机器学习及其MATLAB实现—从基础到实践 第6课

DATAGURU专业数据分析社区

法律声明

【声明】本视频和幻灯片为炼数成金网络课程的教学资料，所有资料只能在课程内使用，不得在课程以外范围散播，违者将可能被追究法律和经济责任。

课程详情访问炼数成金培训网站

http://edu.dataguru.cn

课程目录

支持向量机分类原理概述

- **Support vector machines** (**SVMs**) are a set of related **supervised learning** methods that analyze data and recognize patterns, used for classification and regression analysis.
- The original SVM algorithm was invented by **Vladimir Vapnik** and the current standard incarnation (soft margin) was proposed by Corinna Cortes and Vladimir Vapnik.
- A support vector machine constructs a **hyperplane** or set of hyperplanes in a high or infinite dimensional space, which can be used for classification, regression, or other tasks.
- a good separation is achieved by the hyperplane that has the largest distance to the nearest training data points of any class, since in general the larger the margin the lower the generalization error of the classifier.

支持向量机分类原理概述

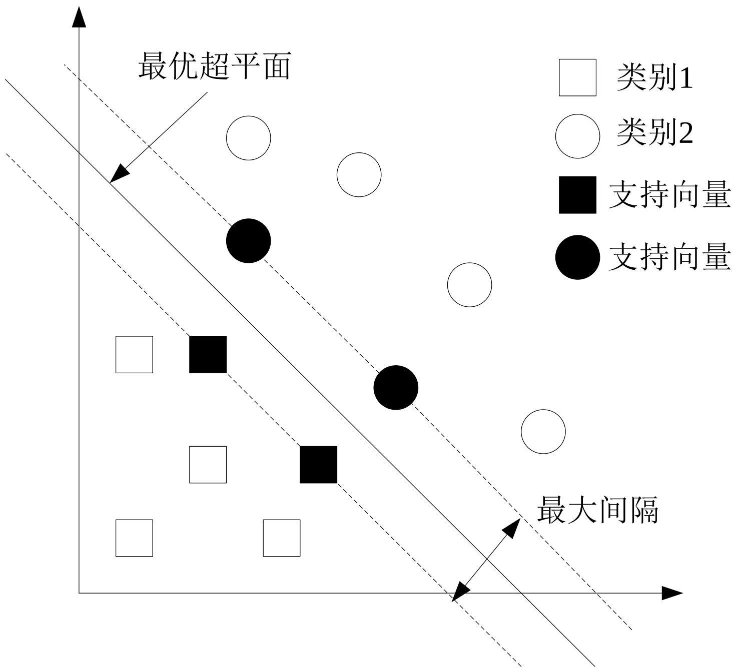
- We want to find the **maximum-margin hyperplane** that divides the points having $y_i = 1$ from those having $y_i = -1$. Any hyperplane can be written as the set of points satisfying

$$w \cdot x + b = 0$$

- We want to choose the w and b to maximize the margin, or distance between the parallel hyperplanes that are as far apart as possible **while still separating the data.**

$$\min \quad \frac{\|w\|^2}{2}$$

$$s.t \quad y_i(w \cdot x_i + b) \geq 1, i = 1, 2, \ldots, l$$



支持向量机分类原理概述

- **Primal form**

$$\min \frac{\|w\|^2}{2}$$

$$s.t\ \ y_i(w\cdot x_i+b)\ge 1, i=1,2,\ldots,l$$

- **Dual form**

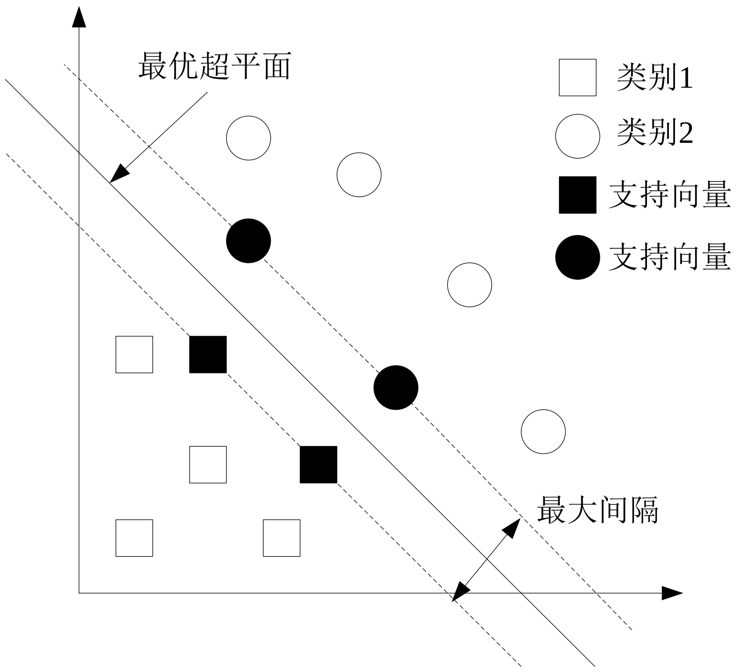
$$\max\ \ Q(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j(x_i\cdot x_j)$$

