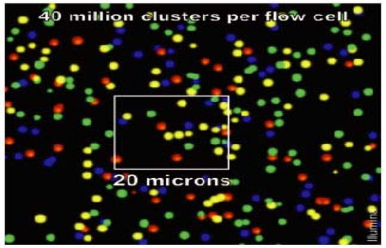
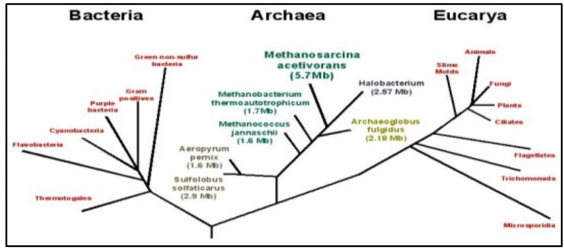
生物信息学：导论与方法

Bioinformatics: Introduction and Methods

TAACCCTAACCCTAACCCTAACCCTAACCCTA
CCTAACCCTAACCCTAACCCTAACCCTAACCC
CCCTAACCCCTAACCCTAACCCTAACCCTAAC
AACCCTAACCCTAACCCTAACCCCTAACCCTA
ACCCTAACCCCAACCCCAACCCCAACCCCAAC
CTACCCTAACCCTAACCCTAACCCTAACCCTA
ACCCTAACCCTAACCCTAACCCCTAACCCTAA



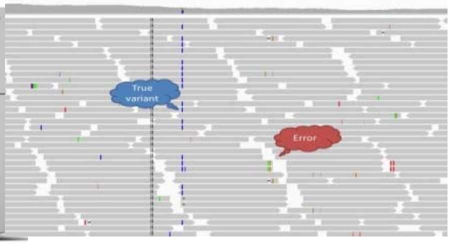


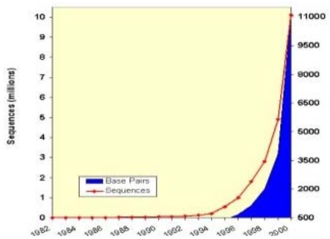
Support Vector Machine(SVM)

