支持向量机中的基本概念

张建国

电子与信息工程学院

2011年 10月 14日

提纲

关于学习的几个问题

1. 为什么要学习？
   许多实际任务不能用传统编程技术完成。如手写字母的识别。[Cristianini2000]
2. 什么是学习？
   使用样本综合出计算机程序的过程称为学习方法。
3. 输入输出关系
   样本的输入/输出对通常反映了把输入映射到输出的函数关系。输入与输出间的内在函数称为**目标函数**。
   学习实际上是在一类（组）候选函数（称之为假设集合）中选择一个与目标函数近似的函数作为问题的解。
4. 训练数据如何生成及如何送给学习器
   批量学习与在线学习。

学习结果的评价

1. 学习的目标是给出一个能正确完成训练数据分类的假设（函数）。
2. 要生成可验证的假设存在两个问题。
   - 待学习的目标函数或许没有一个简单的表达，不容易验证。
   - 训练数据通常都是有噪声的。
3. 一个更基本的问题
   即使可以找到一个与训练数据一致的假设函数，对未见数据也可能无法正确分类。一个假设正确分类非训练集数据的能力称为泛化。

提高学习结果的泛化能力

泛化准则对学习算法施加了一种完全不同的约束。

1. 经典的机器学习算法存在的问题
   - 机械式学习（rote learning）
     对未见数据做出根本无关的预测。
   - 过拟合（overfit）
     为了一致性而使假设变得过渡复杂。只有能显著改善分类正确率的复杂性才是值得的。
2. 支持向量机方法通过使用泛化误差的统计边界实现折衷。
   - 缺点，算法不会好于统计结果。
   - 优点，统计结果为算法提供了一个有事实根据的基础，因此能避免启发式方法可能基于错误直觉的危险。
3. 计算复杂度原则分析
   从玩具问题（toy problems）到实际应用（real-world applications）。

学习方法的前景及面临的困难

1. 学习方法的前景是诱人的
   - 可以用此方法解决的问题很多；
   - 可望避免传统解决方法中艰苦的设计与编程；
   - 借此可以洞察人类的内在工作方式。
2. 学习方法固有的困难
   - 输入输出映射函数类的选择；要足够大又不能过大；
   - 存在局部极小时，学习算法是低效的；
   - 输出假设的规模经常大到不切实际；
   - 训练样本有限时，过多的假设类会导致过拟合；
   - 学习算法通常由许多需要试探调整的参数控制，难以可靠使用。

支持向量机

支持向量机是基于泛化理论所提供的洞察力，利用优化理论在核特征空间中有效训练线性学习机的学习系统。系统的重要特征是在确保受泛化理论启发的学习倾向性的同时，给出假设的稀疏对偶表达，从而获得极其有效的算法。另一个重要特征是满足Mercer条件的核对应的优化问题是凸的，因此无局部极小。

Rosenblatt感知器算法—原始形式

给定线性可分的数据集S和步幅$\alpha \in \mathbb{R}^+$
$\mathbf{w}_0 \leftarrow \mathbf{0}; b_0 \leftarrow 0; k \leftarrow 0$
$R \leftarrow \max_{1 \leqslant i \leqslant N} \|\mathbf{x}_i\|$
  重复
    for $i = 1$ to N
      if $d_i \left(\langle \mathbf{w}_k \cdot \mathbf{x}_i \rangle + b_k\right) \leqslant 0$ then
        $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \alpha d_i \mathbf{x}_i$
        $b_{k+1} \leftarrow b_k + \alpha d_i R^2$
        $k \leftarrow k + 1$
      end if
    end for
  直到for循环中没有错误
  返回$(\mathbf{w}_k, b_k)$