$$s.t\ \ \sum_{i=1}^{l}\alpha_i y_i=0, \alpha_i\ge 0$$

- **Decision function**

$$f(x)=\mathrm{sgn}\left(\sum_{i=1}^{l}\alpha_i^* y_i(x\cdot x_i)+b^*\right)$$

$$w^*=\sum_{i=1}^{l}\alpha_i^* x_i y_i \qquad b^*=-\frac{1}{2}w^*\cdot(x_r+x_s)$$



支持向量机分类原理概述

- **Primal form**

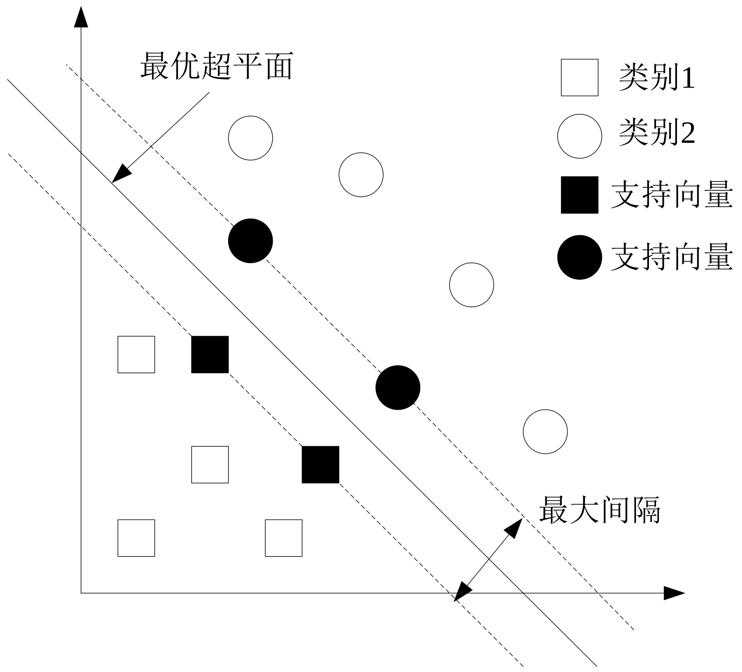
$$\min \quad \frac{\|w\|^2}{2} + C\sum_{i=1}^{l}\xi_i$$

$$s.t \quad \begin{cases} y_i(w\cdot x_i + b) \geq 1 - \xi_i \\ \xi_i > 0 \end{cases} \quad i = 1, 2, \ldots, l$$

- **Dual form**

$$\max \ \mathrm{Q}(\alpha) = \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i \cdot x_j)$$

$$s.t \quad \begin{cases} \sum_{i=1}^{l}\alpha_i y_i = 0 \\ 0 \leq \alpha_i \leq C \end{cases} \quad i = 1, 2, \ldots, l$$



支持向量机分类原理概述

- **Dual form**

$$K(x_i,x_j)=\Phi(x_i)\cdot\Phi(x_j)$$

$$\max\ Q(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j)$$

$$s.t\ \begin{cases}\sum_{i=1}^{l}\alpha_i y_i=0 \\ 0\le\alpha_i\le C\end{cases},i=1,2,\dots,l$$