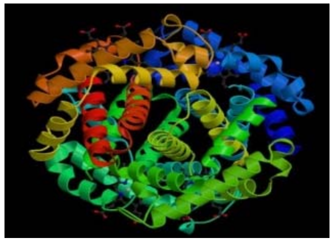
北京大学生物信息学中心 王萌

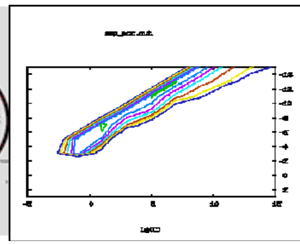
Meng Wang

Center for Bioinformatics, Peking University

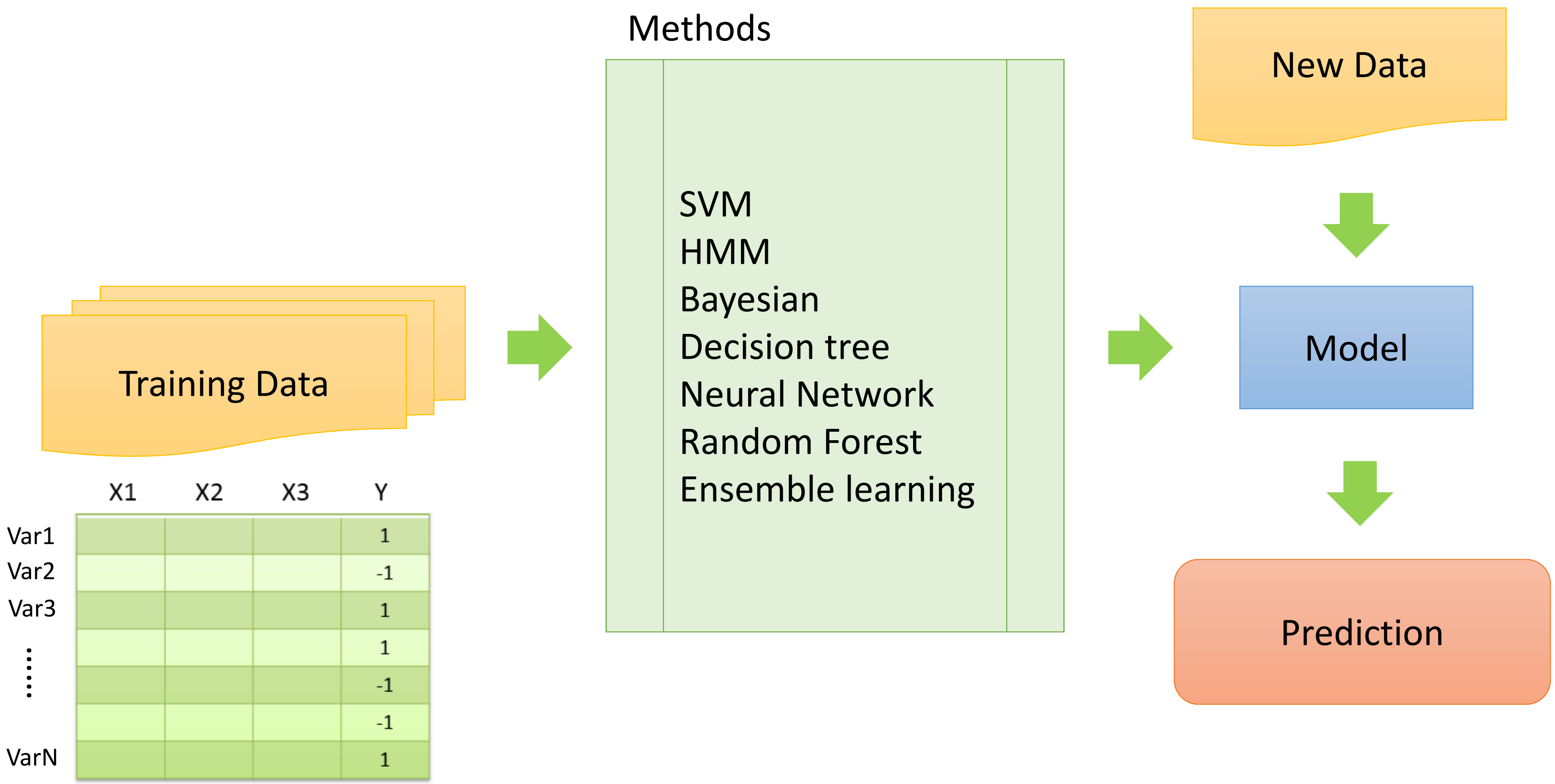








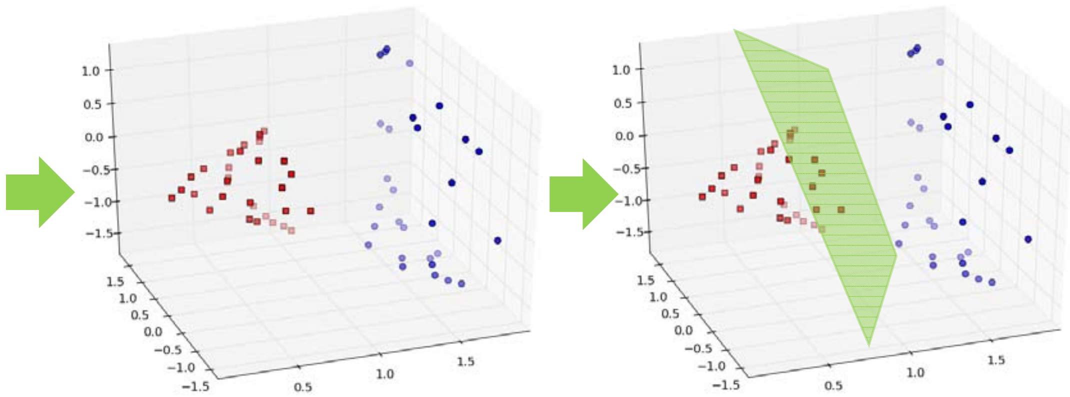
Machine learning model



Classification

- Classifying data is a common task in machine learning. Suppose some given data points each belong to one of two classes, and the goal is to decide which class a new data point will be in.

