Support vector machines

Support vector machines, I

- Developed for binary classification (Vapnik, Statistical Learning Theory, 1998)
- Linear SVM aims to find the separating hyperplane $\beta' x - \beta_0 = 0$, that has the maximal distance from the learning sets of the two classes.
- If classes are linerarly separable, separating hyperplane with biggest "margin" can be found.
- If classes are not linearly separable, separating hyperplane cannot be found. Each choice includes some misclassification. Total error (distance from hyperplane on the wrong side) is bounded above.
- Formally, linear SVM solves optimization problem:
  $min_{(\beta,\beta_0,\xi)} \beta'\beta$ subject to
  $x_i(\beta' x - \beta_0) \geq 1 - \xi_i, \forall i, \ and \ \xi_i \geq 0, \ with \ \sum \xi_i \leq C$

- Separating hyperplanes
- Maximal margin hyperplane
- Soft margin hyperplanes
- Non-linear separation

A separating hyperplane

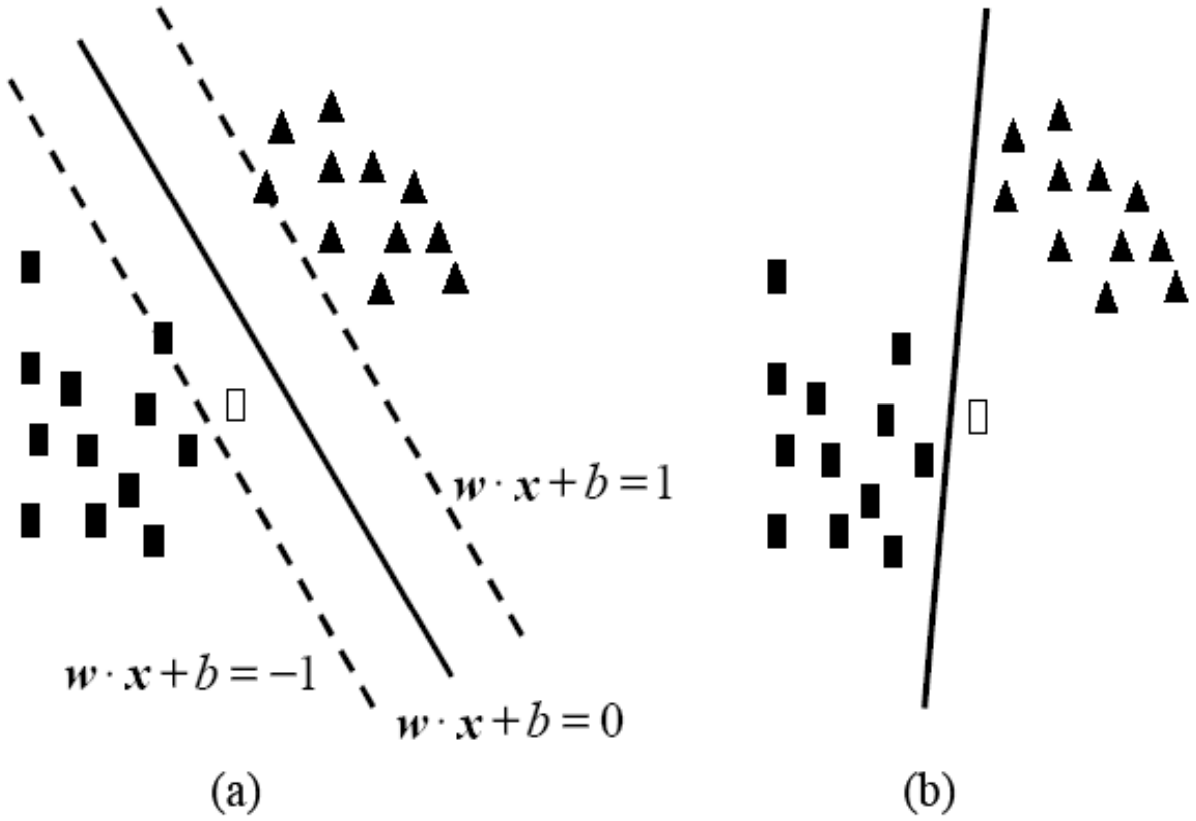


Figure 9.3.(a) The maximum margin hyperplane separating two classes. The solid black line is the hyperplane ($\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$). The two dashed lines are those for the points in the two classes closest to the hyperplane ($\boldsymbol{w} \cdot \boldsymbol{x} + b = \pm 1$). A new point, the blank rectangle, is classified correctly in (a). Note, the larger the margin the greater the deviation allowed or margin for error. (b) A non-maximum margin hyperplane separating the two classes. Note, that the same new point is now classified incorrectly. There is less margin for error.

