机器学习导论

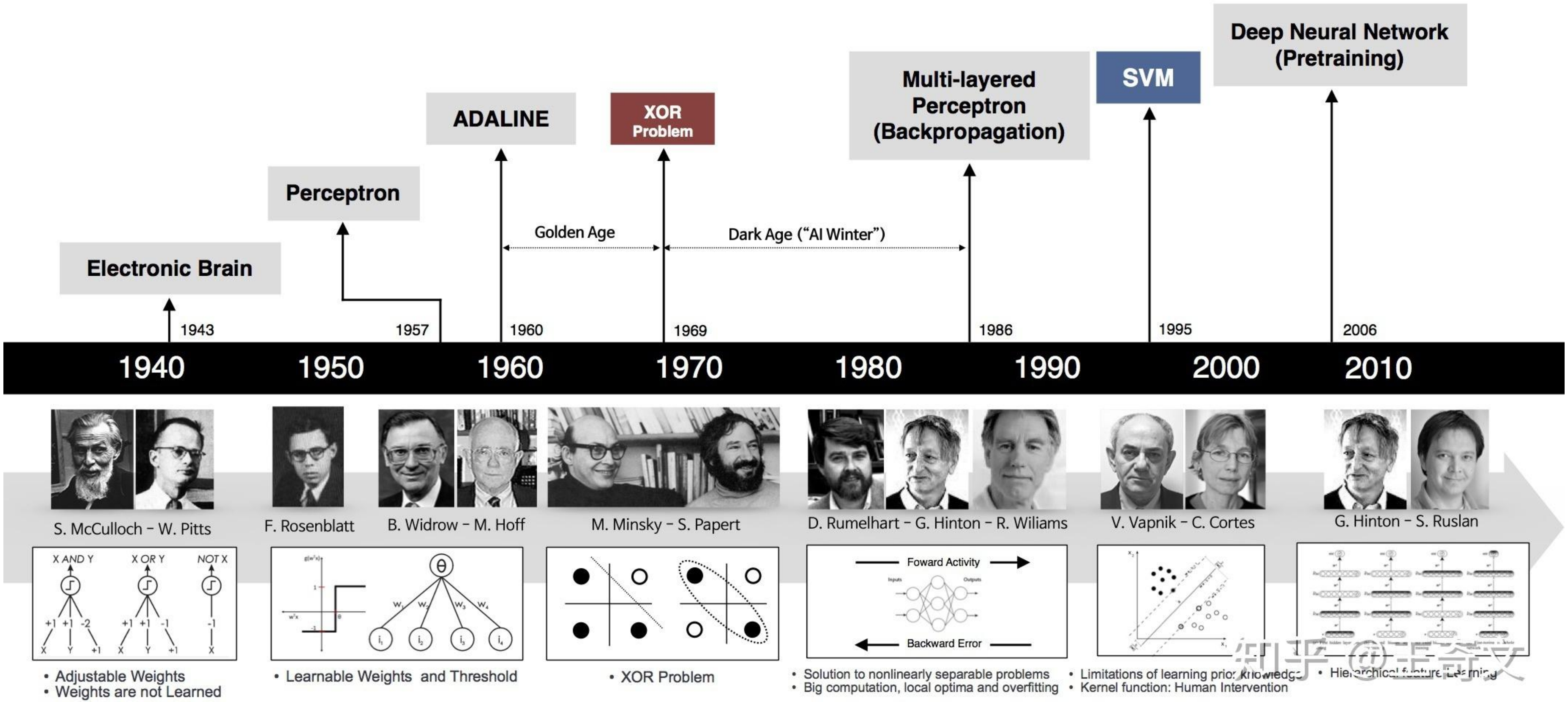
第4章 支持向量机

谢茂强

南开大学软件学院

本章主要内容来自周志华《机器学习》和吴恩达的Coursera网课

SVM在两轮神经网络间的辉煌



SVM历史

- 从90年代后期至2015年前后深度学习兴起之前，SVM被认为是机器学习当时最成功、表现最好的方法。

休息一会儿

小故事：统计学习理论之父弗拉基米尔•瓦普尼克

弗拉基米尔 • 瓦普尼克 (Vladimir N. Vapnik, 1936—) 是杰出的数学家、统计学家、计算机科学家. 他出生于苏联, 1958 年在乌兹别克国立大学获数学硕士学位, 1964 年在莫斯科控制科学学院获统计学博士学位, 此后一直在该校工作并担任计算机系主任. 1990 年(苏联解体的前一年)他离开苏联来到新泽西州的美国电话电报公司贝尔实验室工作, 1995 年发表了最初的 SVM 文章. 当时神经网络正当红, 因此这篇文章被权威期刊 *Machine Learning* 要求以“支持向量网络”的名义发表.

实际上, 瓦普尼克在 1963 年就已提出了支持向量的概念, 1968 年他与另一位苏联数学家 A. Chervonenkis 提出了以他们两人的姓氏命名的“VC 维”, 1974 年又提出了结构风险最小化原则, 使得统计学习理论在二十世纪七十年代就已成型. 但这些工作主要是以俄文发表的, 直到瓦普尼克随着东欧剧变和苏联解体导致的苏联科学家移民潮来到美国, 这方面的研究才在西方学术界引起重视, 统计学习理论、支持向量机、核方法在二十世纪末大红大紫.

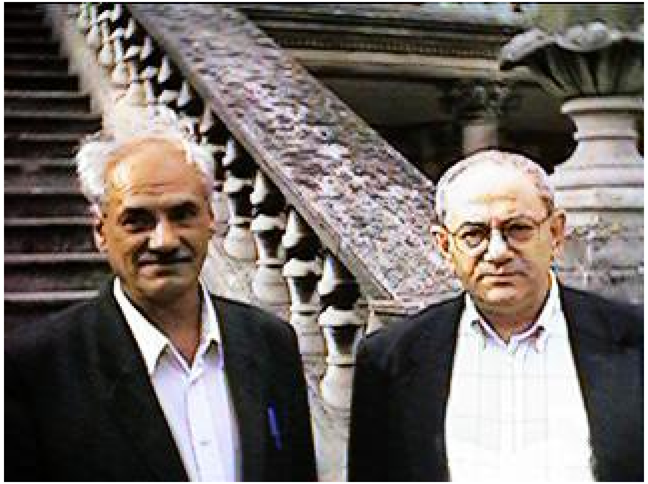
瓦普尼克 2002 年离开美国电话电报公司加入普林斯顿的 NEC 实验室, 2014 年加盟脸书(Facebook)公司人工智能实验室. 1995 年之后他还在伦敦大学、哥伦比亚大学等校任教授. 据说瓦普尼克在苏联根据一本字典自学了英语及其发音. 他有一句名言被广为传诵: “*Nothing is more practical than a good theory.*”

SVM 的确与神经网络有密切联系: 若将隐层神经元数设置为训练样本数, 且每个训练样本对应一个神经元中心, 则以高斯径向基函数为激活函数的 RBF 网络(参见 5.5.1 节)恰与高斯核 SVM 的预测函数相同.

