—武大本科生课程

第6讲　支持向量机(I)

(Lecture 6 Support Vector Machines: Part 1)

武汉大学计算机学院机器学习课程组

2023.03

第5章　支持向量机

内容目录 (以下红色字体为本讲3学时讲授内容)

Review：

1. 贝叶斯定理(Bayes theorem)

$$posterior = \frac{likelihood \times prior}{evidence}$$

$$P(\omega_i \mid \mathbf{x}) = P(\mathbf{x} \mid \omega_i)P(\omega_i) / P(\mathbf{x}) = P(\mathbf{x} \mid \omega_i)P(\omega_i) \Big/ \sum_{j=1}^{M} P(\mathbf{x} \mid \omega_j)P(\omega_j)$$

$P(\omega_i|\mathrm{x})$ 称为后验概率，对模式识别而言可理解为x来自ω_i类的概率，即 x已知的情况下其类别属于ω_i的概率为$P(\omega_i|\mathrm{x})$；$P(\mathrm{x}|\omega_i)$称为类条件概率密度(简称似然)； $P(\omega_i)$称为先验概率。

2. 最小错误率Bayes决策/最大后验Bayes决策

3. 最小风险Bayes决策

$$R(\alpha_i \mid \mathbf{x}) @ E\left[\lambda(\alpha_i, \omega_j)\right] = \sum_{j=1}^{M} \lambda(\alpha_i, \omega_j) P(\omega_j \mid \mathbf{x})$$

若$R(\alpha_k \mid \mathbf{x}) = \min_{i=1,2,\mathrm{L},M} R(\alpha_i \mid \mathbf{x})$， 则判决$\mathbf{x} \in \omega_k$。

4. Naïve Bayes决策

5. 正态分布Bayes决策

6. 最大似然估计

求最大似然函数主要步骤:

(1)写出似然函数或对数似然函数

$$L(\boldsymbol{\theta}) = p(X^N \mid \boldsymbol{\theta}) \text{或} H(\boldsymbol{\theta}) = \ln p(X^N \mid \boldsymbol{\theta})$$

(2)对似然函数或对数似然函数求偏导，令$\frac{\partial h(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$或$\frac{\partial H(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$，求出$\boldsymbol{\theta}$出最大似然估计(解析法)。

或者梯度下降法等迭代法求出$\boldsymbol{\theta}$最大似然估计。

7. Parzen窗估计/KDE (*: 选学)

除了本课件中介绍的Parzen窗估计(KDE)方法，关于KDE的Python案例实践可参考本人CSND博文：

https://yuanyx.blog.csdn.net/article/details/115175706。

8. 最大似然估计在Logistic回归模型训练中的应用

$$H(\mathbf{w}) = \ln L(\mathbf{w}) = \sum_{i=1}^{N} [y_i(\mathbf{w} \cdot \mathbf{x}_i) - \ln(1 + \exp(\mathbf{w} \cdot \mathbf{x}_i))]$$

$$J(\mathbf{w}) = -H(\mathbf{w}) = \sum_{i=1}^{N} [\ln(1 + \exp(\mathbf{w} \cdot \mathbf{x}_i)) - y_i(\mathbf{w} \cdot \mathbf{x}_i)]$$

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} H(\mathbf{w}) \Leftrightarrow \hat{\mathbf{w}} = \arg\min_{\mathbf{w}} J(\mathbf{w})$$

目标函数/准则函数$J(\mathbf{w})$对$\mathbf{w}$的梯度：

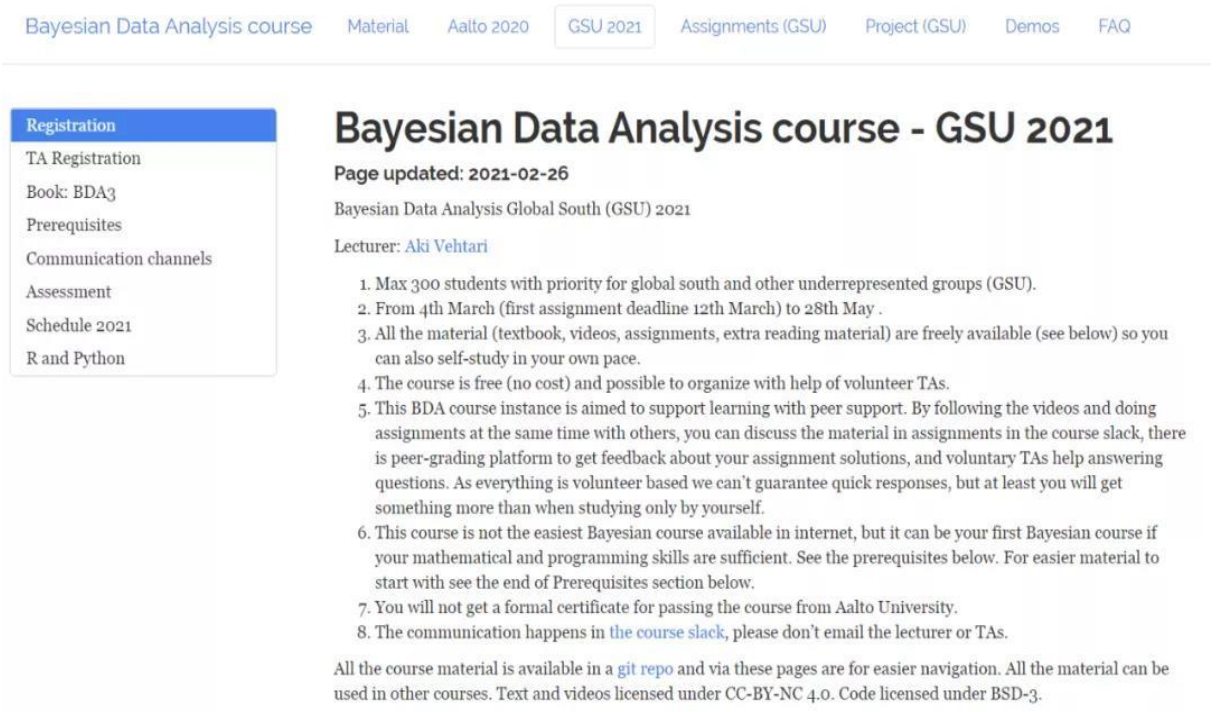
$$\nabla J(\mathbf{w}) = \sum_{i=1}^{N} \left(\frac{\mathbf{x}_i}{1 + \exp(\mathbf{w} \cdot \mathbf{x}_i)} - y_i \mathbf{x}_i \right) = \sum_{i=1}^{N} (\pi(\mathbf{x}_i) - y_i)\mathbf{x}_i$$

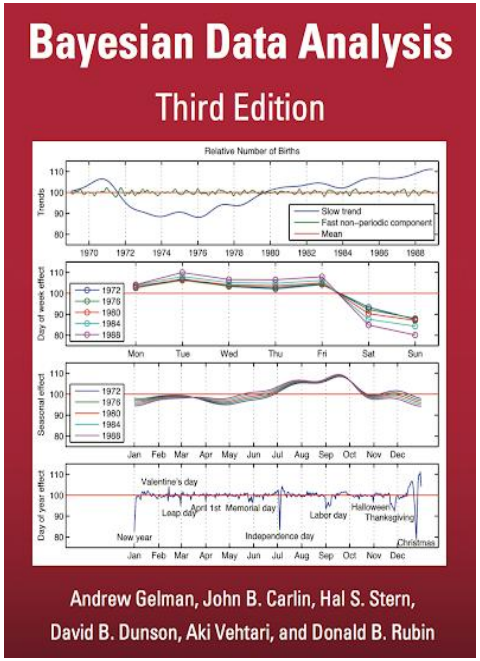
用梯度下降法迭代训练求$\mathbf{w}$最优解 $\hat{\mathbf{w}}$：

$$\mathbf{w}(k+1) = \mathbf{w}(k) - \rho \nabla J(\mathbf{w})$$

$$= \mathbf{w}(k) - \rho \sum_{i=1}^{N} (\pi(\mathbf{x}_i) - y_i)\mathbf{x}_i$$

关于贝叶斯分析 **(Bayesian Data Analysis)**学习与研究，近期芬兰阿尔托大学(Aalto University)有一门贝叶斯数据分析课程已经正式开课，对贝叶斯分析感兴趣的同学可以有选择性地进行学习。





- **课程主页：** https://avehtari.github.io/BDA_course_Aalto/gsu2021.html
- **课程资料：** https://github.com/avehtari/BDA_course_Aalto

该课程使用的教材是由 Andrew Gelman、John Carlin、Hal Stern、David Dunson、Aki Vehtari 和 Donald Rubin 编著的<Bayesian Data Analysis, 3rd ed>。课程讲师也是本书作者之一 Aki Vehtari。