X1	X2	X3	Y
			1
			-1
			1
			1
			-1
			-1
			1



Introduction

- SVM is supervised learning model that analyze data and recognize patterns, used for **classification** and **regression** analysis.

- It selects a small number of critical boundary instances called support vectors from each class and build a linear discriminant function that separates them as widely as possible.

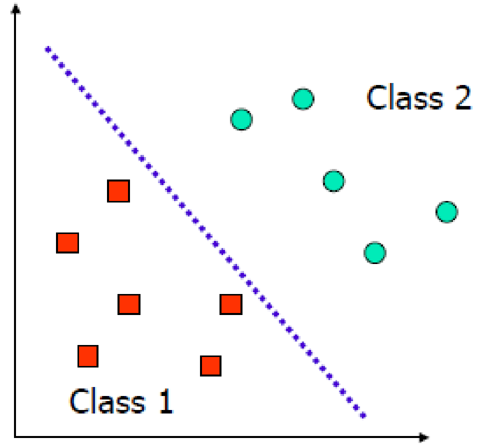
- SVMs can efficiently perform non-linear classification using what is called the **kernel** trick, implicitly mapping their inputs into high-dimensional feature spaces.

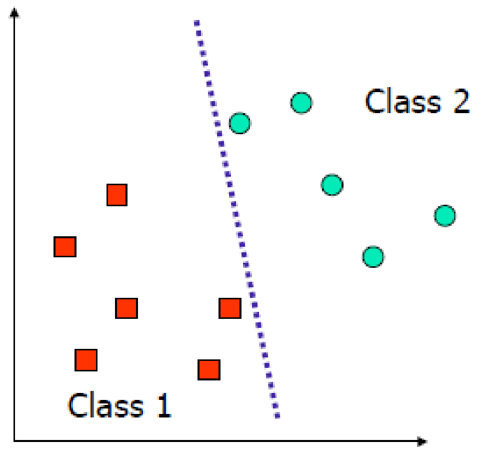
What is a good Decision Boundary?

- Consider a two-class, linearly separable classification problem

- Many decision boundaries!

- Are all decision boundaries equally good?



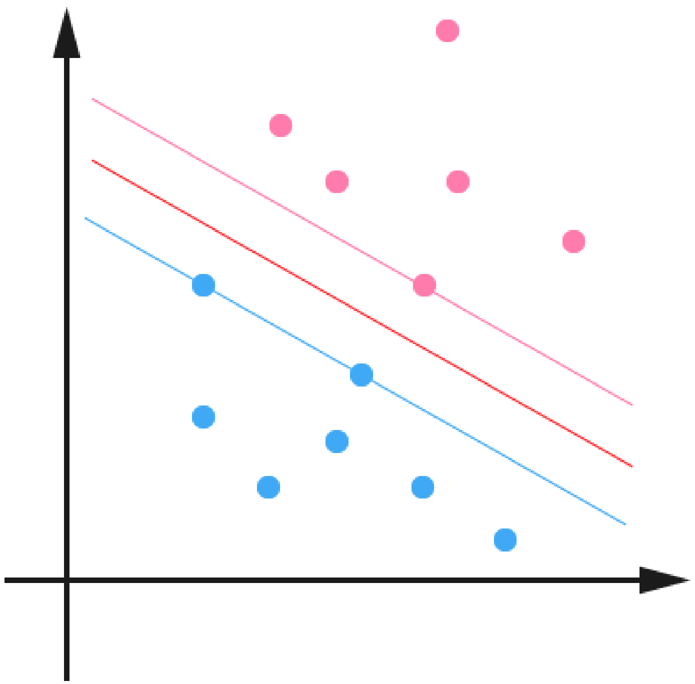


http://blog.csdn.net/v_july_v/article/details/7624837

Decision Boundary

- Intuitively, the best hyperplane is the one that represents the largest separation, or margin, between the two classes,