间隔（Margin）I

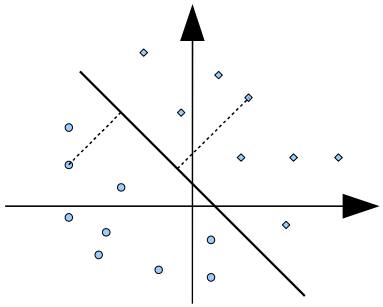
算法的迭代次数与一个称为间隔的量有关。

定义

定义某个样本$(\mathbf{x}_i, d_i)$到超平面$(\mathbf{w}, b)$的（泛函）间隔为：

$$\gamma_i = d_i \left(\langle \mathbf{w} \cdot \mathbf{x}_i \rangle + b\right)$$

- 间隔有正有负
- 正间隔意味着超平面可以正确分类该样本



间隔（Margin）II

间隔分布 超平面$(\mathbf{w}, b)$关于训练集S的间隔分布是S中样本的间隔分布。

间隔 把间隔分布的最小值称为超平面$(\mathbf{w}, b)$关于训练集S的间隔。

最大几何间隔 训练集S的间隔是所有超平面的最大几何间隔。

最大间隔超平面 实现最大间隔的超平面称为最大间隔超平面。

定理

令S为一非平凡训练集，并令$R = \max_{1 \leqslant i \leqslant N} \|\mathbf{x}_i\|$。设存在向量$\mathbf{w}_{opt}$，其$\|\mathbf{w}_{opt}\| = 1$，且对$1 \leqslant i \leqslant N$有：

$$d_i \left(\langle \mathbf{w}_{opt} \cdot \mathbf{x}_i \rangle + b_{opt}\right) \geqslant \gamma$$

则感知器算法的最大误分次数为：$(2R/\gamma)^2$

结果权向量的表示 I

从感知器的学习算法可以看出，如果令初始权向量为零，则学习所得的权向量可以表示为：

$$\mathbf{w} = \sum_{i=1}^{N} \rho_i d_i \mathbf{x}_i.$$

- 样本$\mathbf{x}_i$对应系数的符号由其所属的类别（d_i）确定。
- ρ_i是一个与对应样本被误分次数成正比的值。通常称该量为样本$\mathbf{x}_i$的固有强度（embedding strength）。可以用它根据数据的信息内容对数据进行排序。
- 若样本集S固定，向量$\boldsymbol{\rho}$可以看作是假设在不同（对偶）坐标中的另一种表示。
- 对非线性可分数据，误分点的系数会无限增加。

结果权向量的表示 II

在对偶坐标中，决策函数可以重写为：

$$\begin{aligned} h(\mathbf{x}) &= \operatorname{sgn}(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b) \\ &= \operatorname{sgn}\left(\left\langle \sum_{j=1}^{N} \rho_j d_j \mathbf{x}_j \cdot \mathbf{x} \right\rangle + b\right) \\ &= \operatorname{sgn}\left(\sum_{j=1}^{N} \rho_j d_j \langle \mathbf{x}_j \cdot \mathbf{x} \rangle + b\right) \end{aligned}$$

感知器算法—对偶形式

给定线性可分的数据集S和步幅$\alpha \in \mathbb{R}^+$
$\mathbf{w}_0 \leftarrow \mathbf{0}; b_0 \leftarrow 0; k \leftarrow 0$
$R \leftarrow \max_{1 \leqslant i \leqslant N} \|\mathbf{x}_i\|$
  重复
    for $i = 1$ to N
      if $d_i \left(\sum_{j=1}^{N} \rho_j d_j \langle \mathbf{x}_j \cdot \mathbf{x}_i \rangle + b \right) \leqslant 0$ then
        $\rho_i \leftarrow \rho_i + 1$
        $b \leftarrow b + d_i R^2$
      end if
    end for
  直到for循环中没有错误
  返回$(\boldsymbol{\rho}, b)$