教材P145

History

- 从90年代后期至2010年前后深度学习兴起之前，SVM被认为是机器学习当时最成功、表现最好的方法。
- Vapnik(苏联人)于1963年提出支持向量概念
- Vapnik 和 Chervonenkis于1968年提出VC维，是统计学习理论的基础。
- 目前普遍使用的SVM(支持向量机) 由 Cortes 和 Vapnik于1993年提出，并在1995年发表。



Vladimir N. Vapnik(右) & Alexey Ya. Chervonenkis(左

Vapnik, V.N. and Lerner, A.Y.. Recognition of patterns with help of generalized portraits. Automation and Remote Control. 24(6), pp.774-780. 1963.
Vapnik, V.N. and A. Y. Chervonenkis. The necessary and sufficient conditions for consistency in the empirical risk minimization method. Pattern Recognition and Image Analysis. 1 (3), 283-305. 1991.
Cortes, C. and Vapnik, V.. Support vector networks. Machine learning, 20(3), pp.273-297. 1995.

目录

1.1 线性可分问题 & 最大化间隔策略

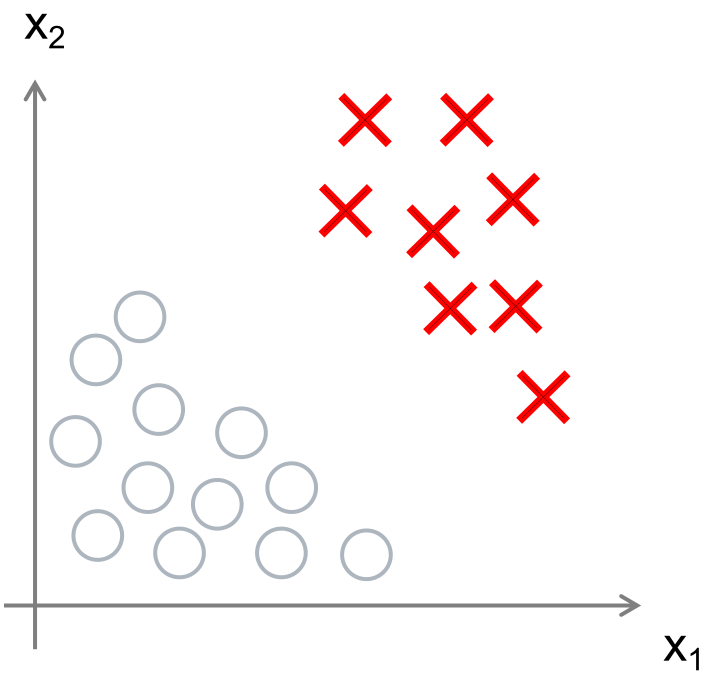
$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$$

$$s.t. \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geqslant 1,$$

$$i = 1, 2, \cdots, m$$

注意：$y_i \in \{-1, 1\}$

支持向量：满足约束条件，距离超平面最近的训练样本



1.1最大化间隔与结构风险

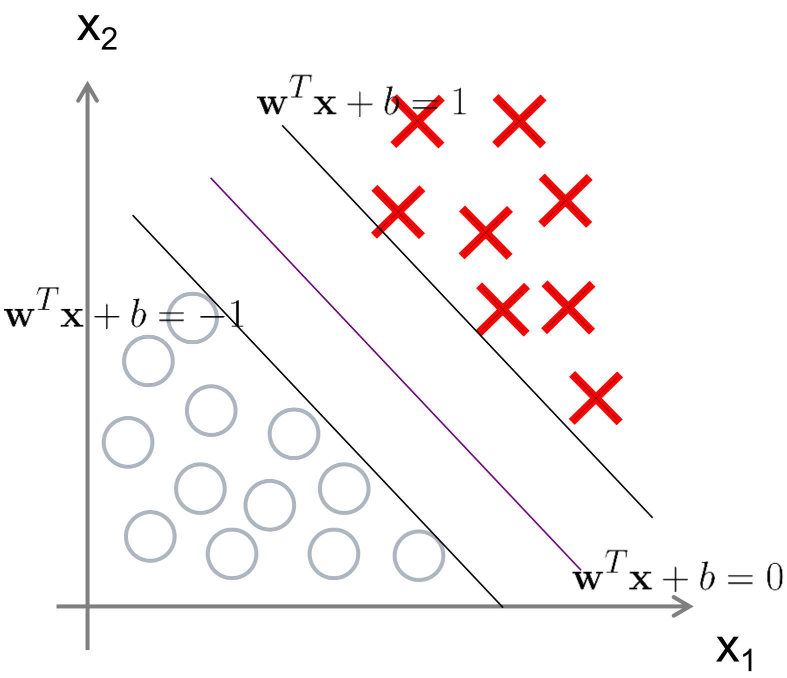
$$\frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_{j=1}^{n}\mathbf{w}^2$$

$\dfrac{2}{\|\mathbf{w}\|}$ ： 两类样本间的margin

点到直线距离： $\dfrac{|\mathbf{w}^T\mathbf{x}+b|}{\|\mathbf{w}\|}$

所以， $\mathbf{w}^T\mathbf{x}+b=1$ 上的点到

$\mathbf{w}^T\mathbf{x}+b=0$ 是 $\dfrac{1}{\|\mathbf{w}\|}$ 。



1.1 上述最优化问题的拉格朗日函数

$$\max_{\mathbf{w},b} \frac{2}{\|\mathbf{w}\|}$$

$$s.t. \quad y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b) \geqslant 1,$$

$$i = 1, 2, \cdots, m$$

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$$

$$\sum_{i=1}^{m} \alpha_i(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$$

经验风险：度量预测模型与训练数据的误差

$$\frac{1}{2}\|\mathbf{w}\|^2 = \frac{1}{2}\sum_{j=1}^{n} \mathbf{w}^2$$

结构风险：描述模型的某些性质，以避免过拟合。(Maximizing the margin)