What is a hyperplane

- A hyperplane is a flat subspace of dimension p-1 in a p-dimensional space
- In 2-D space this is a straight line
- In 3-D space this is a plane
- In high $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0.$ visualize but it can be easily written as a linear constraint:

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p = 0$$

- A hyperplane then divides the p-dimensional space to two halves

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p > 0 \qquad \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p < 0$$

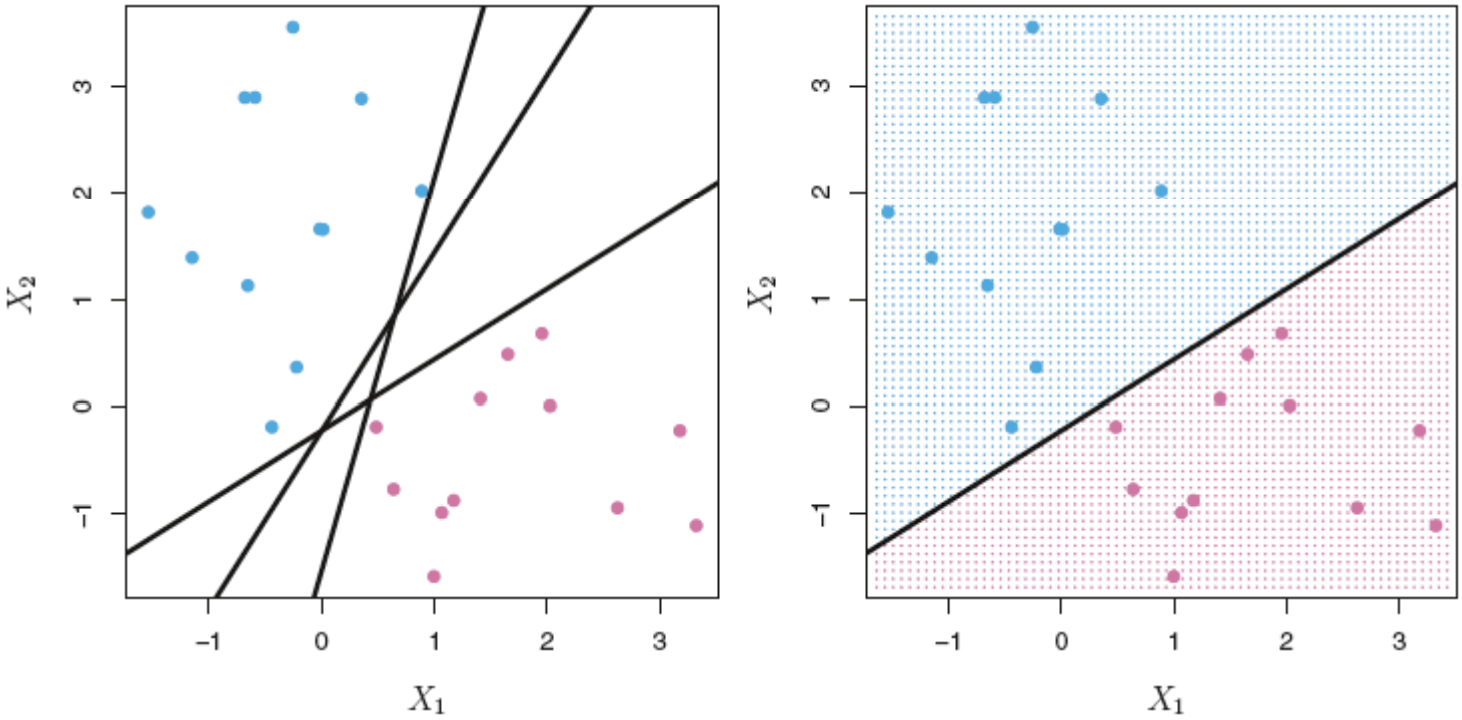
- Consider two classes: $y_1, y_{2,\ldots} y_n \in \{-1,1\}$
- Each outcome has features (p-dimensional)