为什么需要特征空间与核函数 I

1. 线性学习机的计算能力有限
   复杂的实际问题需要表达力更强的假设空间，目标通常不能表示为给定属性的简单的线性组合。使用多层网络可以解决此问题。
2. 核函数表示提供了另一种方法
   - 核函数表示通过将数据投影到高维特征空间增加线性学习机的计算能力。
   - 在对偶表示中使用线性学习机使投影过程可以隐式进行。
     使用对偶表示的学习机的优点是可调参数个数与所用的属性个数无关；用合适的核函数代替内积可以隐式地将数据映射到高维空间而不增加可调参数个数。
   - 核函数计算两个输入对应的特征向量的内积。
3. 核方法的引人之处
   学习方法和理论在相当大的程度上与应用的特性无关。为神经网络应用选择结构的问题变成了为支持向量机选择合适核函数的问题。

为什么需要特征空间与核函数 II

4. 核方法的应用范围
   并不仅仅局限于输入空间是欧氏空间的子空间，还可以用于诸如文本这样的离散结构。核方法甚至可以用于无法定义线性函数的输入空间。
5. 维度魔咒（the curse of dimensionality）
   使用核可以克服计算和泛化的维度魔咒。

通过特征映射简化分类任务 I

学习任务的难度
需要学习的目标函数的复杂度取决于它的表示方式。理想地，应选择与特定学习问题匹配的表示。机器学习通常通过预处理改变数据的表示：

$$\mathbf{x} = (x_1 \dots x_n) \mapsto \boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}) \dots \phi_N(\mathbf{x}))$$

例如，对万有引力定律：

$$f(m_1, m_2, r) = C \frac{m_1 m_2}{r^2}$$

如果使用如下坐标变换：

$$(m_1, m_2, r) \mapsto (x, y, z) = (\ln m_1, \ln m_2, \ln r)$$

通过特征映射简化分类任务 II

则表达变为:

$$g(x,y,z) = \ln f(m_1,m_2,r) = \ln C + \ln m_1 + \ln m_2 - 2\ln r = c + x + y + 2z$$

特征 为描述数据引入的量通常称为特征。

属性 数据原始的量称为属性。

特征选择 选择最合适表达式的任务称为特征选择。

输入空间 输入向量所在的空间。

特征空间 特征映射所得特征向量所在的空间。

通过特征映射可以把非线性可分类问题映射成线性可分类问题。

问题是：如何能映射成线性可分的？

特征选择方法 I

1. 维数约减
   通常寻求确定仍能表达原始属性中所含本质信息的最小的特征集合。

$$\mathbf{x} = (p_1^x, p_1^y, p_1^z, p_2^x, p_2^y, p_2^z, m_1, m_2) \mapsto \boldsymbol{\phi}(\mathbf{x}) = \left(\sqrt{\sum_{i\in\{x,y,z\}} (p_1^i - p_2^i)^2}, m_1, m_2 \right).$$

2. 无关特征的检测与消除
   如使用牛顿万有引力定律时，天体的颜色、温度等特征。
3. 主分量分析
   将数据映射到特征空间中，使得新特征是原属性的线性函数且按数据在每个方向上呈现出的方差的大小排序。

特征选择方法 II

- 特征集合越大，使用标准学习机就可能更准确的表示待学习的函数。
- 令人遗憾的是，特征数的增加会增加计算量、降低泛化性能。

使用支持向量机能避免这种性能的下降。

- 理解支持向量机避免泛化性能的下降需要对泛化理论的深入理解。
- 支持向量机通过隐式映射的手段避免特征数增加带来的计算问题。

到特征空间的隐式映射

用一个固定的非线性映射将数据映射到特征空间，然后在特征空间中使用线性学习机。假设集合是如下类型的函数。

$$f(\mathbf{x}) = \sum_{i=1}^{M} w_i \phi_i(\mathbf{x}) + b$$

在对偶表达中，假设用训练点的线性组合表示。因此，决策规则（函数）可以用测试点和训练点的内积计算:

$$f(\mathbf{x}) = \sum_{i=1}^{N} \rho_i d_i \langle \boldsymbol{\phi}(\mathbf{x}_i) \cdot \boldsymbol{\phi}(\mathbf{x}) \rangle + b.$$