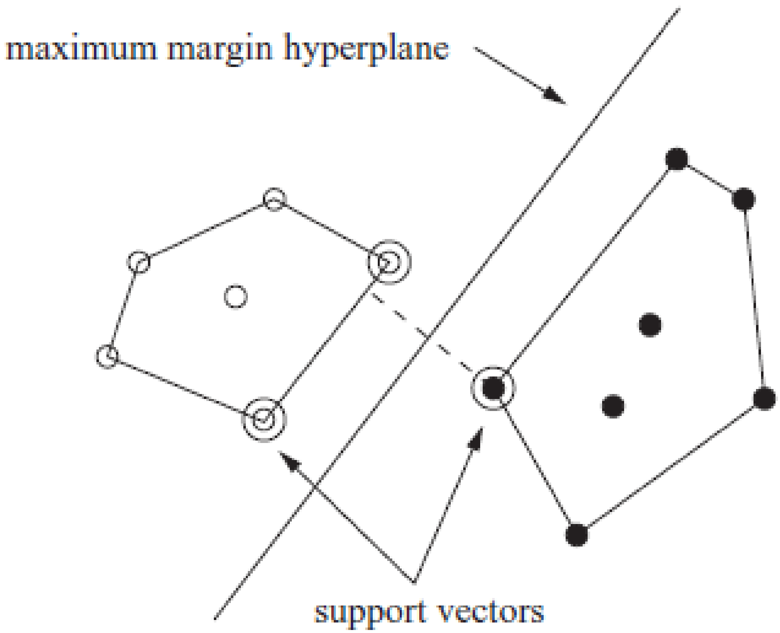
- since the larger the margin is, the lower the generalization error of the classifier will be.



http://blog.csdn.net/v_july_v/article/details/7624837

Support Vector

- The instances that are closest to the maximum-margin hyperplane—the ones with the minimum distance to it—are called **support vectors**.



SVM - mathematics

The data point is donated by x_i, which is a n dimension vector, and y_i is the 1 or -1 to represent the two different class. The hyperplane is

$$w^T x + b = 0$$

So the classification function is

$$f(x) = w^T x + b$$

And

$$y = \begin{cases} 1, & f(x) > 0 \\ -1, & f(x) < 0 \end{cases}$$

SVM - mathematics

The confidence of a classification can be measured by the functional margin, which is $|\mathrm{f}(\mathrm{x})|$, and whether the classification is right can be determined by the consistence of signs of $\mathrm{f}(\mathrm{x_i})$ and $\mathrm{y_i}$. And in fact, $|\mathrm{f}(\mathrm{x})| = \mathrm{y_i} f(x_i)$. So **functional margin** is:

$$\widehat{r_i} = y_i(w^T x_i + b)$$

The functional margin of a hyperplane is measured by

$$\hat{r} = \min \widehat{r_i}$$

However, the **functional margin can be scaled** even if the hyperplane remain the same, for example, w and b changed into 2w and 2b.

SVM - mathematics

A intuitional measurement can be obtained using the distance from the point to the hyperplane, which is called geometrical margin

$$\mathrm{r} = \frac{|\mathrm{f}(\mathrm{x})|}{||w||} = \frac{\hat{r}}{||w||}$$

In this maximum margin classifier, we want to $\max r$. Because the functional margin is scalable, we can assume $\hat{r} = 1$ without influence the optimal result.