周志华《机器学习》附录 B.1 拉格朗日乘子法

1.1 代价函数中的经验风险

$$\sum_{i=1}^{m} \alpha_i (1 - y_i(\mathbf{w}^T \mathbf{x}_i + b))$$

而 $\mathbf{w}^T \mathbf{x}_i + b$ 应该 ≥ 1 或 ≤ -1

最理想的是 y_i 和 $\mathbf{w}^T \mathbf{x}_i + b$ 输出相同，损失为0，

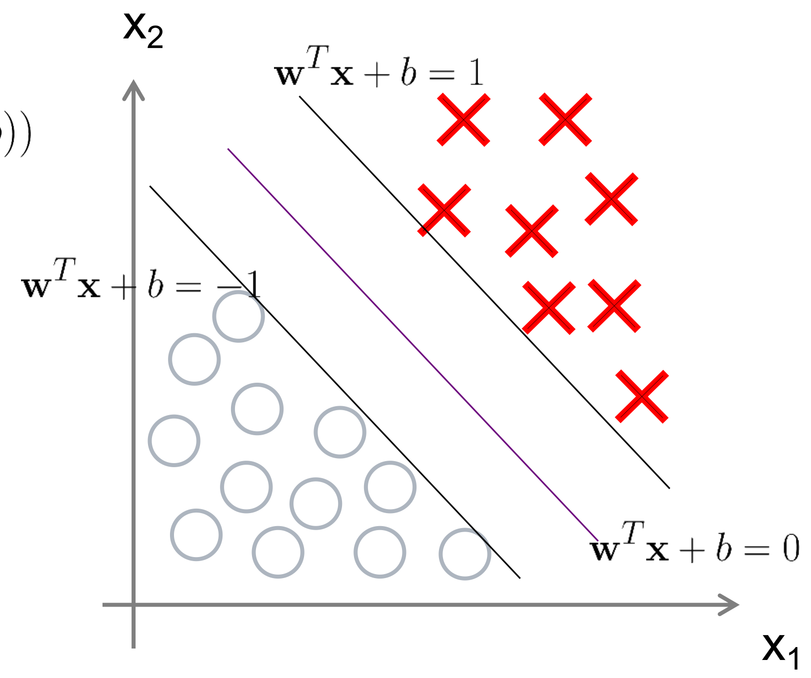
1.2 SVM的判决函数(之一)

当使用训练数据集优化代价函数

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\mathbf{w}^T \mathbf{x}_i + b))$$

得到模型参数 $\mathbf{w}, b$ 后，判决函数为：

$$f(\mathbf{x}_i; \mathbf{w}) = \begin{cases} 1, & \mathbf{w}^T \mathbf{x}_i + b \geq 1 \\ -1, & \mathbf{w}^T \mathbf{x}_i + b \leq -1 \end{cases}$$



目录

1.3 SVM的优化

最小化： $$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\mathbf{w}\|^2 + \sum_{i=1}^{m} \alpha_i(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$$

可以通过优化其对偶形式（下式）加以求解：

$$max_\alpha\{min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)\}$$

（1）先通过求导法，求 $\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$

（2）然后再求 $\min_{\mathbf{w},b} L(\mathbf{w}, b, \alpha)$ 对 α 的极大。

具体的证明请见 李航《统计学习方法》.附录C. 拉格朗日对偶性.

1.3.1通过求导法，给定α求 $\min_{\mathbf{w},b} L(\mathbf{w},b,\alpha)$

给定α最小化：$L(\mathbf{w},b,\alpha)=\frac{1}{2}\|\mathbf{w}\|^2+\sum_{i=1}^{m}\alpha_i(1-y_i(\mathbf{w}^T\mathbf{x}_i+b))$

- 分别对**w**和b偏导等于0

$$\frac{\partial}{\partial \boldsymbol{w}}L(\boldsymbol{w},b,\alpha)=\boldsymbol{w}-\sum_{i=1}^{m}\alpha_i y_i \boldsymbol{x}_i=0 \quad 得 \quad \mathbf{w}=\sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i$$

$$\frac{\partial}{\partial b}L(\boldsymbol{w},b,\alpha)=-\sum_{i=1}^{m}\alpha_i y_i=0 \quad 得 \quad 0=\sum_{i=1}^{m}\alpha_i y_i$$

- 将 $\mathbf{w}=\sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i$ 带入最上式可得

$$L(\boldsymbol{w},b,\alpha)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j-\sum_{i=1}^{m}\alpha_i y_i\left(\left(\sum_{j=1}^{m}\alpha_j y_j \boldsymbol{x}_j\right)^T \boldsymbol{x}_i+b\right)+\sum_{i=1}^{m}\alpha_i$$

给定α最小化： $$L(\mathbf{w},b,\alpha)=\frac{1}{2}\|\mathbf{w}\|^2+\sum_{i=1}^{m}\alpha_i(1-y_i(\mathbf{w}^T\mathbf{x}_i+b))$$

$$L(\boldsymbol{w},b,\alpha)=\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j-\sum_{i=1}^{m}\alpha_i y_i\left(\left(\sum_{j=1}^{m}\alpha_j y_j \boldsymbol{x}_j\right)^T\boldsymbol{x}_i+b\right)+\sum_{i=1}^{m}\alpha_i$$

$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

（1）先通过求导法，求 $\min\limits_{\mathbf{w},b} L(\mathbf{w},b,\alpha)$

$$\min_{\boldsymbol{w},b} L(\boldsymbol{w},b,\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

1.3.2 然后再求 $\min_{\mathbf{w},b} L(\mathbf{w},b,\alpha)$ 对 α 的极大。

$$\max_{\alpha}\left(\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j\right)$$

$$s.t. \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$\alpha_i \geq 0, \quad i=1,2,\cdots,m$$

关于α_i 的二次规划问题

1）使用通用的二次规划(Quadratic Programming)求解

2) 使用专用的SMO(Sequential Minimal Optimization)

Platt, J. Sequential minimal optimization: A fast algorithm for training support vector machines. 1998.