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$$

- A separating hyperplane has property

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) > 0$$

(i.e. the prediction and the actual y have the same sign for all observations)



Left: multiple separating hyperplanes

Right: one particular hyperplane and the classifier

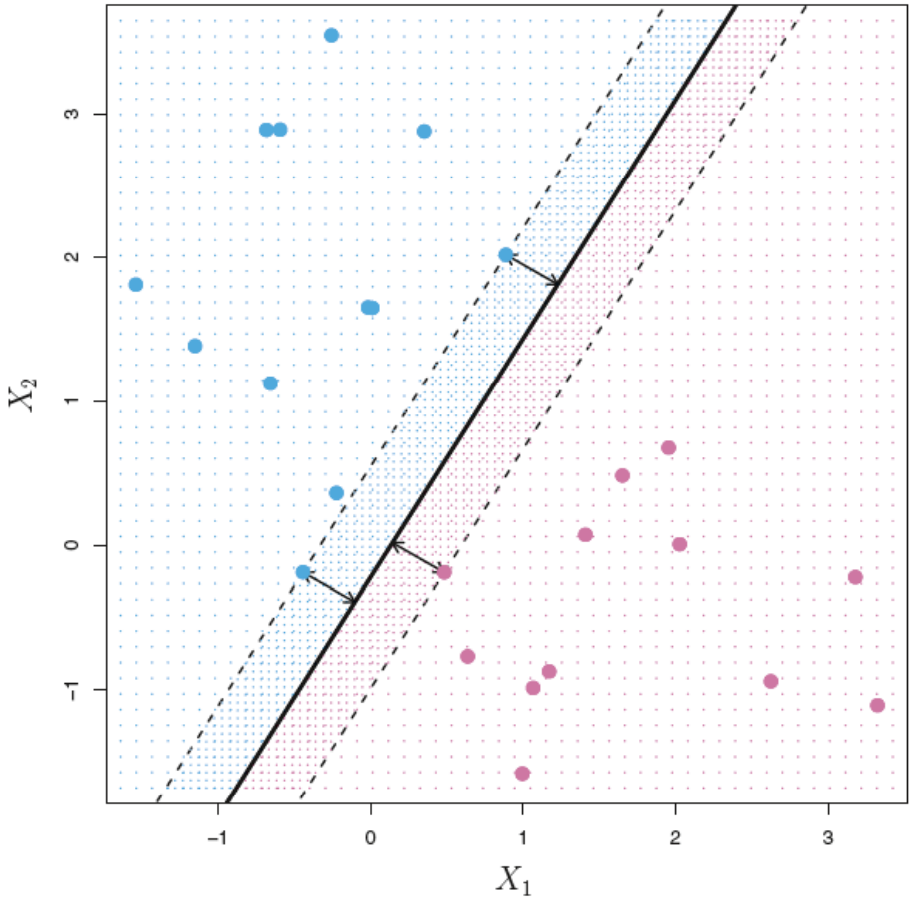
Margin: the minimal distance between observations to the hyperplane

Maximal margin classifier:

The hyperplane that maximizes margin.

Support vectors:

Training observations on the margin plane



Optimization problem:

$$\underset{\beta_0,\beta_1,\ldots,\beta_p}{\text{maximize}}\, M$$

$$\text{subject to} \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M \;\; \forall\, i = 1, \ldots, n.$$

What if it is not separable?