SVM - mathematics

So the objective function is

$$\max \frac{1}{||w||} \quad s.t.\ y_i(w^T x_i + b) \geq 1\,, \qquad i = 1,2,\ldots,n.$$

Which equals to

$$\min \frac{1}{2}\left||w|\right|^2 \quad s.t.\ y_i(w^T x_i + b) \geq 1\,, \qquad i = 1,2,\ldots,n.$$

This is a optimization model with constraints, and can be easily solve by ***Quadratic Programming***.

SVM - mathematics

We can also solve this by Lagrange multipliers

$$\mathrm{L}(\mathrm{w}, \mathrm{b}, \alpha) = \frac{1}{2}\left||w|\right|^2 - \sum_{i=1}^{n} \alpha_i [y_i (w^T x_i + b) - 1]$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{w}} = 0 \quad => \quad w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$\frac{\partial \mathrm{L}}{\partial \mathrm{b}} = 0 \quad => \quad \sum_{i=1}^{n} \alpha_i y_i = 0$$

SVM - mathematics

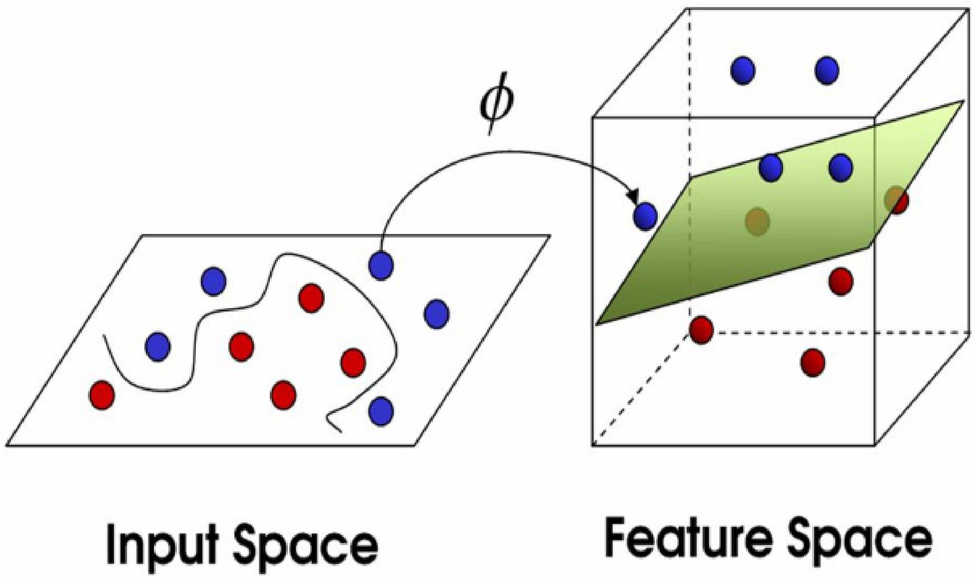
Finally the classification function can be rewritten as

$$\mathrm{f(x)} = \left(\sum_{i=1}^{n} \alpha_i y_i x_i\right)^T x + b$$

$$= \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x \rangle + b$$

SVM - kernel

- The linear learning machine has very limited ability in practice, because of complexity in the real world, which needs more flexible hypothetical space.
- We can use a function ϕ to map x to a higher dimension space, in which all the points can be linear separable.