求解关于αi 的二次规划问题所基于的定理(不要求)

定理C.3 对原始问题（C.1）～（C.3）和对偶问题（C.12）～（C.13），假设函数f(x)和$c_i(x)$是凸函数，$h_j(x)$是仿射函数，并且不等式约束$c_i(x)$是严格可行的，则x^*和a^*,β^*分别是原始问题和对偶问题的解的充分必要条件是x^*,a^*,β^*满足下面的Karush-Kuhn-Tucker(KKT)条件：

$$\nabla_x L(x^*,\alpha^*,\beta^*)=0$$

$$\nabla_\alpha L(x^*,\alpha^*,\beta^*)=0$$

$$\nabla_\beta L(x^*,\alpha^*,\beta^*)=0$$

$$\alpha_i^* c_i(x^*)=0,\quad i=1,2,\cdots,k$$

$$c_i(x^*)\le 0,\quad i=1,2,\cdots,k$$

$$\alpha_i^*\ge 0,\quad i=1,2,\cdots,k$$

$$h_j(x^*)=0,\quad i=1,2,\cdots,k$$

由 $\mathbf{w}=\sum_{i=1}^{m}\alpha_i y_i \mathbf{x}_i$ 和 $\boldsymbol{w}^*>0$

可知存在一个 $\alpha_j^*>0$

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

得： $y_j(\boldsymbol{w}^{*T}\boldsymbol{x}_j+b^*)-1=0$

或： $y_j(\boldsymbol{w}^{*T}\boldsymbol{x}_j+b^*)=y_j^2$

李航《统计学习理论》定理C.3

用于求解关于α的二次规划问题的KKT条件

KKT(Karush-Kuhn-Tucker)条件是对应二次规划问题的充分必要条件。

$$\begin{cases} \alpha_i \geqslant 0; \\ y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b) - 1 \geqslant 0; \\ \alpha_i(y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b) - 1) = 0. \end{cases}$$

若$\alpha_i>0$, 则$y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x} + b)=1$,对应样本位于最大间隔边界上，是一个支持向量。

1.3.3 根据(1)和(2)的结果求解 w^* 和 b^*

可得：

$$\boldsymbol{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \boldsymbol{x}_i$$

$$b^* = y_j - \sum_{i=1}^{m} \alpha_i^* y_i \boldsymbol{x}_i^T \boldsymbol{x}_j$$

注：选择一个 $\alpha_j^* > 0$ 对应的 $(\boldsymbol{x}_j, y_j)$ 来计算 b^*

最后得到的判决函数

$$
\begin{aligned}
f(\mathbf{x}) &= sign(\mathbf{w}^T\mathbf{x} + b) \\
&= sign(\sum_{i=1}^{m} \alpha_i y_i \mathbf{x}_i \mathbf{x} + b)
\end{aligned}
$$

2. 软间隔SVM

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2$$

$$s.t. \quad y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b) \geqslant 1,$$

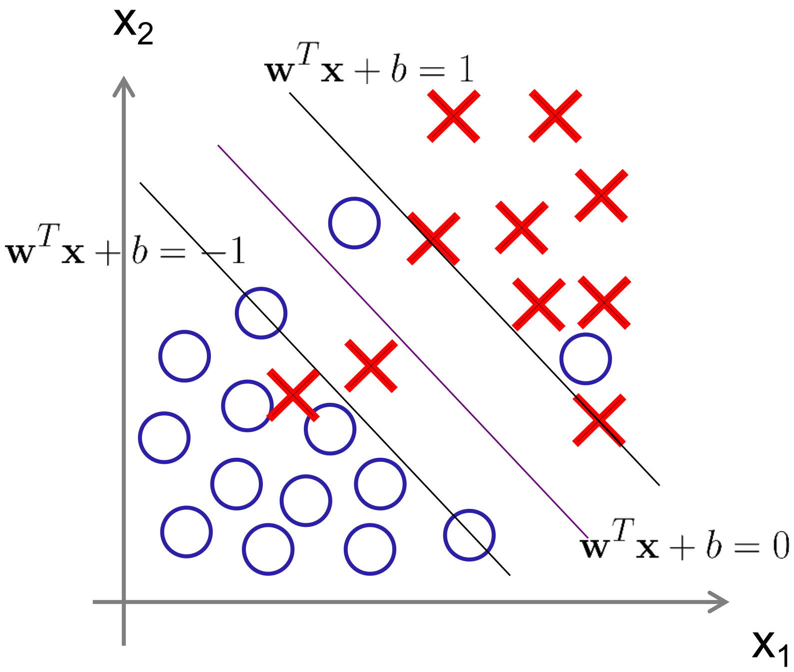
$$i = 1, 2, \cdots, m$$

$$\Downarrow$$

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$s.t. \quad y_i(\mathbf{w}^{\mathbf{T}}\mathbf{x}_i + b) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0, i = 1, 2, \cdots, m$$



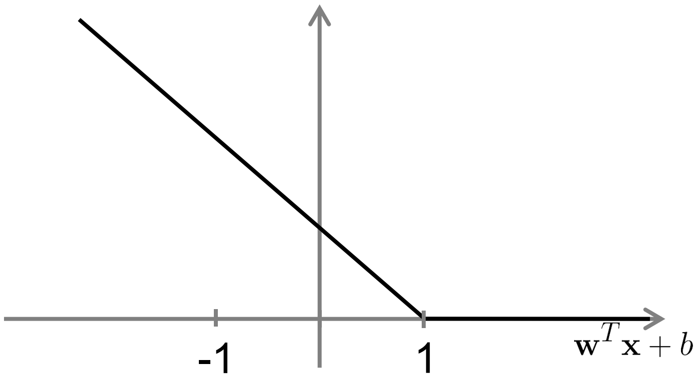
2. 用于软间隔SVM的Hinge损失

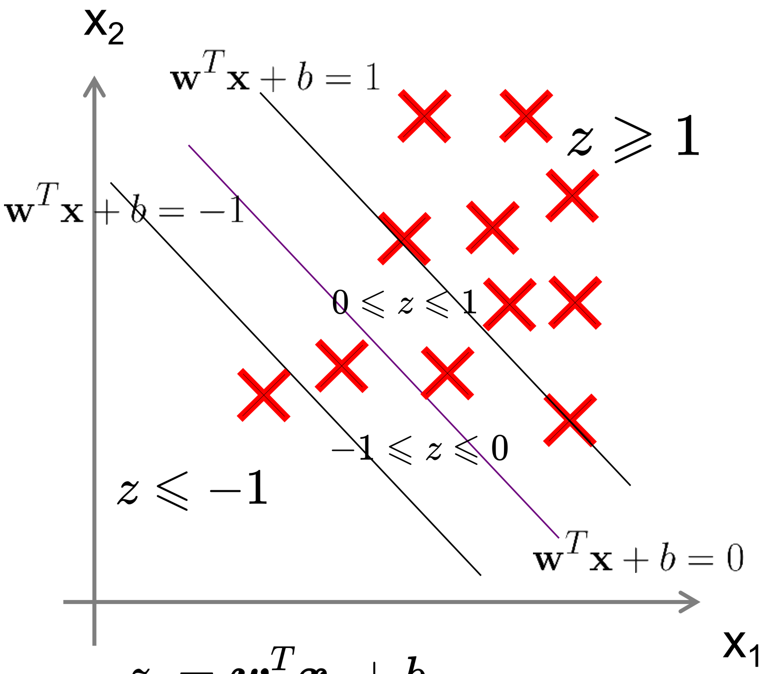
$$\begin{cases} \xi_i = 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \\ \xi_i \geqslant 0 \end{cases}$$