课程录像(5月份课程结束后应有完整录像): https://avehtari.github.io/BDA_course_Aalto/#Videos

书籍电子版链接： https://users.aalto.fi/~ave/BDA3.pdf

Python或R语言程序实现链接: https://avehtari.github.io/BDA_course_Aalto/demos.html

5.1 支持向量机概述

本科数学基础的非数学专业理工科学生深入理解支持向量机理论建议补充下面的一些数学知识：

1.《泛函分析》

2.《最优化理论与方法》

Vapnik、Cortes等人在多年研究统计学习理论基础上对线性分类器提出了另一种设计最佳准则。其原理也是从线性可分讲起，然后再扩展到线性不可分的情况，甚至扩展到使用非线性函数中，这种分类器被称为支持向量机(Support Vector Machines，简称SVM)。SVM在解决小样本、非线性及高维模式识别中表现出许多独特的优势，并能够推广应用到函数拟合等其他机器学习问题中。

- 支持向量机方法是一种具有完备数学理论的经典机器学习方法。
- 高维模式识别是指样本维数很高。如文本的向量表示就是高维模式，如不经过降维处理的话，出现几万维的情况很正常，**SVM**在应对高维数据集和低维数据集方面都有不错的表现。主要是因为**SVM** 产生的分类器很简洁，用到的样本信息很少{仅仅用到那些称之为“支持向量**(Support vectors)**”的样本}。
- 支持向量机在设计时，需要用到条件极值问题的求解，因此需用拉格朗日乘子理论，但对多数人来说，以前学到的或常用的是约束条件为等式表示的方式，但在此处要用到以不等式作为必须满足的条件，此时只需了解拉格朗日理论的有关结论就行。

- 支持向量机方法是建立在俄罗斯著名数学家、统计机器学习大牛Vapnik统计学习理论《Statistical Learning Theory》著作的VC维理论(Vapnik Chervonekis Dimension) 和结构风险最小化原理基础上的，根据有限的样本信息在模型复杂性{即对特定训练样本的学习精度(accuracy)}和学习能力(即无错误地识别任意样本的能力)之间寻求最佳折衷，以期获得最好的泛化能力(generalization ability) {或称推广能力}。

- 所谓VC维是对函数类的一种度量，可简单的理解为问题的复杂程度，VC维越高，一个问题就越复杂。正因为SVM关注的是VC维，采用SVM解决问题时，通常是与样本的维数无关的，甚至样本是上万维的都可以，这就使得SVM很适合用来解决像文本分类这样的问题。当然，拥有这种能力是因为在SVM中引入了核函数 (核函数是足够复杂的分类函数，它的VC维很高，能够精确地记住每一个样本，但对样本之外的数据一律分类错误)。

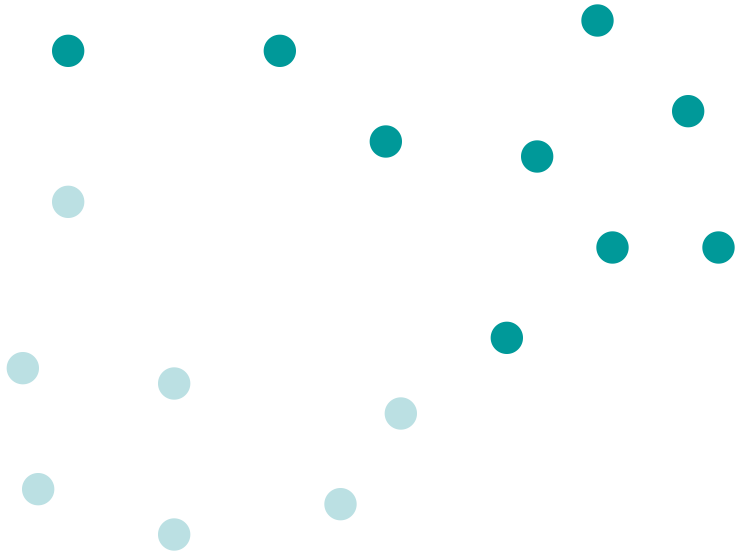
- 支持向量机是一种二类分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器；支持向量机还包括核技巧，这使它成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可形式化为求凸二次规划(Convex quadratic program)的问题；支持向量机的学习算法是求解凸二次规划的最优化算法。
- 核函数(Kernel function)表示将输入从输入空间映射到特征上空间得到时的特征向量之间的内积；通过使用核函数可以学习非线性支持向量机，等价于隐式地在高维的特征空间中学习线性支持向量机，这种方法称为核技巧(Kernel trick)。

- 支持向量机学习方法包含由简至繁的模型：线性可分支持向量机、线性支持向量机以及非线性支持向量机。
- 当训练数据线性可分时，通过硬间隔最大化(Hard margin maximization)学习一个线性分类器，即线性可分支持向量机(SVM in linearly separable case)，又称硬间隔支持向量机。
- 当训练数据近似线性可分时，引入松驰变量，通过软间隔最大化(Soft margin maximization)也学习一个线性分类器，即线性支持向量机(Linear SVM)，又称为软间隔支持向量机。
- 当训练数据线性不可分时，通过使用核技巧及软间隔最大化学习非线性支持向量机(Non-linear SVM)。

- 机器学习从本质上讲就是一种对问题真实模型的逼近（在实际问题中，通常选择一个我们认为较好的近似模型，这个近似模型也称为假设），近似模型/假设与问题真实解之间的误差，就叫做风险（更严格地说，误差的累积叫做风险）。

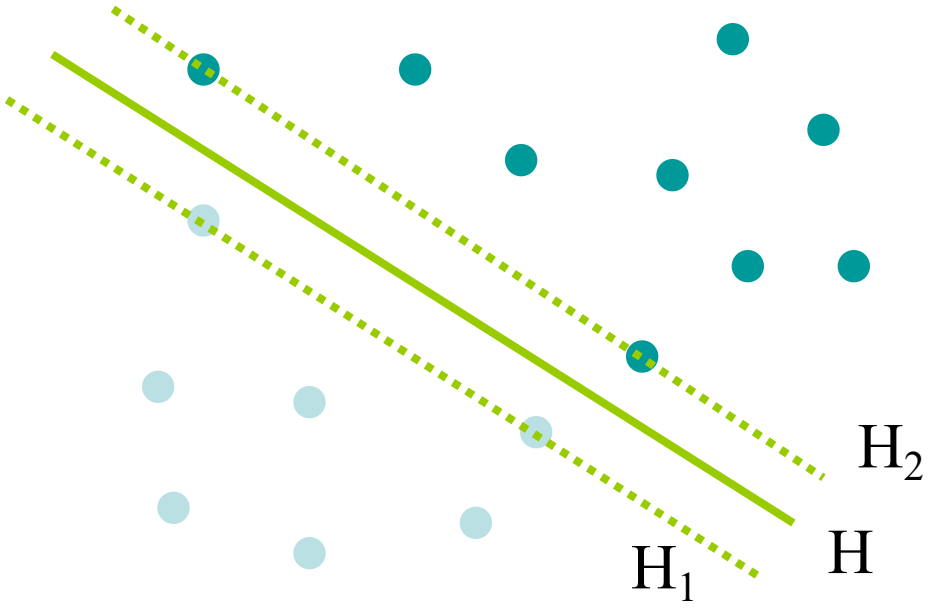
5.2 SVM分类的基本思想

由于两类别训练样本线性可分，因此在两个类别的样本集之间存在一个间隔(Margin)。对一个二维空间的问题可用下图表示。

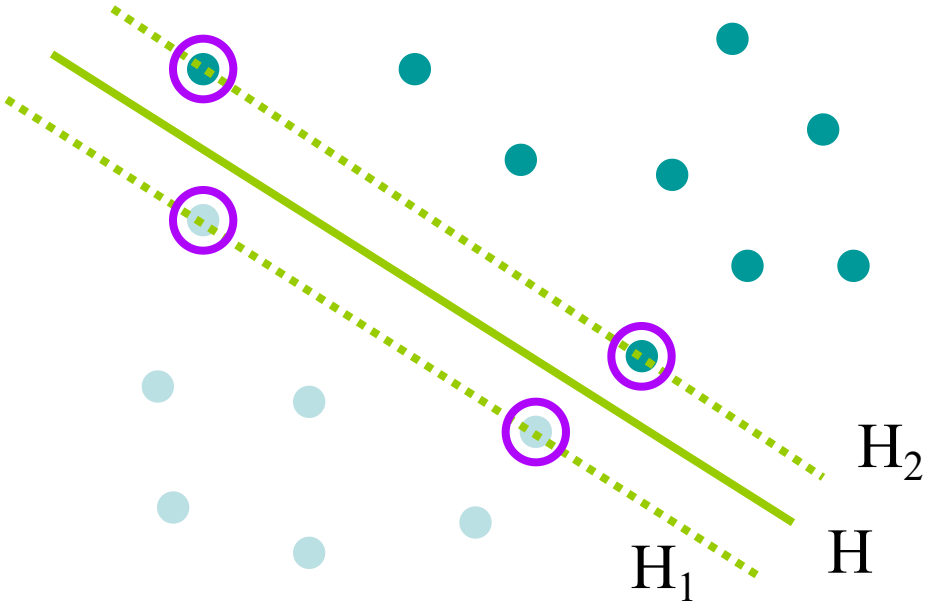


线性可分条件下的支持向量机最优分界面

- 其中H是将两类分开的分界面，而H_1与H_2与H平行，H是其平分面，H_1上的样本是第一类样本到H最近距离的点，H_2的点则是第二类样本距H的最近点。



- 由于这两种样本点很特殊，处在间隔的边缘上，因此再附加一个圈表示。样本中距离超平面最近的一些点称为**支持向量**(Support vectors)，它们决定了这个间隔。