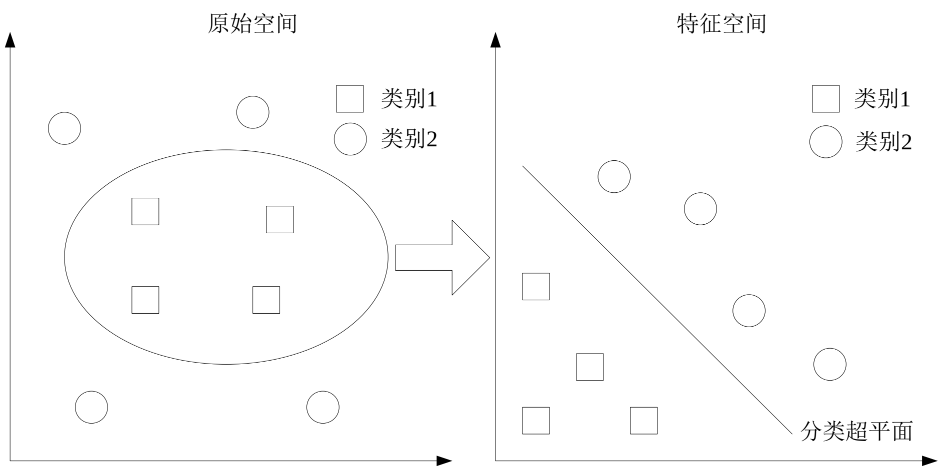
- **Decision function**

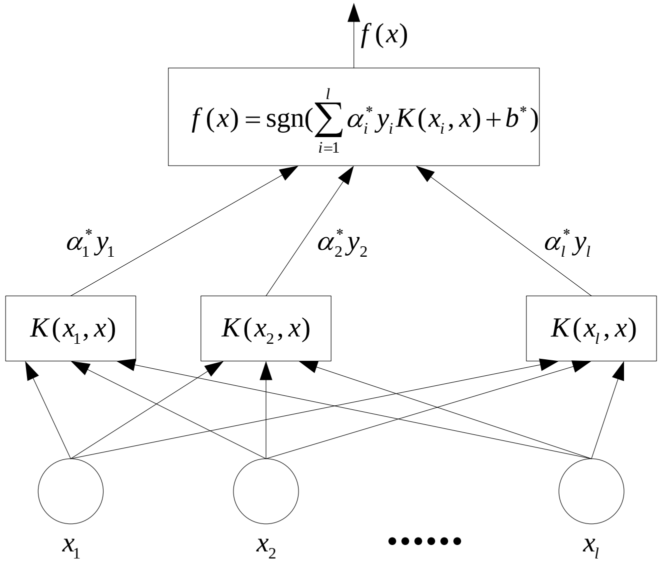
$$f(x)=\mathrm{sgn}(w^*\cdot\Phi(x)+b^*)$$

$$=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^* y_i\Phi(x_i)\cdot\Phi(x)+b^*)$$

$$=\mathrm{sgn}(\sum_{i=1}^{l}\alpha_i^* y_i K(x_i,x)+b^*)$$

$$w^*=\sum_{i=1}^{l}\alpha_i^* y_i\Phi(x_i)$$



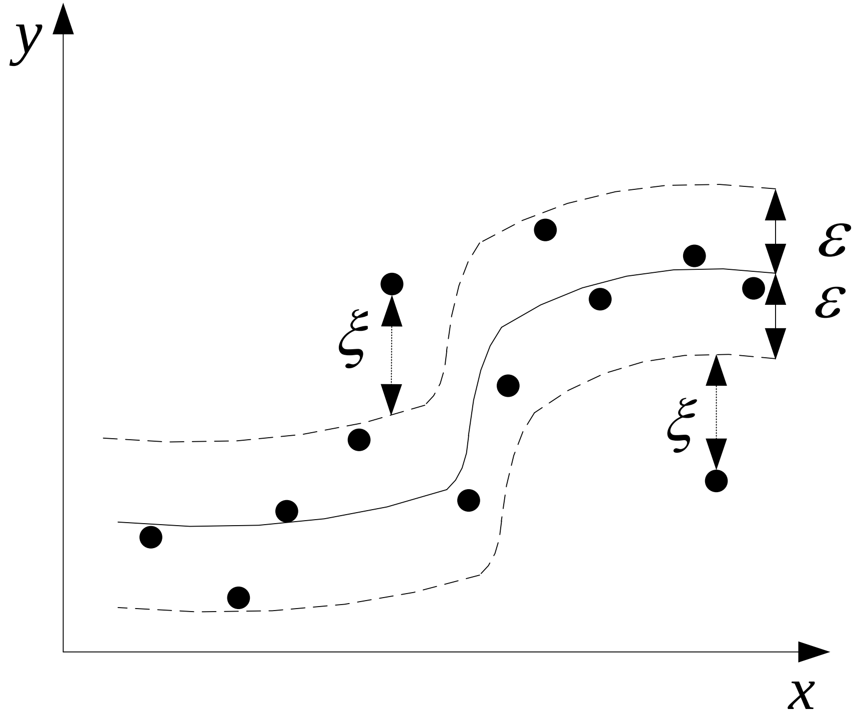


支持向量机回归原理概述

- 为了利用SVM解决回归拟合方面的问题，Vapnik等人在SVM分类的基础上引入了 **不敏感损失函数**，从而得到了回归型支持向量机（Support Vector Machine for Regression，SVR）。
- SVM应用于回归拟合分析时，其基本思想不再是寻找一个最优分类面使得两类样本分开，而是寻找一个最优分类面**使得所有训练样本离该最优分类面的误差最小**。