$$\Downarrow$$

$$\max(0, 1 - y_i(\boldsymbol{w}^T\boldsymbol{x} + b))$$

$1 - y_i(\mathbf{w}^T\mathbf{x}_i + b)$





$$z_i = \boldsymbol{w}^T\boldsymbol{x}_i + b$$
$$y_i = 1$$

2. 软间隔SVM的优化

对于线性不完全可分的情况，引入软间隔的SVM，即允许某些样本不满足约束：

$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geqslant 1$$

$$\Downarrow$$

$$y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geqslant 1 - \xi_i\,, \quad \xi_i \geq 0, i = 1, 2, \cdots, m$$

$$\min_{\boldsymbol{w},b,\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

$$\Downarrow$$

$$\min_{\boldsymbol{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\mu}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i(1 - \xi_i - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\mu_i\xi_i$$

$$\min_{\mathbf{w},b,\xi_i} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$s.t. \quad y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geqslant 1 - \xi_i$$

$$\xi_i \geqslant 0, i = 1, 2, \cdots, m$$

$$\max_{\alpha}\min_{w,b}\mathcal{L}(\mathbf{w},b,\alpha) = \max_{\alpha}\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

$$s.t. \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$\Downarrow$$

$$\max_{\alpha}\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\mu}} L(\mathbf{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\mu}) = \max_{\alpha}\sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

$$s.t. \sum_{i=1}^{m}\alpha_i y_i = 0 \quad , \quad 0 \leqslant \alpha_i \leqslant C, i = 1, 2, \cdots, m$$

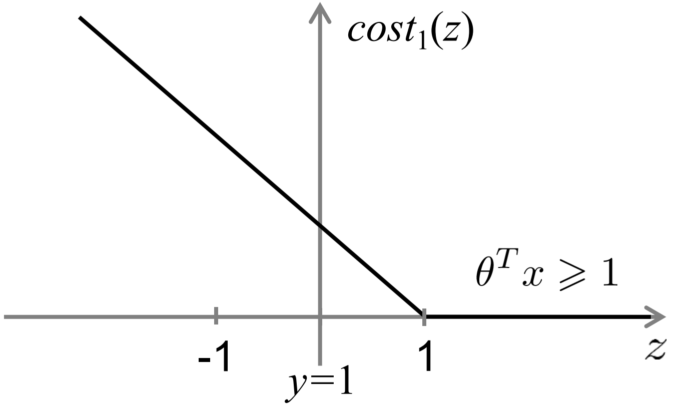
软间隔SVM的KKT条件

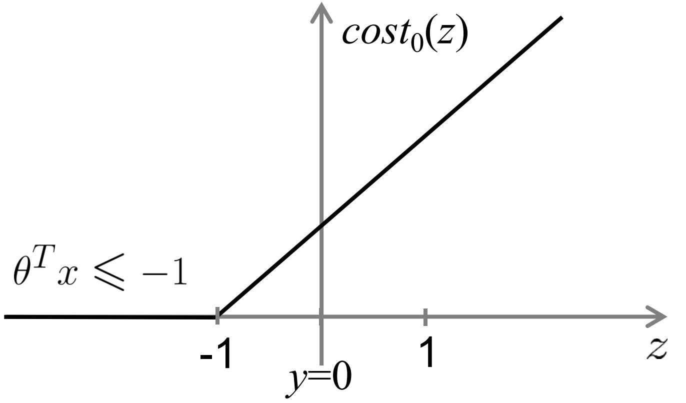
$$\begin{cases} \alpha_i \geqslant 0; \\ y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 \geqslant 0; \\ \alpha_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1) = 0. \end{cases} \Rightarrow \begin{cases} \alpha_i \geqslant 0, \mu_i \geqslant 0, \\ y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + \xi_i \geqslant 0, \\ \alpha_i(y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) - 1 + \xi_i) = 0, \\ \xi_i \geqslant 0, \mu_i\xi_i = 0. \end{cases}$$

2. 吴恩达课程中的软间隔SVM的损失函数

$$y^{(i)} = 1 \ or \ 0$$

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$





If $y = 1$, we want $\theta^T x \geq 1$ (not just ≥ 0)

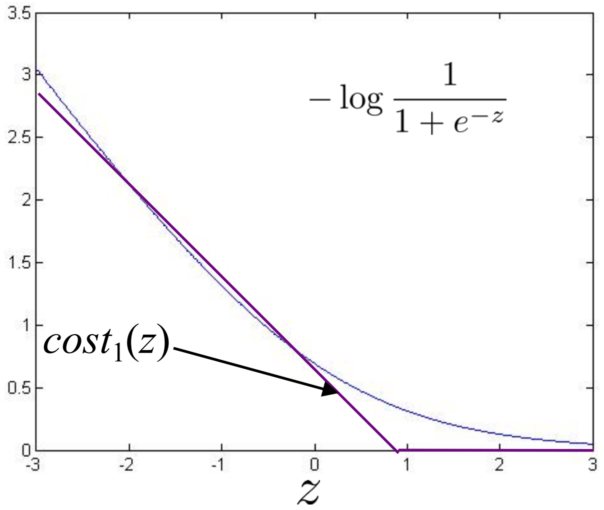
If $y = 0$, we want $\theta^T x \leq -1$ (not just < 0)

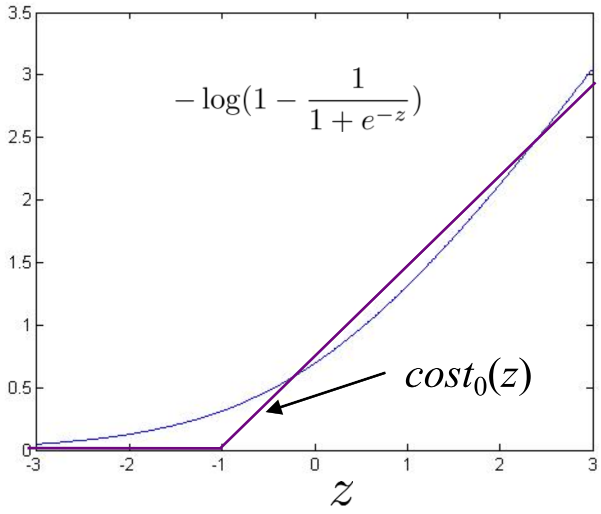
对比logistic回归和软间隔SVM的损失函数

Cost of example: $-(y\log h_\theta(x) + (1-y)\log(1-h_\theta(x)))$

$$= -y\log\frac{1}{1+e^{-\theta^T x}} - (1-y)\log(1-\frac{1}{1+e^{-\theta^T x}})$$