如果有办法用原始输入的函数直接计算特征空间中的内积，就可以将两个步骤合并建立一个非线性学习机。称这种直接计算方法为**核函数（方法）**。

核函数的定义 I

定义

核是一个函数K，对所有的$\mathbf{x}, \mathbf{z} \in X$，有：

$$K(\mathbf{x}, \mathbf{z}) = \langle \boldsymbol{\phi}(\mathbf{x}) \cdot \boldsymbol{\phi}(\mathbf{z}) \rangle,$$

其中$\boldsymbol{\phi}$是从输入空间X到（内积）特征空间F的映射。

- 名字“核”取自积分算子理论，该理论是核与其对应的特征空间之间许多关系理论的基础。
- 对偶表达的一个重大意义是特征空间的维数不再影响计算。通过核函数计算内积所需的运算数不再正比于特征数。
- 核的使用回避（绕过）了计算特征映射时固有的计算问题。

核函数的定义 II

- 唯一使用训练样本的信息是样本在特征空间中的Gram矩阵，又称为核矩阵，用$\mathbf{K}$表示。
- 该方法的关键是找一个能有效计算的核函数。

一旦有这么一个函数，判决规则可以通过最多N次核的计算得到：

$$f(\mathbf{x}) = \sum_{i=1}^{N} \rho_i d_i K(\mathbf{x}_i, \mathbf{x}) + b.$$

使用核的一个事实：为了能在特征空间中学习，我们不再需要了解潜在的特征映射。
核的几个例子

$$\langle \mathbf{A}\mathbf{x} \cdot \mathbf{A}\mathbf{z} \rangle; \qquad \langle \mathbf{x} \cdot \mathbf{z} \rangle^2; \qquad (\langle \mathbf{x} \cdot \mathbf{z} \rangle + c)^d.$$

构造核函数的方法

- 使用核函数是有吸引力的计算捷径。
  使用方法？首先构建一个复杂的特征空间；然后在该空间中计算出内积；最后寻找一种能根据原始输入直接计算内积的方法。
- 实际使用的是直接定义核函数，从而隐式定义特征空间。
  采用这种方法，在计算内积或者设计学习器时均避开了特征空间。直接为输入空间定义一个核函数往往比构造一个复杂的特征空间更自然。
- $K(\mathbf{x}, \mathbf{z})$是特征空间中的一个核所必需的性质。

$$K(\mathbf{x}, \mathbf{z}) = K(\mathbf{z}, \mathbf{x}), \qquad \text{对称}$$
$$K(\mathbf{x}, \mathbf{z}) \leqslant K(\mathbf{x}, \mathbf{x}) K(\mathbf{z}, \mathbf{z}). \qquad \text{柯西-施瓦茨不等式}$$

核函数的特性 I

命题

令X是有限输入空间，$K(\mathbf{x}, \mathbf{z})$是$X$上的对称函数。那么$K(\mathbf{x}, \mathbf{z})$是核函数的充分必要条件是核矩阵

$$\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^{n}$$

是半正定的（即特征值非负）。

Mercer定理给出了满足下列表达

$$K(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{\infty} \lambda_i \phi_i(\mathbf{x}) \phi_i(\mathbf{z}) \quad \lambda_i \text{非负}$$

的连续对称函数是特征空间中的内积的充分必要条件。

核函数的特性 II

如此通过核函数隐式引入了一个由特征向量定义的空间。该空间中的线性函数可以如前所述表示为:

$$f(\mathbf{x}) = \sum_{i=1}^{\infty} \lambda_i \psi_i \phi_i(\mathbf{x}) + b = \sum_{j=1}^{N} \alpha_j d_j K(\mathbf{x}, \mathbf{x}_j) + b,$$

其中第一种表达是原始形式，第二种是对偶形式。

- 原始表达的求和项数等于特征空间的维数，而对偶表达中的求和项数是样本的个数。哪种表达更方便取决于所考虑的特征空间的维数大小。
- 从该表达形式可以看出，只要核函数是非线性的，该函数就是非线性的。
- 验证一个对称函数是核的关键是，在任意有限点集上由该对称函数所定义的矩阵是半正定的。