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

$$s.t \begin{cases} y_i - w\cdot\Phi(x_i) - b \le \varepsilon + \xi_i \\ -y_i + w\cdot\Phi(x_i) + b \le \varepsilon + \xi_i^* \quad i = 1, 2, \dots, l \\ \xi_i \ge 0, \xi_i^* \ge 0 \end{cases}$$



支持向量机回归原理概述

- **Primal form**

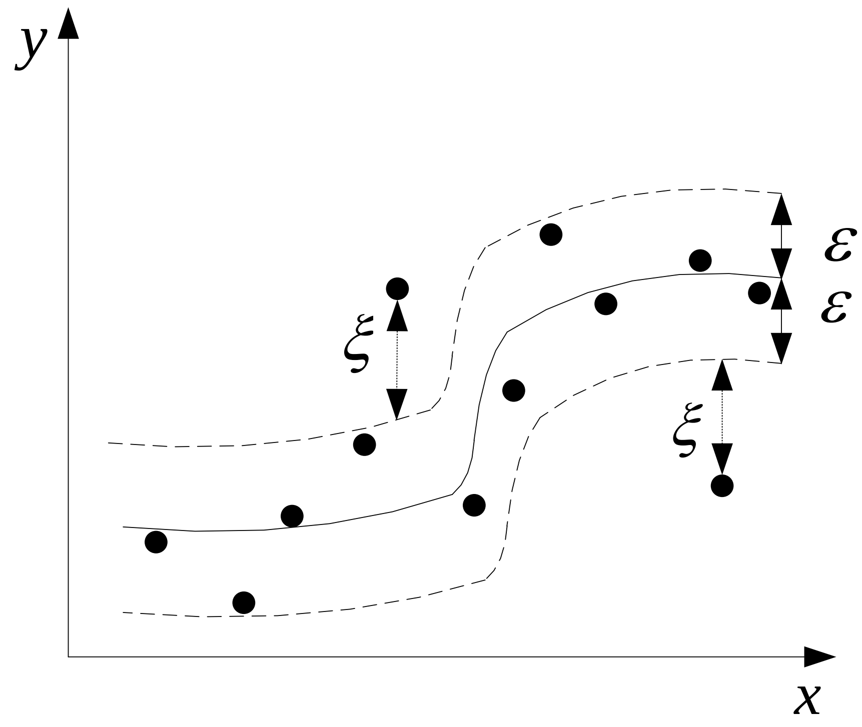
$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*)$$

$$s.t \begin{cases} y_i - w\cdot\Phi(x_i) - b \le \varepsilon + \xi_i \\ -y_i + w\cdot\Phi(x_i) + b \le \varepsilon + \xi_i^* \quad i = 1,2,\dots,l \\ \xi_i \ge 0, \xi_i^* \ge 0 \end{cases}$$

- **Dual form**

$$\max_{\alpha,\alpha^*}\left[-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(x_i,x_j) - \sum_{i=1}^{l}(\alpha_i+\alpha_i^*)\varepsilon + \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)y_i\right]$$

$$s.t \begin{cases} \sum_{i=1}^{l}(\alpha_i-\alpha_i^*) = 0 \\ 0 \le \alpha_i \le C \\ 0 \le \alpha_i^* \le C \end{cases}$$



