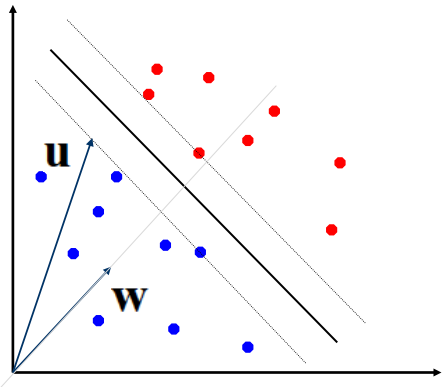
$$f(\mathbf{u}) = \mathbf{w} \cdot \mathbf{u} + b \geq 0$$

Don't have enough constraint here to fix a particular b or a particular $\mathbf{w}$.

$\mathbf{w}$ has to be perpendicular to the median line.

Classification Margin

For a positive sample, insist that the decision function gives the value of one or greater.



For a negative sample, insist that the decision function gives the value of equal to or less than minus 1.

Classify unknown **u** as plus if

$$f(\mathbf{u}) = \mathbf{w} \cdot \mathbf{u} + b \geq 0$$

Constrain

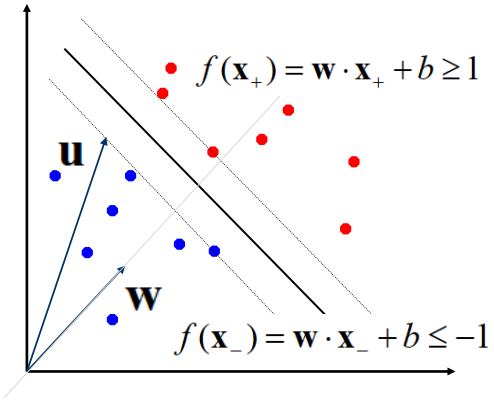
For all plus sample vectors:

$$f(\mathbf{x}_+) = \mathbf{w} \cdot \mathbf{x}_+ + b \geq 1$$

For all minus sample vectors:

$$f(\mathbf{x}_-) = \mathbf{w} \cdot \mathbf{x}_- + b \leq -1$$

Classification Margin



Introduce a variable y_i

$y_i = +1$ for pluses

-1 for minuses

Multiply each equation by the variable y_i.

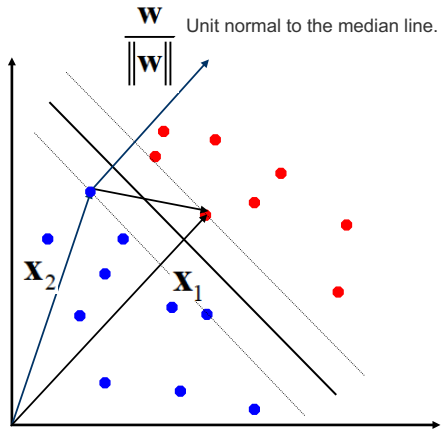
$$y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1$$

$$y_i(\mathbf{w}.\mathbf{x}_i + b) - 1 \geq 0$$

For all points on the boundary

$$y_i(\mathbf{w}.\mathbf{x}_i + b) - 1 = 0$$

Classification Margin



The dot product of the unit normal and the difference vector, would be the margin.

Given the constraints

$$y_i(\mathbf{w}.\mathbf{x}_i + b) - 1 = 0$$

Width?

$$\mathbf{w}\cdot\mathbf{x}_1 + b = +1$$

$$\mathbf{w}\cdot\mathbf{x}_2 + b = -1$$

Width of the margin

$$\frac{\mathbf{w}}{\|\mathbf{w}\|}\cdot(\mathbf{x}_1 - \mathbf{x}_2) = \frac{2}{\|\mathbf{w}\|}$$

Linear Support Vector Machine

- To maximize the width of the separation, one need to minimize $\mathbf{w}$, while still honoring constraints with regards to the values on the edges of the separation.
- One possible approach to finding the minimum is to use the method devised by Lagrange.

Lagrange multipliers; would give us a new expression, which we can maximize or minimize without thinking about the constraints anymore.

- Maximizing the width is ensured if one minimize the following, while honoring constraints on edge values.

$$\frac{1}{2}\|\mathbf{w}\|^2$$

Translation of the previous formula into this one, with ½ and squaring, is a mathematical convenience.

Linear Support Vector Machine

- To minimize $\frac{1}{2}\|\mathbf{w}\|^2$ subject to the constraint

$$y_i(\mathbf{w}.\mathbf{x}_i + b) - 1 \geq 0$$

- The Lagrangian is written; using lagrangian multipliers. Give us new expressions without constraints:

Each or those constraints is going to have a multiplier, alpha$_i$

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} a_i(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1)$$

Linear Support Vector Machine

- To minimize $\frac{1}{2}\|\mathbf{w}\|^2$ subject to the constraint

$$y_i(\mathbf{w}.\mathbf{x}_i + b) - 1 \geq 0$$

- Find where the Lagrangian has zero derivatives

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} a_i(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1)$$

i.e., $\frac{\partial L}{\partial \mathbf{w}} = 0$ and $\frac{\partial L}{\partial b} = 0$

Linear Support Vector Machine

$$L = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{l} a_i(y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1)$$

Decision vector is a linear sum of the samples.