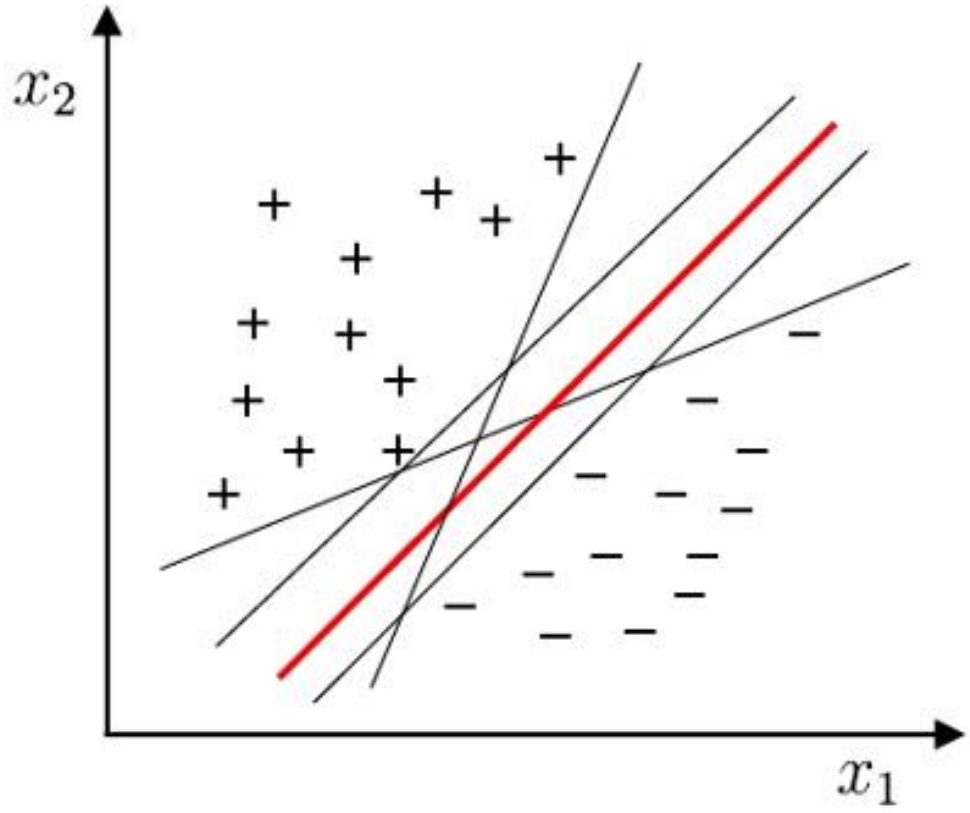


线性可分条件下的支持向量机最优分界面

- 从图中可以看出能把两类分开的分界面并不止H这一个，如果略改变H的方向，则根据H_1、H_2与H平行这一条件，H_1、H_2的方向也随之改变，这样一来，H_1与H_2之间的间隔(两条平行线的垂直距离)会发生改变。
- 显然使H_1与H_2之间间隔最大的分界面H是最合理的选择，因此最大间隔准则就是支持向量机的最佳准则。

支持向量机最早是运用于二分类问题，且大多数情况是非线性的。其基本思想是：首先通过一个非线性变换将输入空间中的向量映射到一个高维的特征空间，然后再在这个高维特征空间中求解最优分类超平面，而这种非线性变换是通过定义适当的内积(核)函数来实现的，也就是说，将高维特征空间中的特征向量和特征空间中向量的内积化为原空间中的核函数计算。