- Allowing misclassifications of training observations

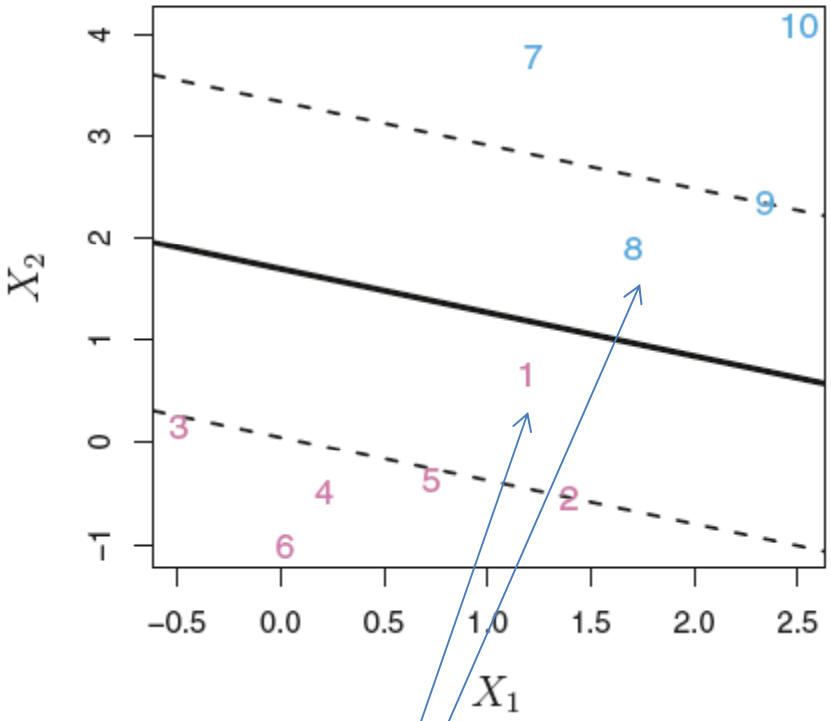
$$\underset{\beta_0,\beta_1,\ldots,\beta_p,\epsilon_1,\ldots,\epsilon_n}{\text{maximize}} \quad M$$

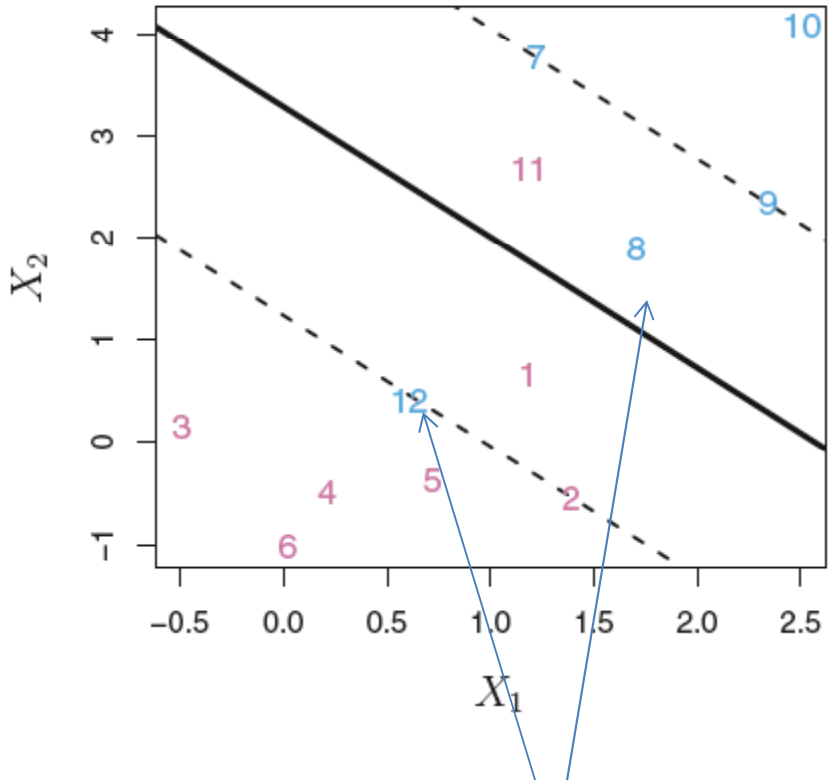
$$\text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 = 1,$$

$$y_i(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}) \geq M(1 - \epsilon_i)$$

$$\epsilon_i \geq 0, \quad \sum_{i=1}^{n} \epsilon_i \leq C,$$

- Slack variable
  - 0: correct sideϵ_i margin
  - (0,1) correct side of plane but wrong side of margin
  - >1: wrong side of plane
- C: tuning parameter for error allowance
  - Maximum number of observations on the wrong side of the plane





Vialates margin but not plane

violates the plane

- Soft margin classifier: Only the observations that either
  - Lie on the margin
  - Violate the margin

  affect the hyperplane

  Observations that affect the hyperplane are the **support vectors**. The tuning parameter C controls how many support vectors are included.