If $y=1$ (want $\theta^T x \gg 0$): If $y=0$ (want $\theta^T x \ll 0$):





调整C对判决模型的影响:

$$\min_{\theta} C \sum_{i=1}^{m} \left[y^{(i)} cost_1(\theta^T x^{(i)}) + (1 - y^{(i)}) cost_0(\theta^T x^{(i)}) \right] + \frac{1}{2} \sum_{i=1}^{n} \theta_j^2$$

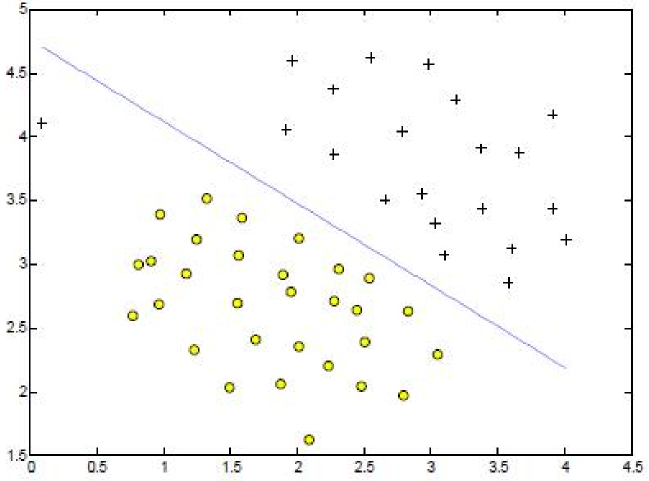


Figure 2: SVM Decision Boundary with $C = 1$ (Example Dataset 1)

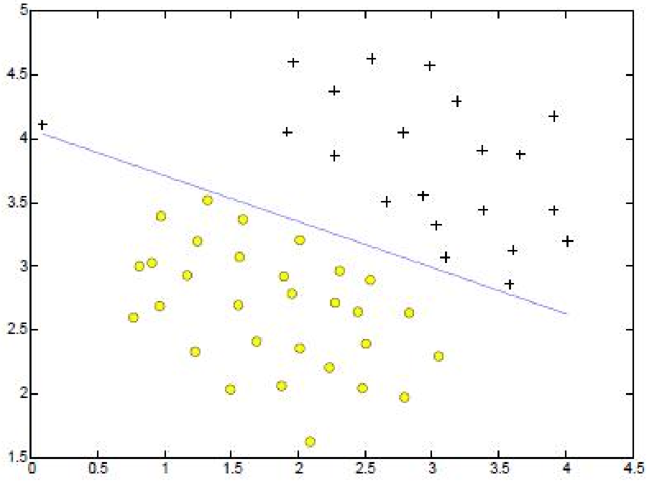
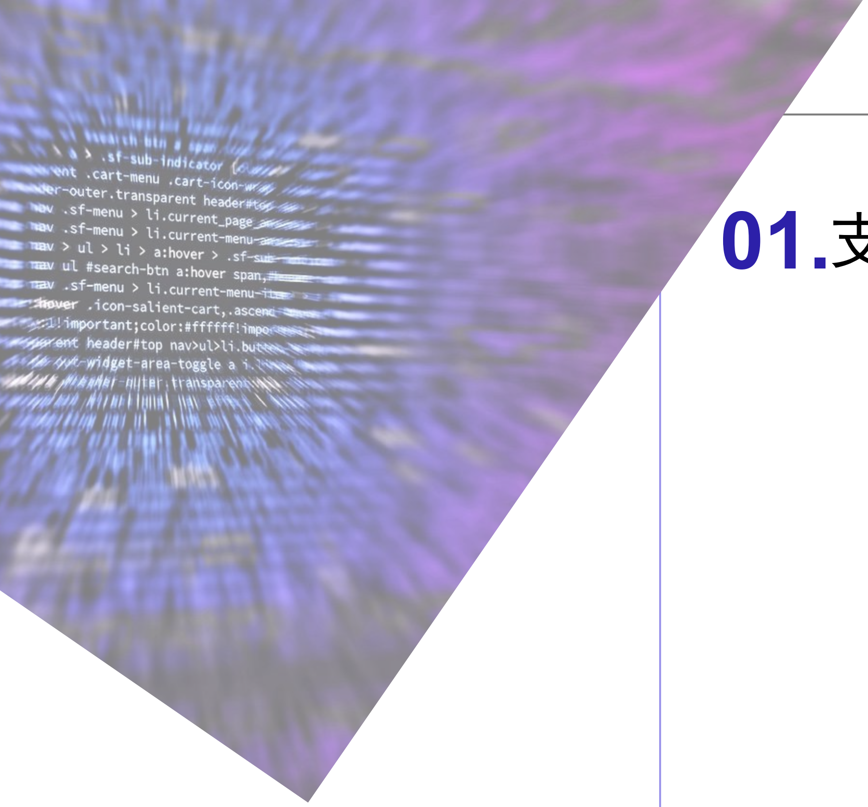


Figure 3: SVM Decision Boundary with $C = 100$ (Example Dataset 1)