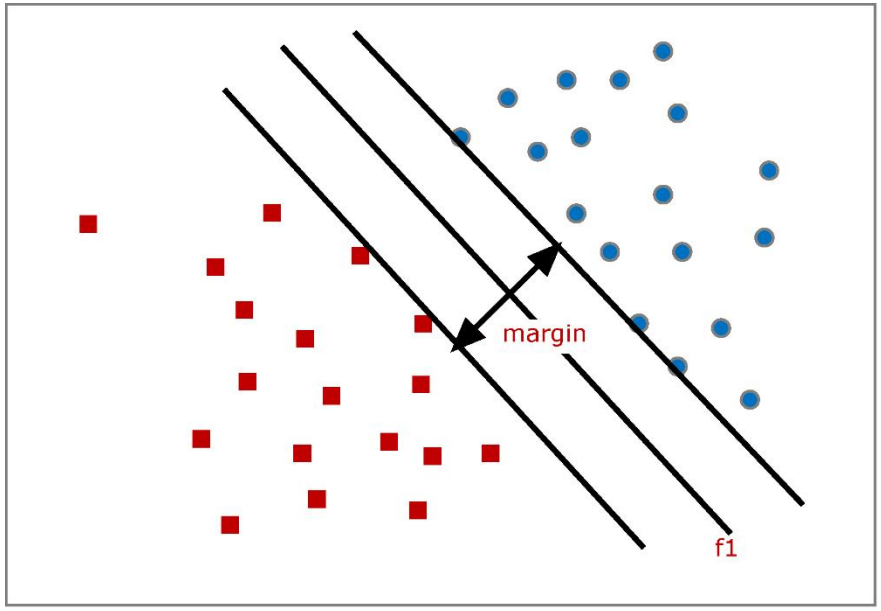
“正中间”分类器相对更好，但不唯一

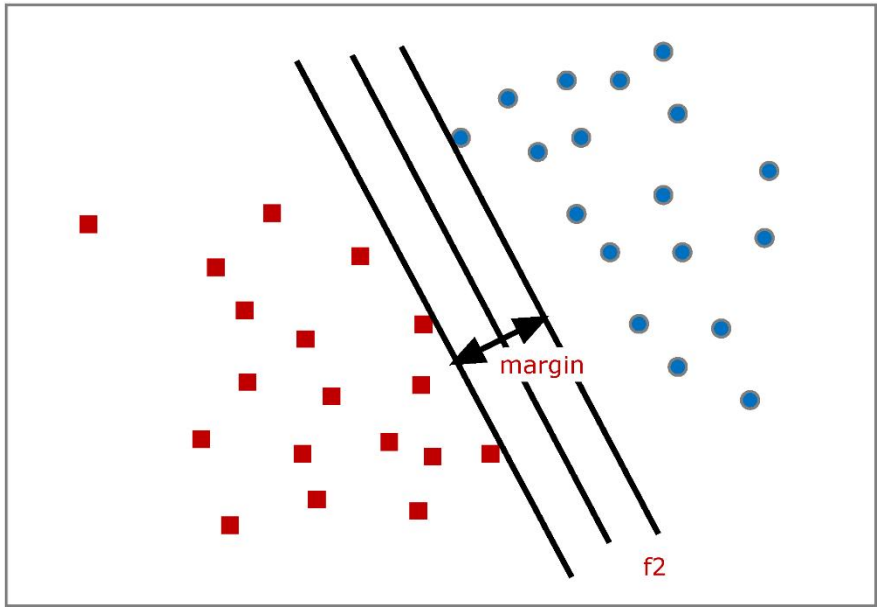


“正中间”分类器：对两类样本交汇区域的分类能力更强

稳健性更好，泛化能力更强

“最大分隔间隔”是“正中间”分类器选择标准





都是“正中间”分类器，但是左边的比右边的更好

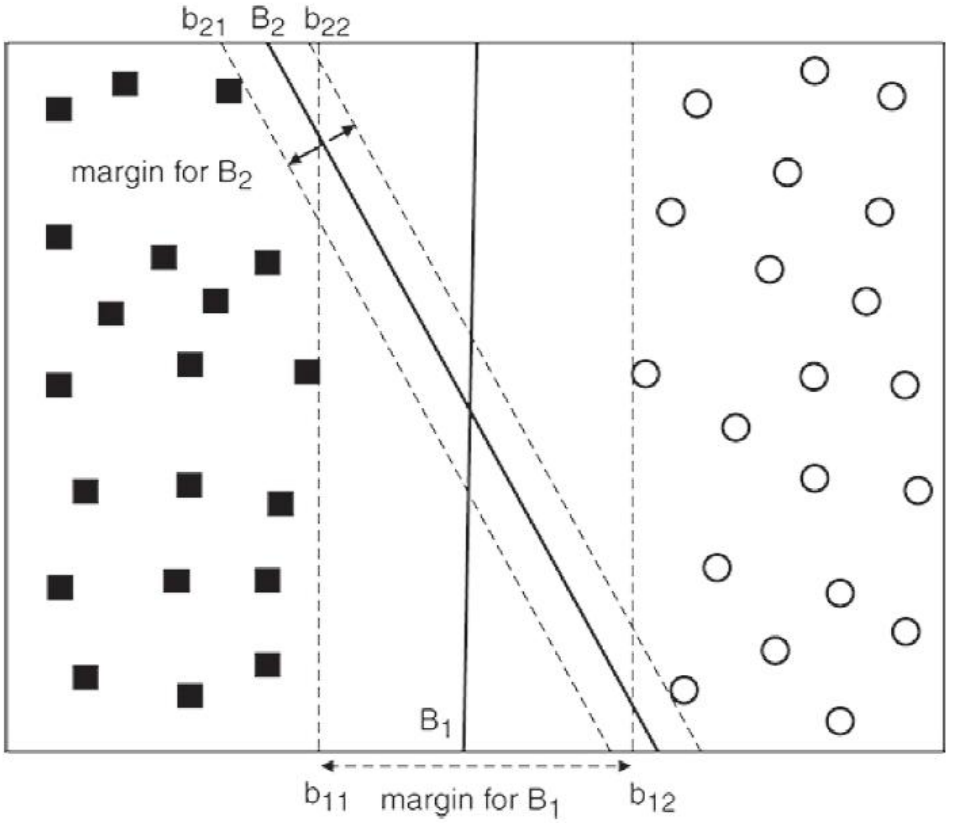
与两类当前边界越远，对未来两类样本的**越界点**更少误判

如何寻找具有最大间隔（缓冲区宽度）的线性分类器？

分类超平面、间隔边界、分类面缓冲区、分类面间隔

- 假设某二分类训练集**线性可分**，则存在一个**分类超平面**将其线性分开。分别向正类和负类区平移分类超平面，在与两个类的最近样本点接触后即停止，形成两个间隔边界 (支持超平面)

- 两个支持面之间的无样本数据区称为**分类面缓冲区**，缓冲区的宽度称为**分类面间隔**，是衡量分类超平面的一个重要指标
- 虽然分类超平面 B_1 和 B_2的训练错误率都是0，但是在未来的测试数据上，两者的泛化性能（测试错误率）表现会有不同。Who is better ?



最佳分类超平面的选择——最大间隔的基本思想

- 直观上看，如果**分类超平面的间隔比较小**，那么训练数据的轻微扰动，或计算过程的数值扰动，都会对分类器性能产生较大影响。因此，具有较小间隔的分类超平面更容易出现过拟合 (对训练数据敏感)。同时，因为较小的分类面间隔，测试数据在支持面 (类边界) 附近的微小扰动 (这在实际情况中几乎是不可避免的)，就会带来泛化性能的下降 (较高的测试错误率)。因此，小间隔分类器的稳健性和泛化性能都较差
- 另一方面，具有**较大间隔的分类超平面**同时远离两个类的训练实例，因此训练数据和计算过程的轻微扰动，对分类器性能的影响不大，分类器具有较高的稳健性。同时，大间隔分类器还有充分的余地，在测试数据在支持面 (类边界) 附近有微小扰动时，仍能有较高的分类性能。相比较于小间隔分类器，大间隔分类器具有更高的稳健性 (对训练数据的扰动不敏感) 和更好的泛化性能 (较低的测试错误率)

硬间隔与线性可分支持向量机

- 支持向量机 (Support Vector Machine) 的基本思想是
  - 找到能够正确划分训练数据集、并且**间隔最大的分类超平面**
- 对**线性可分**训练数据集
  - 线性可分分离超平面可以有无穷多，但是支持面间隔最大的分离超平面则是唯一的。这时的分类超平面称为 **硬间隔最大分类超平面**
- 对**轻微线性不可分**训练数据集
  - 存在唯一的间隔最大分离超平面，称为 **软间隔最大分类超平面**
- 间隔最大化的直观解释是：
  - 对训练数据集找到几何间隔最大的超平面，意味着：不仅能将正负训练样例**正确地**分开，而且对最难分的样例点 (类边界之间的点) 也有**足够大的确信度**将它们分开
  - 对用于计算分类面的训练数据的微小变化不敏感，不易过拟合

线性可分支持向量机：定义

定义：给定线性可分训练数据集, 通过间隔最大化或等价地求解相应的凸二次规划问题学习得到的分离超平面为

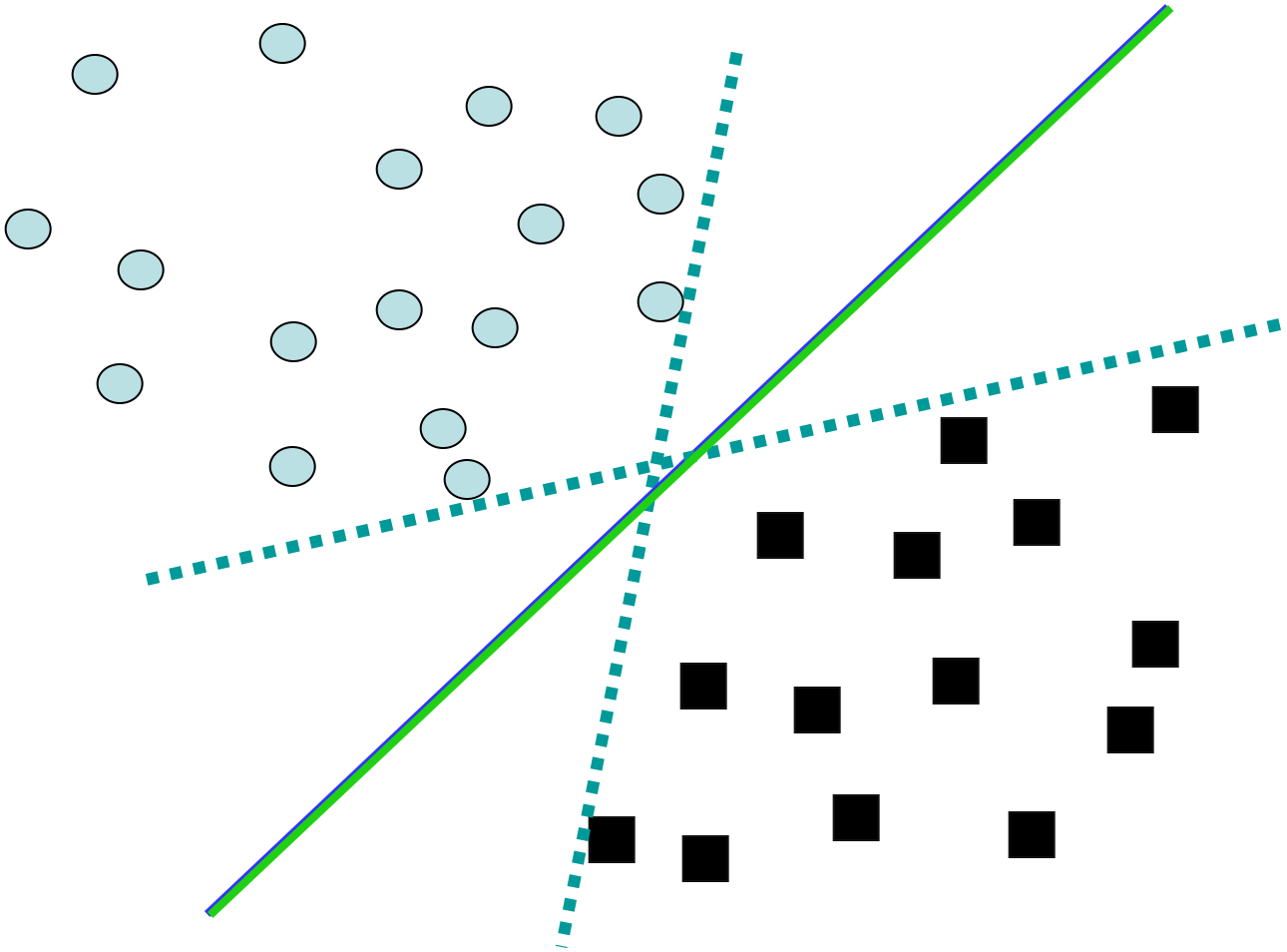
$$\mathbf{w}^* \cdot \mathbf{x} + b^* = 0$$