Bias-variance trade off again

- C large:
  - Allow for more violation of margin
  - More support vectors
  - Less variance, more stable
  - Potential high bais

- C small:
  - less violation on training data allowed
  - Low training error, less bias
  - Fewer support vectors
  - Higher variance

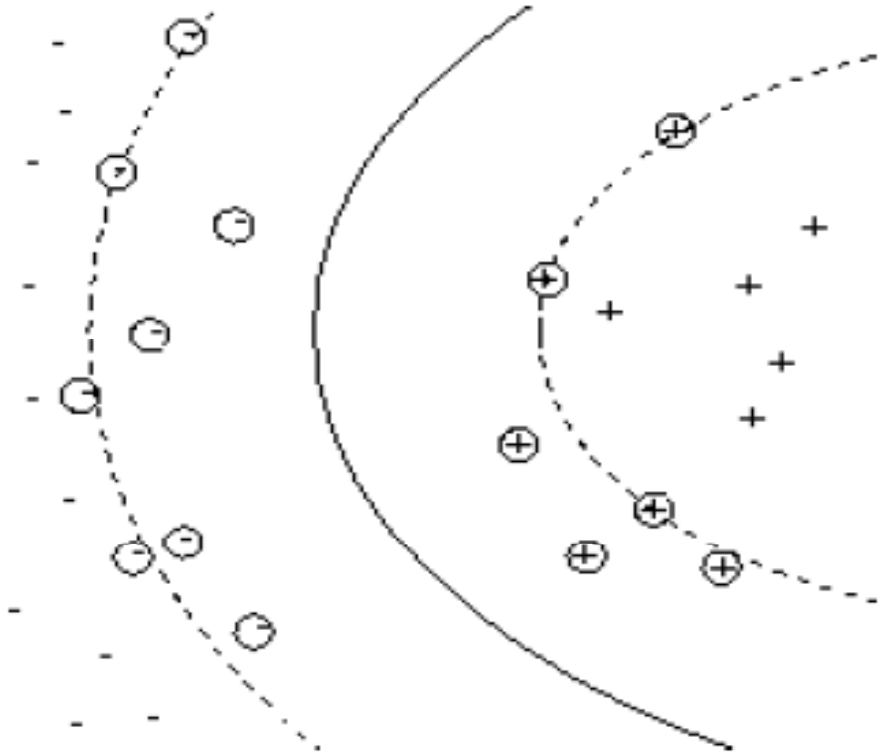
Extending the linear support vector classifier

- It turns out that the optimization problem only involves the *inner products* of the observations.
- For p-dimensional vectors the inner product is

$$\langle x_i, x_{i'} \rangle = \sum_{j=1}^{p} x_{ij} x_{i'j}$$

- The inner product is just one way of measuring similarity. We could redefine the similarity with a *kernal* , for example

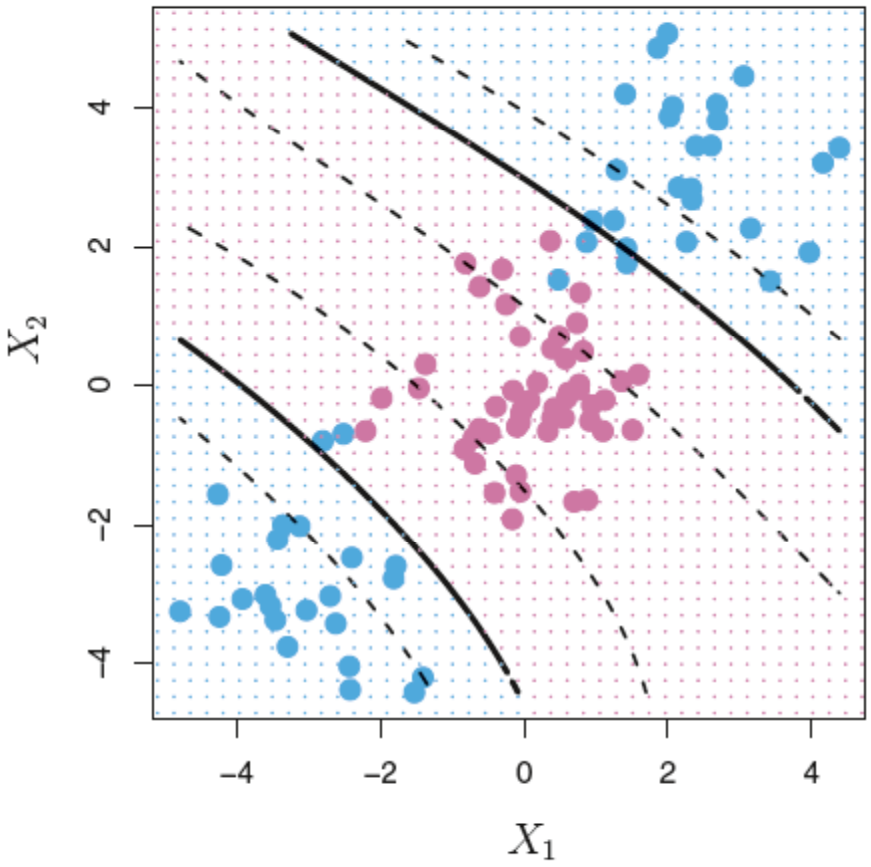
$$K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_i - \mathbf{x}_j\|^2 / 2\sigma^2}$$