详见吴恩达网课ex6中在数据集1上的实验结果



目录

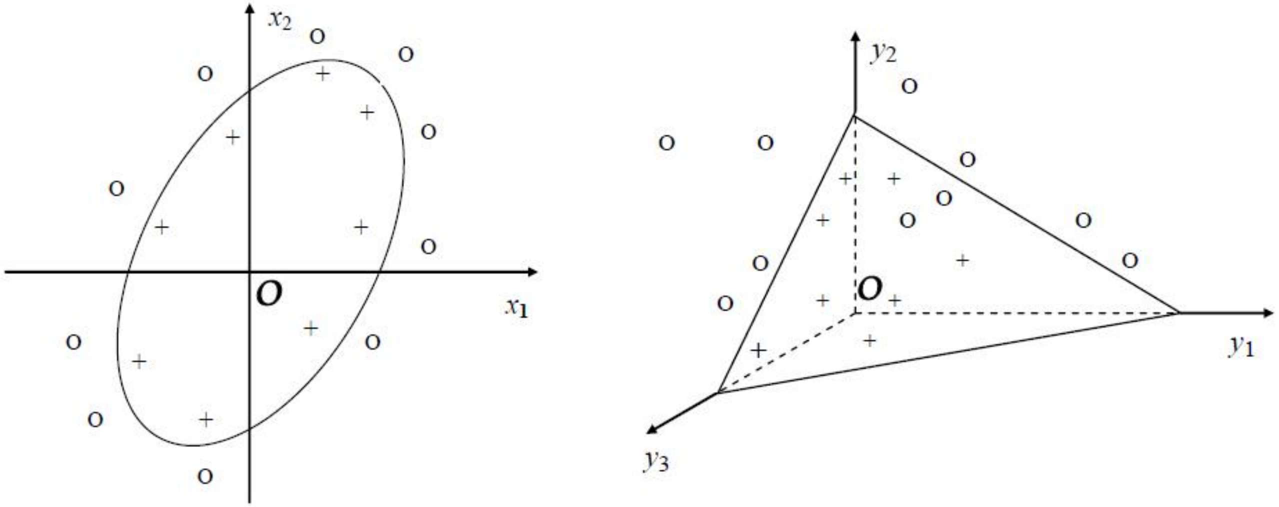
- **01.**支持向量机(Support Vector Machine)
  - 代价函数
    - 结构风险+经验风险
    - 结构风险: 最大化分割面的间隔
  - 预测模型（判决函数）
  - SVM代价函数的优化
    - 代价函数的对偶形式
    - 半正定二次规划
- **02.**软间隔SVM
- **03.**核函数(Kernel Function)
- **04.**支持向量回归

3. 通过空间变换将非线性可分问题转变为线性可分问题

$$w_1 x_1^2 + w_2 x_2^2 + \sqrt{2} w_3 x_1 x_2 + b = 0$$

$$\phi(\vec{x}) = \phi(x_1, x_2) = (y_1, y_2, y_3) = (x_1^2, x_2^2, \sqrt{2} x_1 x_2)$$

$$w_1 y_1 + w_2 y_2 + w_3 y_3 + b = 0$$



（a）在 x_1 和 x_2 张成的空间中　　（b）在 $x_1^2, x_2^2, \sqrt{2}x_1x_2$ 张成的空间中

3. 核函数：考虑空间变换前后的相似度度量变换

核函数（kernel function）$K(\vec{x},\vec{y})$ 是一个内积函数，它对于所有 $\mathbf{x},\mathbf{y}\in X\subseteq\mathbf{R}^d$ 满足：

$$K(\mathbf{x},\mathbf{y})=\langle\phi(\mathbf{x}),\phi(\mathbf{y})\rangle$$

其中，ϕ 是从 X 到（内积）特征空间 F 的一个映射。

$$\phi:\mathbf{x}\mapsto\phi(\mathbf{x})\in F$$

1. 多项式核：$K(\mathbf{x},\mathbf{y})=(\gamma\langle\mathbf{x},\mathbf{y}\rangle+c)^d$

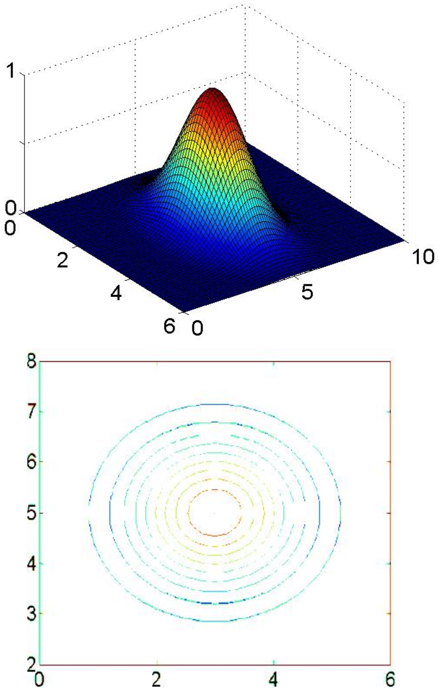
2. 高斯径向基核：$K(\mathbf{x},\mathbf{y})=\exp(\frac{-\|\mathbf{x}-\mathbf{y}\|}{\sigma^2})$

3. Sigmoid 核：$K(\mathbf{x},\mathbf{y})=\tanh(\gamma\langle\mathbf{x},\mathbf{y}\rangle+c)$

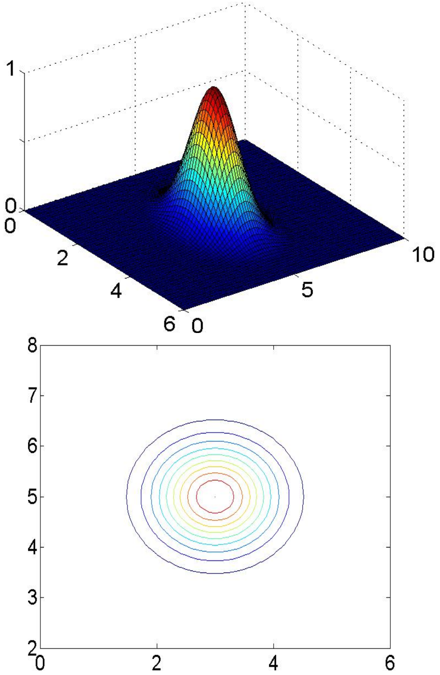
3. 最常用的核函数：高斯核函数

$$l^{(1)} = \begin{bmatrix} 3 \\ 5 \end{bmatrix}, \quad f_1 = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right)$$

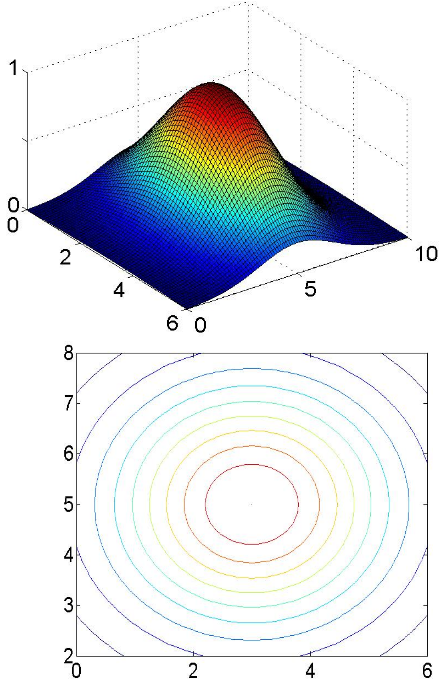
$\sigma^2 = 1$



$\sigma^2 = 0.5$



$\sigma^2 = 3$



SVM with Kernels

$$\max_{\alpha}\min_{w,b}\mathcal{L}(\mathbf{w},b,\alpha) = \max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

$$s.t.\sum_{i=1}^{m}\alpha_i y_i = 0$$

$$\max_{\alpha}\min_{\mathbf{w},b,\boldsymbol{\xi},\boldsymbol{\mu}} L(\mathbf{w},b,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\mu}) = \max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T\mathbf{x}_j$$

$$s.t.\sum_{i=1}^{m}\alpha_i y_i = 0 \quad , \quad 0\leqslant\alpha_i\leqslant C, i=1,2,\cdots,m$$