支持向量机回归原理概述

- **Decision function**

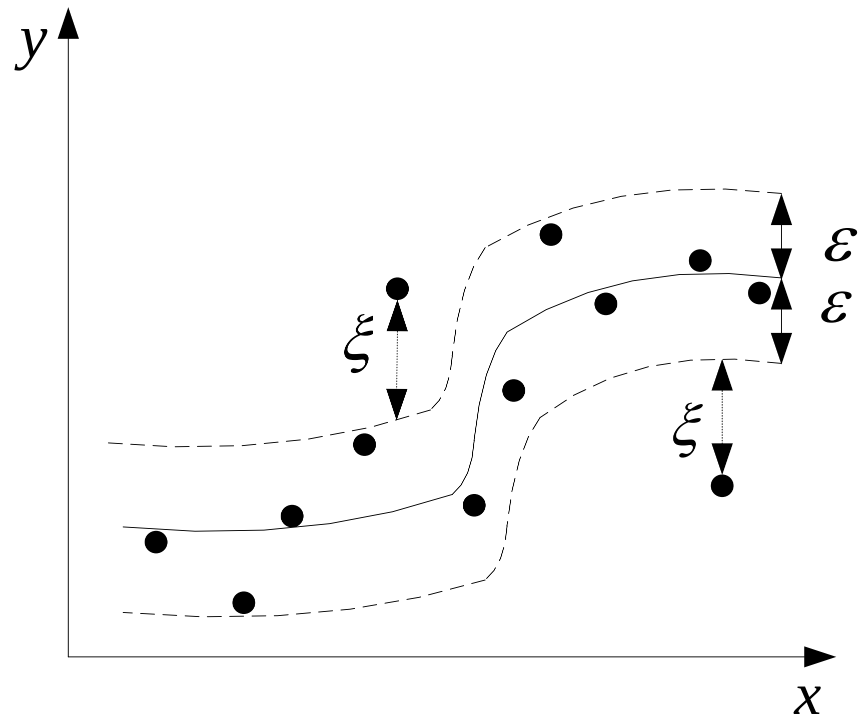
$$f(x)=w^*\cdot\Phi(x)+b^*=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\Phi(x_i)\cdot\Phi(x)+b^*$$

$$=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(x_i,x)+b^*$$

$$w^*=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\Phi(x_i)$$

$$b^*=\frac{1}{N_{nsv}}\sum_{0<\alpha_i<C}\left[y_i-\sum_{x_i\in SV}(\alpha_i-\alpha_i^*)K(x_i,x_j)-\varepsilon\right]$$

$$+\frac{1}{N_{nsv}}\sum_{0<\alpha_i<C}\left[y_i-\sum_{x_j\in SV}(\alpha_j-\alpha_j^*)K(x_i,x_j)+\varepsilon\right]$$



多分类SVM及QP求解概述

- **Multiclass SVM** aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements.
- The dominating approach for doing so is to **reduce** the single multiclass problem into multiple **binary classification** problems.
- **one-against-all**
  - Classification of new instances for one-against-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class.
- **one-against-one**
  - For the one-against-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with most votes determines the instance classification.

多分类SVM及QP求解概述

- 分块算法（Chunking）
- Osuna算法
- 序列最小优化算法（Sequential Minimal Optimization，SMO）
- 增量学习算法（Incremental Learning）

$$\max\ Q(\alpha)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j)$$

$$s.t\ \begin{cases}\sum_{i=1}^{l}\alpha_i y_i=0 \\ 0\le\alpha_i\le C\end{cases},i=1,2,\ldots,l$$

Repeat till convergence {

1. Select some pair α_i and α_j to update next (using a heuristic that tries to pick the two that will allow us to make the biggest progress towards the global maximum).
2. Reoptimize $W(\alpha)$ with respect to α_i and α_j, while holding all the other α_k's $(k \neq i,j)$ fixed.

}

Libsvm工具箱概述

- **LIBSVM -- A Library for Support Vector Machines**
- **Install Tips**
  - Current Directory
  - mex -setup
  - make
  - compiler (VC 6.0)

重点函数解读

- **meshgrid**
  - Generate X and Y arrays for 3-D plots
  - [X,Y] = meshgrid(x,y)
  -
- **svmtrain**
  - Train support vector machine classifier
  - model = svmtrain(train_label,train_matrix,’ libsvm_options’);
- **svmpredict**
  - Predict data using support vector machine
  - [predict_label,accuracy] = svmpredict(test_label,test_matrix,model);

案例分析

乳腺癌诊断

混凝土抗压强度预测

炼数成金逆向收费式网络课程

- Dataguru（炼数成金）是专业数据分析网站，提供教育，媒体，内容，社区，出版，数据分析业务等服务。我们的课程采用新兴的互联网教育形式，独创地发展了逆向收费式网络培训课程模式。既继承传统教育重学习氛围，重竞争压力的特点，同时又发挥互联网的威力打破时空限制，把天南地北志同道合的朋友组织在一起交流学习，使到原先孤立的学习个体组合成有组织的探索力量。并且把原先动辄成千上万的学习成本，直线下降至百元范围，造福大众。我们的目标是：低成本传播高价值知识，构架中国第一的网上知识流转阵地。
- 关于逆向收费式网络的详情，请看我们的培训网站 http://edu.dataguru.cn

Thanks

FAQ时间