以及相应的分类决策函数

$$f(\mathbf{x}) = sign(\mathbf{w}^* \cdot \mathbf{x} + b^*)$$

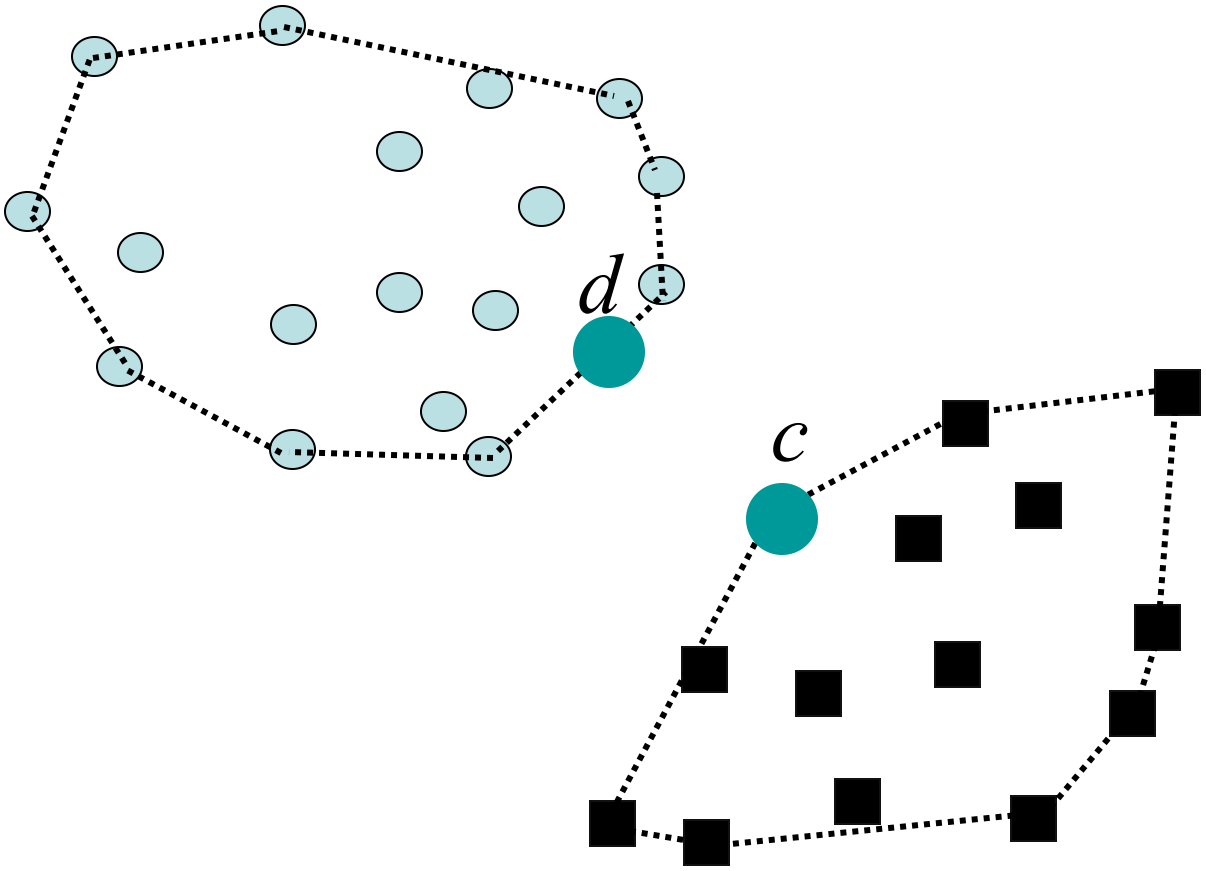
称为线性可分支持向量机。

5.3 支持向量机理论基础

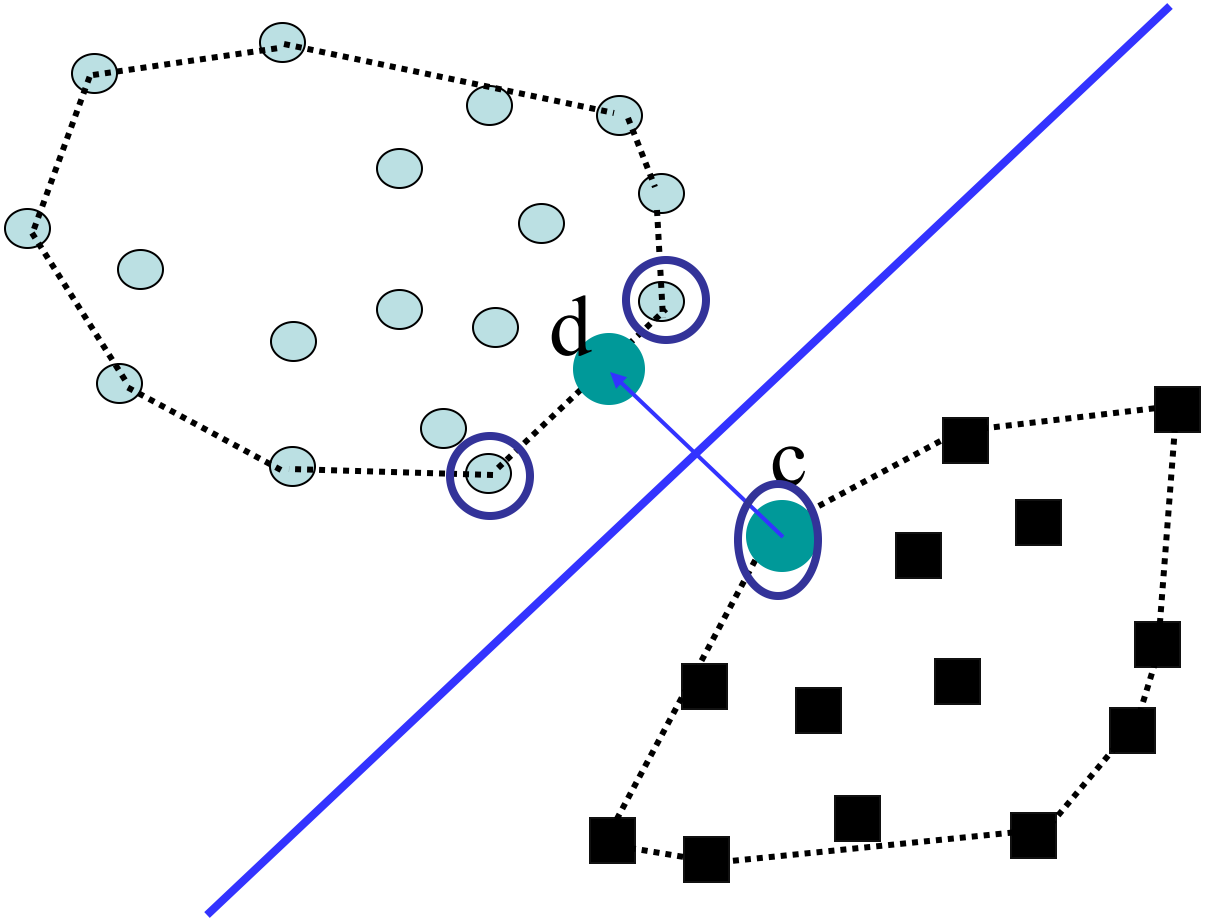


Best Linear Separator?