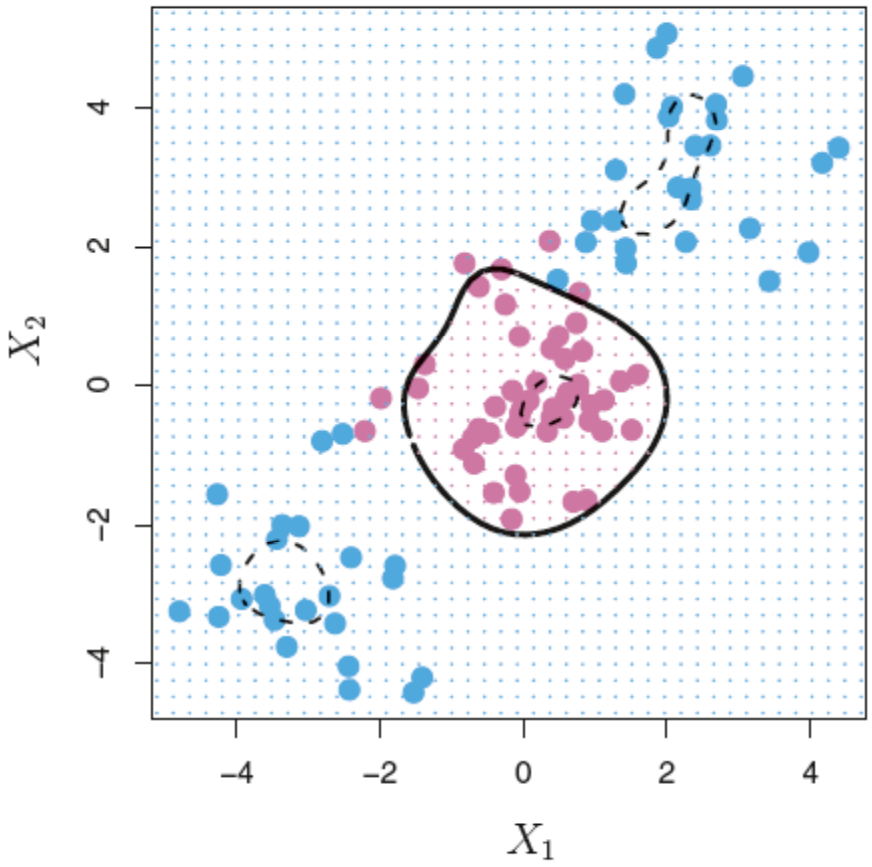
A separating hyperplane in the feature space may correspond to a non-linear boundary in the input space. The figure shows the classification boundary (solid line) in a two-dimensional input space as well as the accompanying soft margins (dotted lines). Positive and negative examples fall on opposite sides of the decision boundary. The support vectors (circled) are the points lying closest to the decision boundary.