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} a_i y_i \mathbf{x}_i = 0 \qquad \mathbf{w} = \sum_{i=1}^{l} a_i y_i \mathbf{x}_i$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{l} a_i y_i = 0$$

Find a maximum of this expression.

$$L = \sum_{i=1}^{l} a_i - \frac{1}{2}\sum_{i,j=1}^{l} a_i a_j y_i y_j \mathbf{x}_i \cdot \mathbf{x}_j$$

Optimization depends only on the dot product of pairs of samples.

Linear Support Vector Machine

- Recall that the decision rule was to

  Classify unknown $\mathbf{u}$ as plus if

$$f(\mathbf{u}) = \mathbf{w} \cdot \mathbf{u} + b \geq 0$$

- With $\mathbf{w} = \sum_{i=1}^{l} a_i y_i \mathbf{x}_i$ the decision rule is

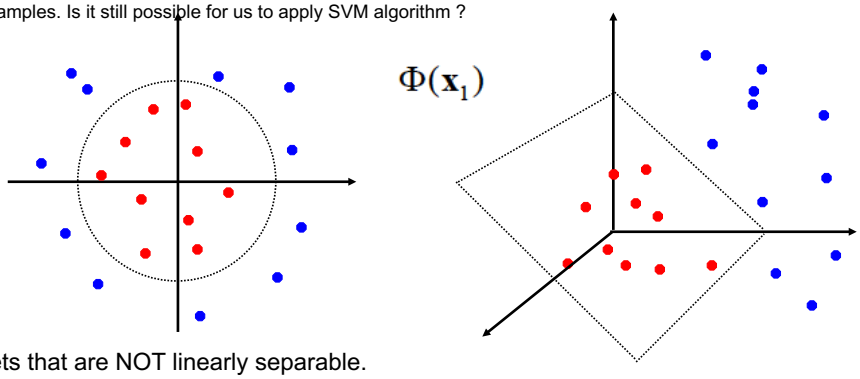
The decision rule, also, depends only on the dot product of the sample vectors and the unknown.

If $\sum_{i=1}^{l} a_i y_i \mathbf{x}_i \cdot \mathbf{u} + b \geq 0$ then classify $\mathbf{u}$ as plus.

Nonlinear Classification

- General idea: the **original feature space** can always be mapped to some **higher-dimensional feature space** where the training set is separable:

Impossible to draw a line in the 2D plot which could separate the blue from the red samples. Is it still possible for us to apply SVM algorithm ?



Datasets that are NOT linearly separable.

Nonlinear Classification

- Optimization and also the decision rule, require inner product.
  - $\Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$ for optimization.
  - $\Phi(\mathbf{x}_i).\Phi(\mathbf{u})$ for decision rule.
- All that is required is a **way to compute dot products in high-dimensional space** as a function of vectors in original space!
- A **kernel function** is a function that is equivalent to an inner product in some feature space.
  - $K(\mathbf{x}_i,\mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$

We just need to know *K* and not the mapping function itself. Doesn't need to know what Φ is; having *K* is enough!

Nonlinear Classification

The kernel trick avoids the explicit mapping to get linear learning algorithms to learn a nonlinear function or decision boundary.

1. Linear Kernel $K(\mathbf{x}_i,\mathbf{x}_j)= \mathbf{x}_i.\mathbf{x}_j$

   Linear Kernel is used when the data is **Linearly** separable, that is, it can be separated using a single Line. It is one of the **most common kernels** to be used. Used when there are a Large number of Features in a particular Data Set.

2. Polynomial Kernel $K(\mathbf{x}_i, \mathbf{x}_j)= (1+ \mathbf{x}_i.\mathbf{x}_j)^n$

   Represents the **similarity of vectors in a feature space over polynomials** of the original variables,. Intuitively, the kernel looks not only at the given features of input samples, but also combinations of these.

3. Gaussian Kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}}$

   Gaussian kernel computed with a support vector is an **exponentially decaying function** in the input feature space, the maximum value of which is attained at the support vector and which decays uniformly in all directions around the support vector, leading to **hyper-spherical contours of the kernel function**.

Main Ideas

- **Maximum-Margin Classifier**
  - Formalize notion of the best linear separator
  - Best hyperplane is the one that represents the largest separation, or margin, between the two classes - distance from it to the nearest data point on each side is maximized.
  - If such a hyperplane exists, it is known as the maximum-margin hyperplane and the linear classifier it defines is known as a maximum-margin classifier
- **Lagrangian Multipliers**
  - Way to convert a constrained optimization problem to one that is easier to solve
- **Kernels**
  - Projecting data into higher-dimensional space makes it linearly separable.

Final Comments

- The original SVM algorithm was invented by Vladimir N. Vapnik in 1963; **gained increasing popularity in late 1990s.**
- SVMs are **currently among the best performers** for a number of classification tasks ranging from text to genomic data.
- SVMs **can be applied to complex data types** beyond feature vectors (e.g. graphs, sequences, relational data) by designing kernel functions for such data.
- SVM **techniques have been extended** to a number of tasks such as regression, principal component analysis etc.