http://blog.csdn.net/v_july_v/article/details/7624837

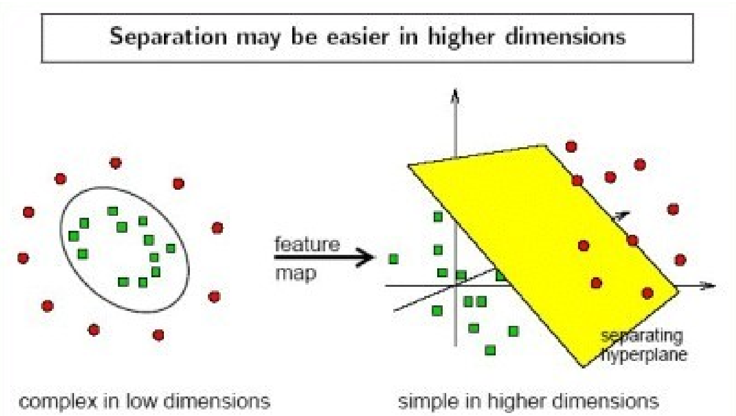
kernel

- So the classification function can be extended as

$$f(x) = \sum_{i=1}^{n} \alpha_i y_i \langle \phi(x_i), \phi(x) \rangle + b$$

- Here we get the kernel function:

$$\mathrm{K}(\mathrm{x}, \mathrm{z}) = \langle \phi(x), \phi(z) \rangle$$



http://blog.csdn.net/v_july_v/article/details/7624837

kernel

- Take points in the picture for example, the two classes can be separated by a circle

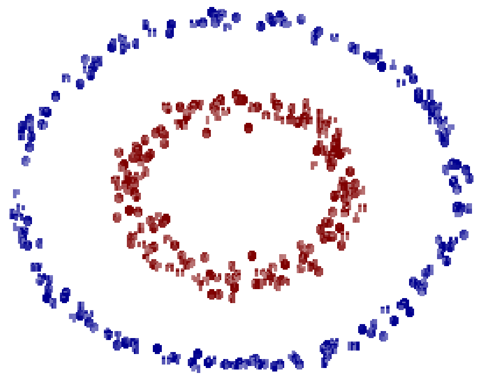
$$a_1X_1 + a_2X_1^2 + a_3X_2 + a_4X_2^2 + a_5x_1x_2 + a_6 = 0$$

- The we can construct a 5-dimension space, where

$$Z_1 = x_1, \qquad Z_2 = x_1^2, \qquad Z_3 = x_2, \qquad Z_4 = x_2^2, \qquad Z_5 = x_1x_2$$