软间隔SVM的超参数与核函数参数

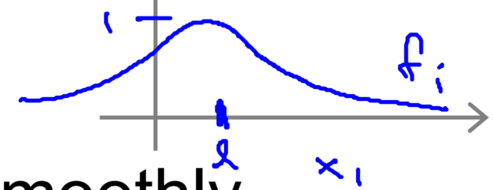
C ($=\frac{1}{\lambda}$) → Large C: Lower bias, high variance. (small λ)

→ Small C: Higher bias, low variance. (large λ)

σ^2 : Large σ^2 : Features f_i vary more smoothly.

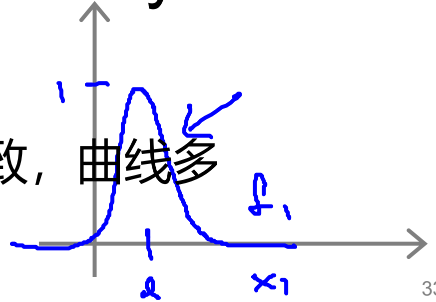
→ Higher bias, lower variance.

$\exp\left(-\frac{\|x-l^{(i)}\|^2}{2\sigma^2}\right)$



Small σ^2 : Features f_i vary less smoothly.

Lower bias, higher variance.



Lower bias, higher variance.	Large C	Small σ^2 分类面细致，曲线多
Higher bias, lower variance	Small C	Large σ^2

调整 σ^2 对判决模型的影响:

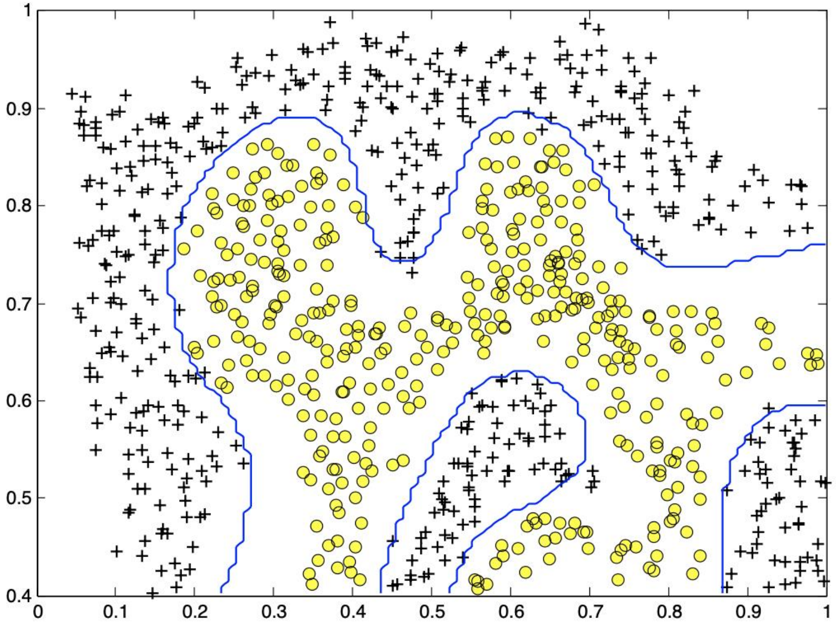


Figure 5: SVM (Gaussian Kernel) Decision Boundary (Example Dataset 2)

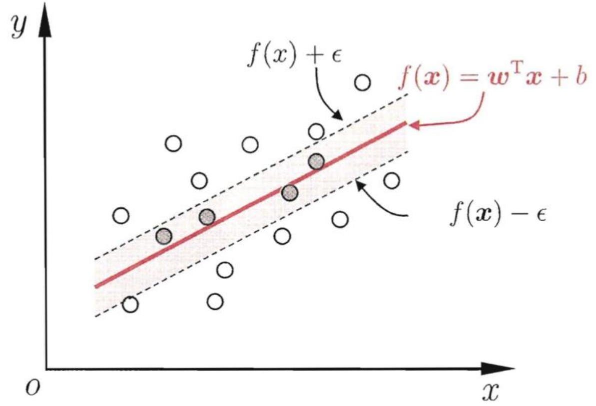
详见吴恩达网课ex6中在数据集2上的实验结果

4. 支持向量回归-SVR

传统回归：f(x)与y完全相等时损失为0；

SVR： f(x)与y的差异大于ϵ时才计算损失。

落入 ϵ 间隔带中不计算损失。



$$\min_{\mathbf{w},b} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m} \max(0, |f(x_i)-y|-\epsilon)$$

详见周志华《机器学习》6.5节

4. SVR的代价函数

$$\max_{\theta} = \frac{2}{\|\theta\|} \quad \text{或} \quad \min_{\theta} \frac{1}{2}\|\theta\|^2$$

$$s.t. \quad y_i(\mathbf{\Theta}^{\mathbf{T}}\mathbf{x}_i + b) \geqslant 1, \quad i = 1, 2, \cdots, m$$

- 拉格朗日乘子法可得如下损失函数:

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{i=1}^{m} \alpha_i(1 - y_i(\mathbf{w}^T\mathbf{x}_i + b))$$

小 测

1 $$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$s.t. \quad 1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \le 0$$

2 $$\min_{\boldsymbol{w},b,\boldsymbol{\alpha}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

3 $$\max_{\alpha} \min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i(1 - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b))$$

4 $$\max_{\alpha} \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

5 $$\boldsymbol{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \boldsymbol{x}_i$$

A)这是根据求解到的alpha最优解后所得到的w的最优解

B)原始的代价函数

C)使用拉格朗日乘子法将原始代价函数及约束转换为单一的优化问题

D)这是原始代价函数的最大最小对偶

E)这是最大最小对偶第1步的结果，该问题是关于alpha的二次规划问题，其最优解满足KKT条件

作业

原题：让直线 $x+y=1$ 与椭圆 $x^2+2y^2=C$（C 为常数）相切，求 C 取值。

题目：请使用拉格朗日乘子法（Lagrange Multiplier Method）求解。

提示：可将其转换为如下优化问题:

$$\min_{x,y}(x^2+2y^2)$$

$$s.t. \quad x+y=1$$