用核函数构造核函数 I

命题

令K_1与K_2是$X \times X$上的核，$X \subseteq \mathbb{R}^n$，$a \in \mathbb{R}^+$，$f(\cdot)$是X上的实值函数，$\boldsymbol{\phi} : X \longrightarrow \mathbb{R}^m$ 且K_3是$\mathbb{R}^m \times \mathbb{R}^m$上的核，$\mathbf{B}$ 是一个对称半正定$n \times n$矩阵。则下面的函数是核。

$$K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z}) + K_2(\mathbf{x}, \mathbf{z})$$
$$K(\mathbf{x}, \mathbf{z}) = aK_1(\mathbf{x}, \mathbf{z})$$
$$K(\mathbf{x}, \mathbf{z}) = K_1(\mathbf{x}, \mathbf{z})K_2(\mathbf{x}, \mathbf{z})$$
$$K(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})f(\mathbf{z})$$
$$K(\mathbf{x}, \mathbf{z}) = K_3(\boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{z}))$$
$$K(\mathbf{x}, \mathbf{z}) = \mathbf{x}'\mathbf{B}\mathbf{z}$$

用核函数构造核函数 II

推论

令$K_1(\mathbf{x}, \mathbf{z})$是$X \times X$上的核，$\mathbf{x}, \mathbf{z} \in X$ ，$p(x)$是正系数多项式。则下面的函数也是核。

$$K(\mathbf{x}, \mathbf{z}) = p\left(K_1(\mathbf{x}, \mathbf{z})\right)$$
$$K(\mathbf{x}, \mathbf{z}) = \exp\left(K_1(\mathbf{x}, \mathbf{z})\right)$$
$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\|\mathbf{x} - \mathbf{z}\|^2/\sigma^2\right)$$

为什么需要泛化理论?

- 核的引入极大地增加了学习机的表达能力，它保持了内在线性性从而使学习易于控制。
- 适应性的增加增大了过拟合的风险，因为自由度的增加增大了间隔超平面选择的不适定性。
- 成功控制核特征空间增加的适应性需要一套完善的泛化理论，它能精确描述控制学习机中的哪个因子才能保证好的泛化能力。
- VC理论促进了SVM的出现，在已有的几中理论中用它描述SVM是最合适的。
- VC理论的主要结论给出了线性分类器泛化性的可靠界，由此指出了如何控制核空间中线性函数的复杂度。

引言

- 由泛化理论，假设函数应选为最小或最大化某种泛函。对线性学习机LLM，这相当于在一些约束条件下寻找最小（最大）化特定代价函数的参数向量。
- 优化理论描述此类问题的解并研究找到解的有效算法。
- 根据特定的代价函数及约束的性质，可以分出许多种已被深入研究且存在有效解法的问题。
- 凸二次规划问题，代价函数是凸二次函数，约束是线性的。
- 优化理论不仅提供算法技术，而且定义了给定函数是解的充要条件。对偶理论就是一个例子，它给出了LLM 对偶表达的自然解释。

一般优化问题 I

定义

（原始优化问题）给定在域$\mathbf{\Omega} \subseteq \mathbb{R}^n$上定义的函数$f$，$g_i$，$i = 1, \ldots, k$，及$h_i$，$i = 1, \ldots, m$，

$$\begin{array}{lll} \textit{minimise} & f(\mathbf{w}), & \mathbf{w} \in \mathbf{\Omega}, \\ \textit{subject to} & g_i(\mathbf{w}) \leqslant 0, & i = 1, \ldots, k, \\ & h_i(\mathbf{w}) = 0, & i = 1, \ldots, m \end{array}$$