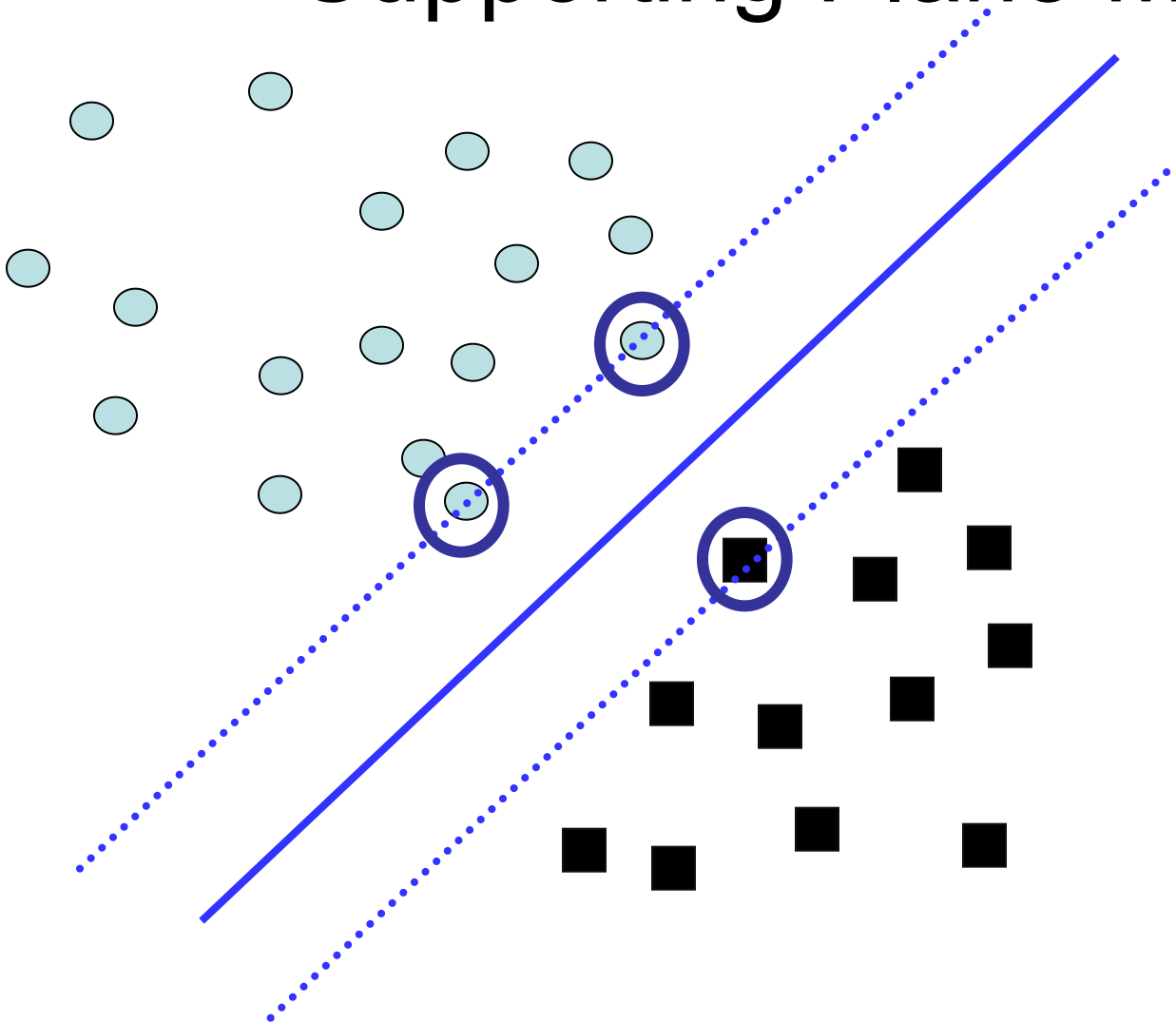
Find Closest Points in Convex Sets



Plane Bisects Closest Points



Best Linear Separator: Supporting Plane Method



Maximize distance between two parallel supporting planes

Distance

= "**Margin**"

$$= \frac{2}{\| \mathbf{w} \|}$$

5.3.1 线性可分支持向量机

- 为了将这个准则具体化，需要用数学式子表达。为方便起见，将训练样本集表示成 $\{\mathbf{x}_i, y_i\}$，$i=1, \ldots, N$，其中$\mathbf{x}_i$为d维向量也就是特征向量，而$y_i \in \{-1, +1\}$，即类标签y_i用+1或-1表示，通常称类标签为+1的样本为正例(positive example)，类标签为-1的样本为反例(negative example)。

- 现对这两类样本进行分类。即根据训练样本确定最大分类间隔的分隔超平面(Separating hyperplane)。设最优超平面的方程为: $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = 0$ (1)

- 根据第3讲 线性判别分析中超平面的几何性质②，样本点 $\mathbf{x}$到超平面H的正交投影$\|\mathbf{r}\|$与$|g(\mathbf{x})|$ 值成正比(点到平面的距离公式)，样本与最优超平面$(\mathbf{w}, b)$的之间的距离为:

$$\|\mathbf{r}\| = \frac{|g(\mathbf{x})|}{\|\mathbf{w}\|} = \frac{|\mathbf{w}^T\mathbf{x} + b|}{\|\mathbf{w}\|}$$

这里$\|\mathbf{w}\|$是$\mathbf{w}$的L2范数(L2 norm)。注意到通过等比例地缩放权向量$\mathbf{w}$和阈值b，最优超平面存在多个解，我们对超平面进行规范化，选择使得距超平面最近的样本$\mathbf{x}_i$满足$|\mathbf{w}^T\mathbf{x}_i + b| = 1$的$\mathbf{w}$和$b$，即得到规范化超平面。

向量的范数

$$\forall \mathbf{x} = (x_1, x_2, \mathrm{L}\ \ , x_n)^T$$

$$L_0\text{-0 norm}: \|\mathbf{x}\|_0 = \sum_{i=1}^{n} \left| x_i \right|^0 = \sum_{i=1}^{n} \left| sign(x_i) \right|$$

$$L_1\text{-1 norm}: \|\mathbf{x}\|_1 = \sum_{i=1}^{n} \left| x_i \right|$$

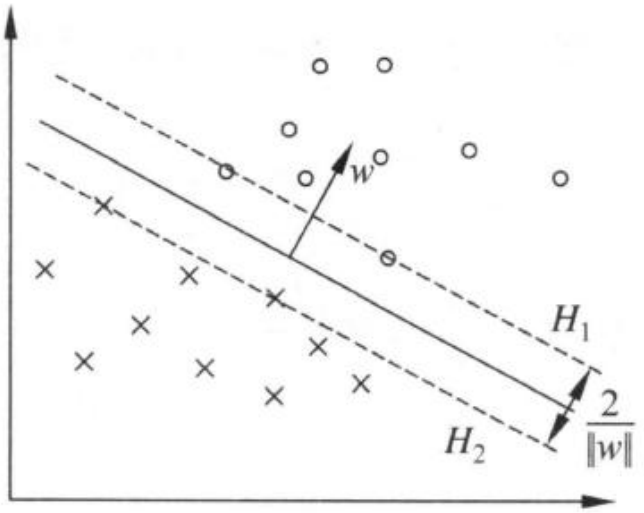
$$L_2\text{-2 norm}: \|\mathbf{x}\|_2 = \left(\sum_{i=1}^{n} \left| x_i \right|^2 \right)^{1/2}$$

$$L_p\text{-}p\text{ norm}: \|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} \left| x_i \right|^p \right)^{1/p}$$

$$L_\infty\text{-}\infty\text{ norm}: \|\mathbf{x}\|_\infty = \max_{1 \le i \le n} \left| x_i \right|$$

几个概念：支持向量、间隔边界

- 在线性可分情况下，样本点中与分类超平面距离最近的样本点的样例称为**支持向量**(support vector)。支持向量是使约束条件式 等号成立的点，也就是**函数间隔等于1的样例点**
- 对$y_i = +1$的正例点，支持向量在超平面 H_1: $\mathbf{w}\cdot\mathbf{x} + b = 1$ 上，超平面H_1称为**正类支持面（正类间隔超平面，正类间隔边界）**
- 对$y_i = -1$的正例点，支持向量在超平面 H_2: $\mathbf{w}\cdot\mathbf{x} + b = -1$ 上，超平面H_2称为**负类支持面（负类间隔超平面， 负类间隔边界）**



- 求取分离面时，**只有支持向量在起作用，其他样例点并不起作用**——Hence the name
- 支持向量个数一般很少，所以SVM是由少量“重要的”训练样本确定的

- 此时从最近样本到超平面H的距离为：

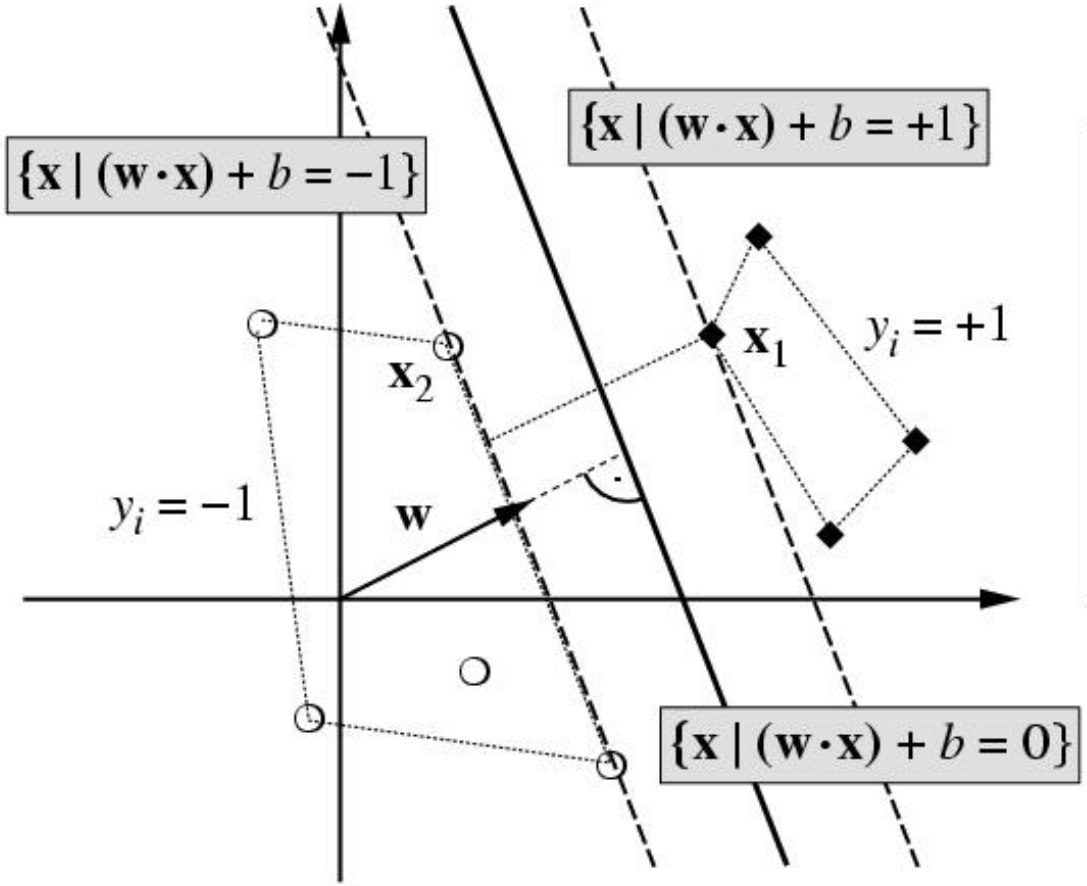
$$\frac{\left|\mathbf{w}^T\mathbf{x}+b\right|}{\|\mathbf{w}\|}=\frac{1}{\|\mathbf{w}\|} \qquad (2)$$