Kernal examples

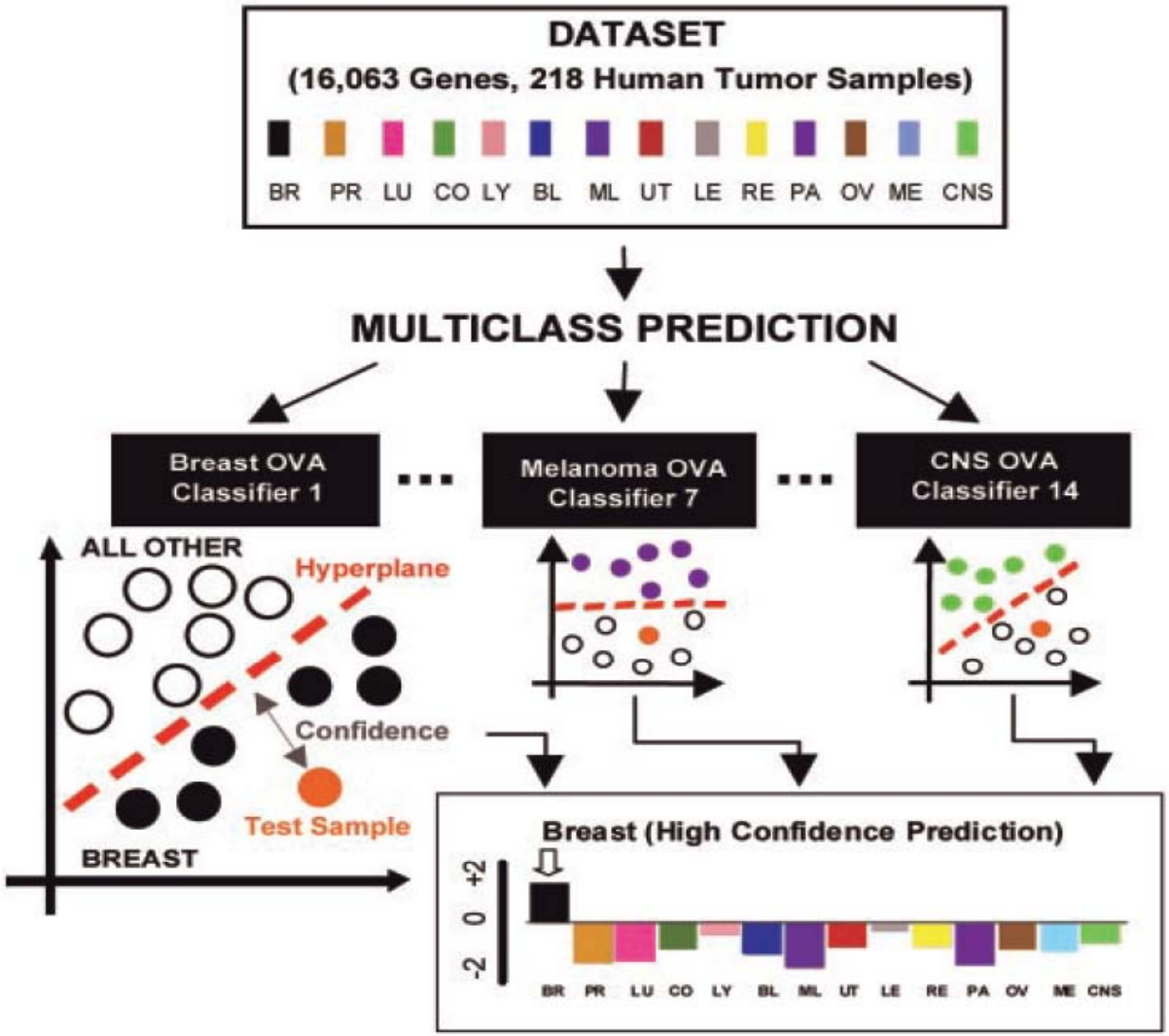
Polynomial kernal



Radial kernal



SVM used in multiclass classification: one versus all scheme



Resources for learning SVM and application in microarrays

- SVM classification and validation of cancer tissue samples using microarray expression data (T S Furey et al, 2000 Bioinformatics)
- **Support Vector Machine Classification of Microarray Gene Expression Data**

http://www.cse.ucsc.edu/research/compbio/genex/genexTR2html/genex.html

- **CLASSIFYING MICROARRAY DATA USING SUPPORT VECTOR MACHINES**

http://cbcl.mit.edu/projects/cbcl/publications/ps/svmmicro.pdf