其中，$f(\mathbf{w})$ 称为目标函数，余下的关系分别称为不等式及等式约束。

简化表示：

用$\mathbf{g}(\mathbf{w}) \leqslant \mathbf{0}$表示$g_i(\mathbf{w}) \leqslant 0, i = 1, \ldots, k$。$\mathbf{h}(\mathbf{w}) = \mathbf{0}$与此同理。

一般优化问题 II

目标函数的定义域与满足约束条件的区域的交集称为**可行区域**，表示如下：

$$R = \{\mathbf{w} \in \mathbf{\Omega} : \mathbf{g}(\mathbf{w}) \leqslant 0, \mathbf{h}(\mathbf{w}) = 0\}$$

点$\mathbf{w}^* \in R$称为最优化问题的解，如果$\forall \mathbf{w} \in R$，当$\mathbf{w} \neq \mathbf{w}^*$时，有$f(\mathbf{w}) < f(\mathbf{w}^*)$。（局部极小）

目标函数、等式以及不等式约束均是线性函数的优化问题称为**线性规划**。目标函数是二次函数，约束是线性函数的优化问题成为**二次规划**。

通过引入松弛变量ξ，可以将不等式约束转换为等式约束，如：

$$g_i(\mathbf{w}) \leqslant 0 \Longleftrightarrow g_i(\mathbf{w}) + \xi_i = 0, \text{其中} \xi_i \geqslant 0.$$

一般优化问题 III

一个实值函数$f(\mathbf{w})$称为$\mathbf{w} \in \mathbb{R}^n$上的凸函数，如果$\forall \mathbf{w}, \mathbf{u} \in \mathbb{R}^n$，对任意$\theta \in (0, 1)$，有

$$f(\theta\mathbf{w} + (1-\theta)\mathbf{u}) \leqslant \theta f(\mathbf{w}) + (1-\theta) f(\mathbf{u}).$$

如果不等号严格成立，则称为严格凸。

如果一个二次可导函数的Hessian矩阵是半正定的，则它是凸的。

仿射函数是指可以用某个矩阵$\mathbf{A}$和向量$\mathbf{b}$表示为$f(\mathbf{w}) = \mathbf{A}\mathbf{w} + \mathbf{b}$形式的函数。仿射函数具有零Hessian矩阵，因此它是凸的。

对$\mathbf{\Omega} \subseteq \mathbb{R}^n$，如果$\forall \mathbf{w}, \mathbf{u} \in \mathbf{\Omega}, \forall \theta \in (0, 1)$，均有：

$$(\theta\mathbf{w} + (1-\theta)\mathbf{u}) \in \mathbf{\Omega},$$

则称该集合是凸的。

一般优化问题 IV

- 如果无约束优化问题的目标函数是凸的，那么局部极小点$\mathbf{w}^*$也是全局最小点。凸函数的这个性质使得，如果优化问题的函数与集合是凸的，则该问题是可解的。
- 如果一个最优化问题的集合$\mathbf{\Omega}$，目标函数以及所有约束都是凸的，则称该问题是凸的。
- 对训练支持向量机来说，我们可以通过限制使得约束是线性的，目标函数是凸的和二次的，且$\mathbf{\Omega} \in \mathbb{R}^n$。因此，我们讨论凸二次规划。

无约束优化问题

拉格朗日理论的最初的目的是刻画只有等式约束的优化问题的解。该理论的主要概念是拉格朗日乘数和拉格朗日函数。

拉格朗日理论（1797年）是对Fermat的研究结果（1629年）进行推广得到的。1951年，Kuhn和Tucker又将该理论推广到有不等式约束的情况。

定理

$\mathbf{w}^*$是函数$f(\mathbf{w}), f \in C^1$的最小值的必要条件是$\partial f(\mathbf{w}^*)/\partial \mathbf{w} = \mathbf{0}$。如果函数$f$是凸的，那么该条件也是充分条件。

对约束问题，需要定义拉格朗日函数，它是目标函数加上约束的线性组合。

等式约束优化问题–拉格朗日乘数法 I

定义

设给定优化问题的目标函数为$f(\mathbf{w})$，等式约束$h_i(\mathbf{w}) = 0, i = 1, \ldots, m$，拉格朗日函数定义为：