H1到H2的间隔，即分类间隔：

$$\text{Distance/Margin}=\frac{2}{\|\mathbf{w}\|} \qquad (3)$$

至此，问题逐渐明朗化，我们的目标是寻找使得(3)式最大化的法向量$\mathbf{w}$，然后将$\mathbf{w}$代入关系式$|\mathbf{w}^T\mathbf{x}_i+b|=1$，即可得到$b$。

计算分类超平面的间隔



Note:

$$(\mathbf{w}\cdot\mathbf{x}_1) + b = +1$$
$$(\mathbf{w}\cdot\mathbf{x}_2) + b = -1$$

$$\Rightarrow \quad (\mathbf{w}\cdot(\mathbf{x}_1-\mathbf{x}_2)) = 2$$

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

函数间隔=2

线性可分条件下的支持向量机最优分界面

- 最大化(3)式等价于最小化下式：

$$\text{Minimize} \qquad J(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2 \tag{4}$$

- 另外，还有如下约束条件：

$$\text{Subject to} \qquad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \cdots, N \tag{5}$$

- 这是因为距离超平面最近的样本点$\mathbf{x}_k$满足$|\mathbf{w}^T\mathbf{x}_k+b|=1$，而其他的样本点$\mathbf{x}_i$到超平面的距离$d(\mathbf{x}_i)$要大于等于$d(\mathbf{x}_k)$，因此有：

$$|\mathbf{w}^T\mathbf{x}_i+b| \geq 1 \tag{6}$$

具体地说，我们设定正例所在的一侧超平面的正方向，则对于正例(对应的类标签y_i为+1的样本$\mathbf{x}_i$)有：$\mathbf{w}^T\mathbf{x}_i+b \geq 1$ **(6a)**；而对于反例(对应的类标签y_i为-1的样本$\mathbf{x}_i$)有：$\mathbf{w}^T\mathbf{x}_i+b \leq -1$ **(6b)**。在式**(6a)**、**(6b)**的两端分别乘以对应其$\mathbf{x}_i$的类标签y_i，由于它们的y_i分别为+1和-1，因此得到**(5)**式统一形式的表达式。

- 因此欲达到Vapnik提出的使间隔最大化准则，则应使||w||最小。

$$\text{Minimize} \quad J(\mathbf{w}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} \tag{4}$$

$$\text{Subject to} \quad y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \cdots, N \tag{5}$$

- 由于式(4)中的目标函数$J(\mathbf{w})$是二次函数，意味着只存在一个全局最小值，因此不必再像在神经网络的优化过程中那样担心搜索陷入局部极小值；
- 现在要做的是在式(5)约束条件下找到能够最小化 (4)式的超平面方向向量**w**，对于这样一个带约束条件为不等式的条件极值问题，需要引用扩展的(广义)拉格朗日乘子理论，按这个理论构造拉格朗日函数的原则为：

- 使目标函数$J(\mathbf{w})$为最小，减去用拉格朗日乘子α_i(乘数值必须不小于0)与约束条件函数的乘积，在讨论的问题中，可将此条件极值转化为下面不受约束的优化问题，即关于$\mathbf{w}$、b和α_i来最小化L值：

$$L(\mathbf{w},b,\boldsymbol{\alpha})=\frac{1}{2}\mathbf{w}^T\mathbf{w}-\sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i+b)-1), \qquad (6)$$

$$\alpha_i \geq 0, \quad i=1,2,...,N$$

(6)式称为广义Langrage函数(generalized Lagrange function)。

- 目标函数是二次函数，而约束条件为线性函数(二次规划问题)，根据拉格朗日理论此二次规划问题问题存在唯一解。根据研究广义拉格朗日理论的Kuhn与Tucker等人的研究，表明以下是该唯一解的充分必要条件：

$$\frac{\partial L}{\partial \mathbf{w}}=\mathbf{0}, \quad \frac{\partial L}{\partial b}=0$$

Definition:

An optimization problem in which the objective function, inequality and equality constraints are all linear functions is called a Linear Program (LP, 线性规划). If the objective function is quadratic while the constraints are all linear, the optimization problem is called a Quadratic Program (QP, 二次规划).

求L对$\mathbf{w}$和b的偏导数，并令其等于0，得到(7)式和(8)式。

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = \mathbf{0}$$

$$\Rightarrow \quad \boxed{\mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i} \tag{7}$$

$$\frac{\partial L}{\partial \mathbf{w}} = \left(\frac{\partial L}{w_1}, \frac{\partial L}{w_2}, \ldots, \frac{\partial L}{w_n} \right)$$

$$\frac{\partial L}{\partial b} = \sum_{i=1}^{N} \alpha_i y_i = 0 \Rightarrow \boxed{\sum_{i=1}^{N} \alpha_i y_i = 0} \tag{8}$$

$$L(\mathbf{w},b,\boldsymbol{\alpha})=\frac{1}{2}\mathbf{w}^T\mathbf{w}-\sum_{i=1}^{N}\alpha_i(y_i(\mathbf{w}^T\mathbf{x}_i+b)-1), \qquad (6)$$

$$\alpha_i \geq 0, \quad i=1,2,...,N$$

展开式(6)，得：

$$L(\mathbf{w},b,a)=\frac{1}{2}\mathbf{w}^T\mathbf{w}-\sum_{i=1}^{N}\alpha_i y_i\mathbf{w}^T\mathbf{x}_i-b\sum_{i=1}^{N}\alpha_i y_i+\sum_{i=1}^{N}\alpha_i \qquad (9)$$

再将式(7)和式(8)代入式(9)，得：

$$L(\mathbf{w},b,a)=\frac{1}{2}\mathbf{w}^T(\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i)-\mathbf{w}^T\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i-0+\sum_{i=1}^{N}\alpha_i$$

$$=-\frac{1}{2}(\sum_{i=1}^{N}\alpha_i y_i\mathbf{w}^T\mathbf{x}_i)+\sum_{i=1}^{N}\alpha_i$$

$$=-\frac{1}{2}(\sum_{i=1}^{N}\alpha_i y_i(\sum_{j=1}^{N}\alpha_j y_j\mathbf{x}_j)^T\mathbf{x}_i)+\sum_{i=1}^{N}\alpha_i$$

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

上式与**w**、b无关，仅为 α_i 的函数，记为：

$$L_D(\boldsymbol{\alpha}) = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j + \sum_{i=1}^{N}\alpha_i \qquad (10)$$

此时约束条件为：

Subject to $\alpha_i \geq 0$ and $\sum_{i=1}^{N}\alpha_i y_i = 0$ (11)

以上带约束条件为不等式的条件极值问题是一个拉格朗日对偶问题(Lagrange duality problem)。

$$L_D(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$= \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i D_{ij} \alpha_j$$

- *Maximize* $L_D(\boldsymbol{\alpha}) = -\frac{1}{2} \boldsymbol{\alpha}^T D \boldsymbol{\alpha} + \sum_{i=1}^{N} \alpha_i$ **(10)**

Subject to $\sum_{i=1}^{N} y_i \alpha_i = 0, \quad \boldsymbol{\alpha} \geq \mathbf{0}$ **(11)**

此拉格朗日对偶问题是一个关于拉格郎日乘子$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_N)^T$的凸二次规划 (Convex quadratic program) 问题，可借助最优化理论中的标准优化方法求解。

- 而 $D=[D_{ij}]_{\mathrm{N}\times\mathrm{N}}$是形如下式的 N×N阶矩阵：

$$D_{ij}=y_i y_j \mathbf{x}_i^T \mathbf{x}_j \tag{12}$$

拉格朗日理论证明：满足上述约束条件式(11) 时，找L_{D}极大值的解就是$L(\mathbf{w}, b, \boldsymbol{\alpha})$公式的条件极小值。