- So the hyperplane in the new feather space is

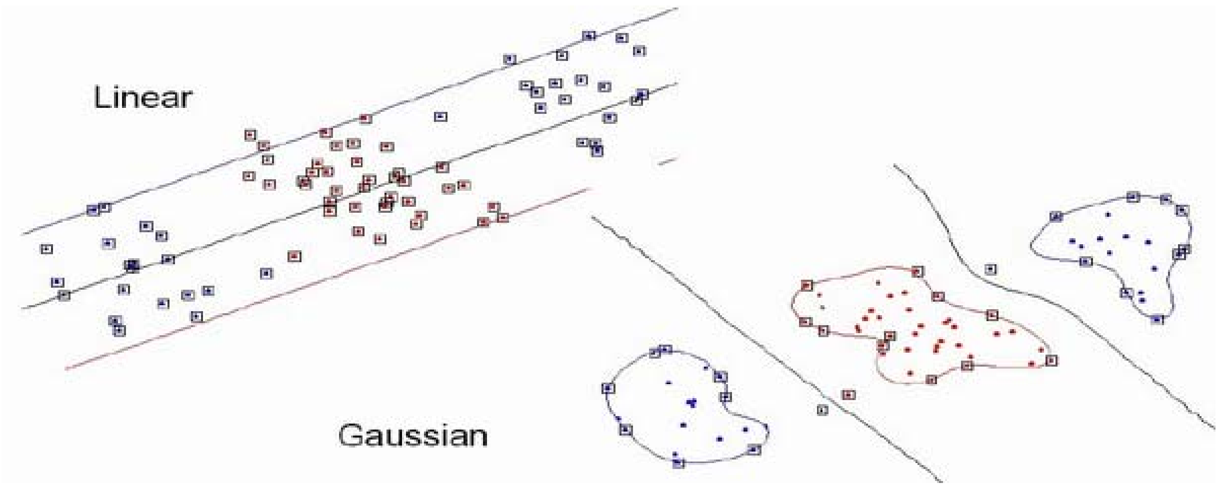
$$\sum_{i=1}^{5} a_iZ_i + a_6 = 0$$



http://blog.csdn.net/v_july_v/article/details/7624837

Kernel function

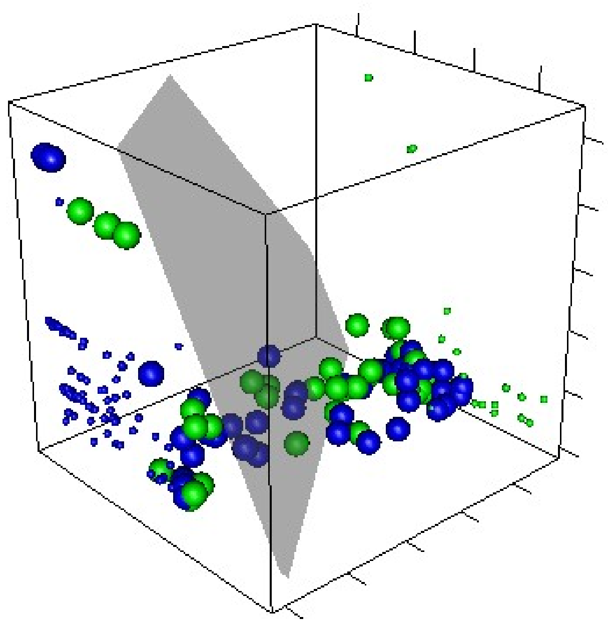
- **Linear kernel:** $\mathrm{K}(\mathrm{x}_1, \mathrm{x}_2) = \langle x_1, x_2 \rangle$
- **Polynomial kernel**: $\mathrm{K}(\mathrm{x}_1, \mathrm{x}_2) = (\langle x_1, x_2 \rangle + d)^n$
- **Gauss kernel**: $\mathrm{K}(\mathrm{x}_1, x_2) = e^{-\frac{||x_1 - x_2||^2}{2\sigma^2}}$



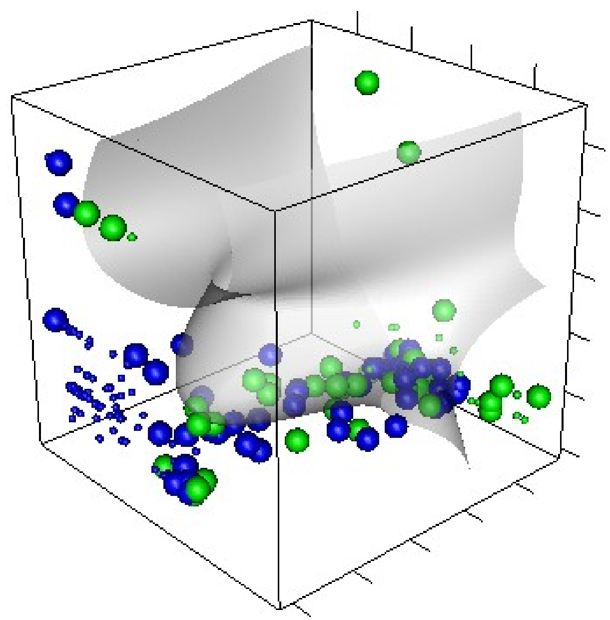
http://blog.csdn.net/v_july_v/article/details/7624837

SVM - example

- Linear kernel



- Gauss kernel



Applications

SVM has been used successfully in many real-world problems

- bioinformatics (Mutation classification,   Cancer classification)
- text (and hypertext) categorization
- image classification – different types of sub-problems
- hand-written character recognition

Pros and Cons

- With support vectors, the maximum-margin hyperplane is relatively stable.
- However, they often produce very accurate classifiers because subtle and complex decision boundaries can be obtained.
- Compared with other methods, even the fastest training algorithms for support vector machines are slow when applied in the nonlinear setting.

生物信息学：导论与方法

Bioinformatics: Introduction and Methods