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

其中系数β_i称为拉格朗日乘数。

当有等式约束时，为了遵守某个约束，必须在垂直于约束法线（$\partial h_i(\mathbf{w})/\mathbf{w}$）的方向上移动。如果是多个约束，为了遵守所有等式约束，必须在垂直于所有约束法线张成的子空间上移动。即：$\partial f(\mathbf{w}^*)/\partial \mathbf{w} + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w}^*) = 0.$

等式约束优化问题–拉格朗日乘数法 II

定理

点$\mathbf{w}^*$是函数$f(\mathbf{w})$在$h_i(\mathbf{w}) = 0, i = 1, \ldots, m$约束下的极小值的必要条件是：对某些$\boldsymbol{\beta}^*$

$$\frac{\partial L(\mathbf{w}^*, \boldsymbol{\beta}^*)}{\partial \mathbf{w}} = \mathbf{0},$$
$$\frac{\partial L(\mathbf{w}^*, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

其中，$f, h_i \in C^1$。如果$L(\mathbf{w}, \boldsymbol{\beta}^*)$是$\mathbf{w}$的凸函数，那么上述条件也是充分的。

由于约束等于零，拉格朗日函数在最优点的值等于目标函数的值，即$L(\mathbf{w}^*, \boldsymbol{\beta}^*) = f(\mathbf{w}^*)$

带不等式约束的优化问题–广义拉格朗日函数

定义

给定域$\mathbf{\Omega} \subseteq \mathbb{R}^n$上的优化问题，

$$\begin{array}{lll} \textit{minimise} & f(\mathbf{w}), & \mathbf{w} \in \mathbf{\Omega}, \\ \textit{subject to} & g_i(\mathbf{w}) \leqslant 0, & i = 1, \ldots, k, \\ & h_i(\mathbf{w}) = 0, & i = 1, \ldots, m \end{array}$$

广义拉格朗日函数定义为：

$$\begin{aligned} L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) &= f(\mathbf{w}) + \sum_{i=1}^{k} \alpha_i g_i(\mathbf{w}) + \sum_{i=1}^{m} \beta_i h_i(\mathbf{w}) \\ &= f(\mathbf{w}) + \boldsymbol{\alpha}' \mathbf{g}(\mathbf{w}) + \boldsymbol{\beta}' \mathbf{h}(\mathbf{w}). \end{aligned}$$

带不等式约束的优化问题–Kuhn-Tucker定理

定理

给定一个凸域$\mathbf{\Omega} \subseteq \mathbb{R}^n$上的优化问题，

$$\begin{array}{lll} \textit{minimise} & f(\mathbf{w}), & \mathbf{w} \in \mathbf{\Omega}, \\ \textit{subject to} & g_i(\mathbf{w}) \leqslant 0, & i = 1, \ldots, k, \\ & h_i(\mathbf{w}) = 0, & i = 1, \ldots, m \end{array}$$

其中，$f \in C^1$是凸的，且g_i, h_i是仿射函数。则点$\mathbf{w}^*$是最优点的充要条件是存在$\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$满足：

$$\begin{aligned} \frac{\partial L(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \mathbf{w}} &= \mathbf{0}, \\ \frac{\partial L(\mathbf{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}} &= \mathbf{0}, \end{aligned} \qquad \begin{aligned} \alpha_i^* g_i(\mathbf{w}^*) &= 0, i = 1, \ldots, k, \\ g_i(\mathbf{w}^*) &\leqslant 0, i = 1, \ldots, k, \\ \alpha_i^* &\geqslant 0, i = 1, \ldots, k. \end{aligned}$$

互补条件

第三个关系称为Karush-Kunhn-Tucker互补条件（KKT条件）。它意味着对积极约束有$\alpha_i^* \geqslant 0$，而对非积极约束有$\alpha_i^* = 0$。