由对偶问题L_D的最优$\boldsymbol{\alpha}^*$值，便可求出L的最优$\mathbf{w}^*$值：

$$\mathbf{w}^*=\sum_{i}^{N}\alpha_i^* y_i \mathbf{x}_i \tag{13}$$

- b can be determined from $\boldsymbol{\alpha}^*$, which is a solution of the dual problem, and from the Kuhn-Tucker conditions

$$\alpha_i^*\left(y_i(\mathbf{w}^*\cdot\mathbf{x}_i+b^*)-1\right)=0, \quad i=1,\ldots,N \qquad \textbf{(14)}$$

$$\frac{\partial L}{\partial \mathbf{w}}=\mathbf{w}-\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i=0 \Longrightarrow \mathbf{w}=\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i$$

$$y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1, \quad i=1,2,\cdots,N \qquad \textbf{(5)}$$

- Note that the only α_i* that can be nonzero in **(14)** those for which the constraints **(5)** are satisfied with the equality sign.

$$\alpha_i^* (y_i (\mathbf{w}^* \cdot \mathbf{x}_i + b^*) - 1) = 0, \quad i = 1, \ldots, N \qquad \textbf{(14)}$$

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i = 0 \Longrightarrow \mathbf{w} = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, 2, \cdots, N \qquad \textbf{(5)}$$

- Most of the constraints in **(5)** are satisfied with inequality signs i.e., most α^* solved from the dual are null.

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \;\; i = 1, 2, ..., N \;\; \textbf{(5)}$$

- Where for all i=1,……,N, either

$$y_i(\mathbf{w}\cdot\mathbf{x}_i+b)>1\Rightarrow\alpha_i=0\rightarrow\mathbf{x}_i \text{ irrelevant}$$

or $y_i(\mathbf{w}\cdot\mathbf{x}_i+b)=1$ (on the margin)

$\rightarrow\mathbf{x}_i$ Support vector

- The solution is determined by the examples on the margin. Thus

$$f(\mathbf{x})=\mathrm{sign}(\mathbf{w}^T\mathbf{x}+b)$$

$$=\mathrm{sign}(\sum_{i=1}^{N}\alpha_i y_i\mathbf{x}_i{}^T\mathbf{x}+b)$$

$$y_i(\mathbf{w}\cdot\mathbf{x}_i+b)\geq 1,\quad i=1,2,\cdots,N \qquad (5)$$

- 只有*满足约束条件(5)中等于的样本数据点，其拉格朗日乘子*α_i*才可能不为零*；而对满足约束条件(5)中大于的样本数据来说，其拉格朗日乘子α_i必须为零，显然只有部分(经常是少量)的样本数据的α_i不为零，而线性分界面的权向量**w**则是这些α_i不为零的样本数据的线性组合，α_i*不为零的样本数据*$\mathbf{x}_i$*也因而被称为*支持向量。

凸二次规划优化问题中的KKT条件：

$$\min_{\mathbf{w}} \quad f(\mathbf{w})$$

$$s.t. \quad g_i(\mathbf{w}) \le 0 \qquad i = 1,\ldots,k$$

$$h_i(\mathbf{w}) = 0 \qquad i = 1,\ldots,m$$

定义广义$Lagrange$函数：

$$L(\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta}) = f(\mathbf{w}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{w}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w})$$

求$f(\mathbf{w})$凸函数最优解$\mathbf{w}^*$的充要条件是著名的KKT条件（Karush-Kuhn-Tucker conditions）：

$$1) \frac{\partial L(\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta})}{\partial w_i} = 0 \qquad i = 1,\mathrm{L}\ ,n$$

$$2) \alpha_i g_i(\mathbf{w}) = 0 \qquad i = 1,\mathrm{L}\ ,k$$

$$3) h_i(\mathbf{w}) = 0 \qquad i = 1,\mathrm{L}\ ,m$$

$$4) g_i(\mathbf{w}) \le 0 \qquad i = 1,\mathrm{L}\ ,k$$

$$5) \alpha_i \ge 0 \qquad i = 1,\mathrm{L}\ ,k$$

其中：1)是驻点条件，2)是互补松弛条件，3)、4)是原条件，5)是KKT乘子条件

至此，我们再来回顾一下线性可分条件下的支持向量机方法。

- 首先支持向量机的方法是明确提出一个间隔(margin)概念，并把使间隔最宽作为确定线性分界面的最佳原则。既然是间隔又有线性可分作条件，只需找到处在间隔边缘上的点，以便确定最优的间隔，而其他数据点的作用，只是要求所确定的间隔能保证把它们置在间隔外确定的一方就行。

- 这样一来，数据点就分成两部分，一种对确定间隔参数(体现在权向量**w**的确定很重要)，而另一类(一般说占数据的大部分)对确定间隔的参数没有直接的影响，从这个意义上说它们对确定间隔参数无关紧要。它们相应的拉格朗日乘子α_i是否为0，就表示了数据的这种重要性，*对确定间隔参数主要的**数据点**应有*$\alpha_i \neq 0$*，并称为***支持向量**，而其余的数据点，它的$\alpha_i=0$。

- 一旦使L_D式达到极大值的数据确定下来(只有少量的$\alpha^*_i>0$，其余都为0)，则最优的权向量**w**就可利用下面公式

$$\mathbf{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \mathbf{x}_i \qquad (13)$$

确定下来，它们是这些支持向量数据$\mathbf{x}_i$的线性求和。

设支持向量集用符号SV表示，由于对支持向量$\mathbf{x}_i$而言其$\alpha^*_i>0$，因此 (13)式还可以表示为：

$$\mathbf{w}^* = \sum_{\mathbf{x}_i \in SV} \alpha_i^* y_i \mathbf{x}_i \qquad (13a)$$

b没有出现在对偶问题中，b可由任意一个支持向量$\mathbf{x}_j$用(5)式取等号时求解：

$$b^* = y_j - \mathbf{w}^* \cdot \mathbf{x}_j = y_j - \sum_{i=1}^{N} y_i \alpha_i^* (\mathbf{x}_i \cdot \mathbf{x}_j) \qquad (13b)$$

为减少误差，也可取多个支持向量计算b。如，取2个支持向量，有：

$b^*=-1/2(\mathbf{w}^*\cdot\mathbf{x}_1 + \mathbf{w}^*\cdot\mathbf{x}_2)$

这里假设上式中，$\mathbf{x}_1$和$\mathbf{x}_2$分别表示第1类和第2类的一个支持向量。

$$\text{Maximize} \quad L_D(\boldsymbol{\alpha}) = -\frac{1}{2}\boldsymbol{\alpha}^T D\boldsymbol{\alpha} + \sum_{i=1}^{N}\alpha_i \qquad (10)$$

- 如果知道哪些数据是支持向量，哪些不是，则问题就简单了。问题在于哪些数据是支持向量事先并不能确定，因此这只有通过求(10)式的极大值来求解。
- 对(10)式的来源不必搞懂，只需知道它的极大值解与(9)式的极小值解是一致的即可。
- 求出α、$\mathbf{w}$、b对应的最优解α^*、$\mathbf{w}^*$、b^*后，得到下面最优分类函数：

$$h(\mathbf{x}) = \mathrm{sign}(\mathbf{w}^* \cdot \mathbf{x} + b^*) = \mathrm{sign}(\sum_{i=1}^{N}\alpha_i^* y_i(\mathbf{x}_i \cdot \mathbf{x}) + b^*)$$

(15)

上面公式中，向量$\mathbf{x}$是待分类的测试样本，向量$\mathbf{x}_i$(i=1, 2, …, N)是全部训练样本。注意在(15)式中，测试样本$\mathbf{x}$与训练样本$\mathbf{x}_i$也是以点积形式出现的。

思考：

1.什么是凸集？

2.支持向量机的最佳准则是什么？什么是支持向量？

3. max margin/dis= $\frac{2}{\|\mathbf{w}\|} \Rightarrow \min J(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|^2$

$$s.t.\ \ y_i(\mathbf{w}\cdot\mathbf{x}_i + b) \geq 1 \qquad i = 1, \mathrm{L}\ \ , N$$

4.什么是拉格朗日对偶问题？

5.凸优化问题的KKT(Karuch-Kuhn-Tucher)条件是什么？

参考答案：

1. 答：凸集是指某类点的集合，其中任取两个点连一条直线，这条线上的点仍然在该集合内部。

2. 答：最大间隔准则；支持向量(Support vectors)是指在确定的最大分类间隔边缘的那些训练样本点。

3. 答：由于目标函数是凸的，根据最优化理论，该问题存在唯一全局最小解。

4. 答：通过将Lagrange函数变换成仅包含拉格朗日乘子的函数，称为拉格朗日对偶问题。

5. 答：见课件介绍，共有五条。