这说明对不等式约束，最优解要么位于边界上，要么位于约束区域的内部。对第一种情况，该不等式是积极约束，可以看作一个等式约束；对第二种情况，该不等式的约束是非积极的，相当于该约束该优化问题没有影响，因此对应的拉格朗日乘数为0。

具有不等式约束的优化问题存在拉格朗日对偶问题，它为线性学习机LLM的对偶表达提供了自然解释。

拉格朗日对偶问题

定义

原始优化问题的拉格朗日对偶问题为：

$$\begin{array}{ll} \textit{maximise} & \theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) \\ \textit{subject to} & \boldsymbol{\alpha} \geqslant \mathbf{0} \end{array}$$

其中$\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \inf_{\mathbf{w} \in \mathbf{\Omega}} L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta})$。目标函数在最优解处的值称为问题的解。

下面的弱对偶定理给出了原始优化问题与对偶优化问题间的基本关系。

强对偶定理说明，满足一定条件时原始优化问题与对偶问题具有相同的值。

弱对偶定理 I

定理

设$\mathbf{w} \in \mathbf{\Omega}$是原始优化问题的可行解，$(\boldsymbol{\alpha}, \boldsymbol{\beta})$是对偶优化问题的可行解，则$f(\mathbf{w}) \geqslant \theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$。

证明： 由$\theta(\boldsymbol{\alpha}, \boldsymbol{\beta})$的定义，对$\mathbf{w} \in \mathbf{\Omega}$，有

$$\begin{aligned}\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \inf_{\mathbf{u} \in \mathbf{\Omega}} L(\mathbf{u}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &\leqslant L(\mathbf{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \\ &= f(\mathbf{w}) + \boldsymbol{\alpha}'\mathbf{g}(\mathbf{w}) + \boldsymbol{\beta}'\mathbf{h}(\mathbf{w}) \leqslant f(\mathbf{w}).\end{aligned}$$

因为$\mathbf{w}$)的可行意味着：$\mathbf{g}(\mathbf{w}) \leqslant \mathbf{0}$且$\mathbf{h}(\mathbf{w}) = \mathbf{0}$，而$(\boldsymbol{\alpha}, \boldsymbol{\beta})$的可行意味着$\boldsymbol{\alpha} \geqslant \mathbf{0}$。

弱对偶定理 II

推论

对偶优化问题值的上界是原始优化问题的值，即

$$\sup\{\theta(\boldsymbol{\alpha}, \boldsymbol{\beta}) : \boldsymbol{\alpha} \geqslant \mathbf{0}\} \leqslant \inf\{f(\mathbf{w}) : \mathbf{g}(\mathbf{w}) \leqslant \mathbf{0}, \mathbf{h}(\mathbf{w}) = \mathbf{0}\}.$$

原始问题与对偶问题值之间的差称为对偶间隙（duality gap）。

推论

如果$f(\mathbf{w}^*) = \theta(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$，其中$\boldsymbol{\alpha}^* \geqslant \mathbf{0}$，且$\mathbf{g}(\mathbf{w}^*) \leqslant \mathbf{0}, \mathbf{h}(\mathbf{w}^*) = \mathbf{0}$，则$\mathbf{w}^*$和$\theta(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ 分别是原始优化问题和对偶优化问题的解。在此情况下$\alpha_i^* g_i(\mathbf{w}^*) = 0, i = 1, \ldots, k$。

强对偶定理

定理

给定一个凸域$\mathbf{\Omega} \subseteq \mathbb{R}^n$上的优化问题，

$$\begin{array}{lll} \textit{minimise} & f(\mathbf{w}), & \mathbf{w} \in \mathbf{\Omega}, \\ \textit{subject to} & g_i(\mathbf{w}) \leqslant 0, & i = 1, \ldots, k, \\ & h_i(\mathbf{w}) = 0, & i = 1, \ldots, m \end{array}$$

其中g_i, h_i是仿射函数，即存在矩阵$\mathbf{A}$与向量$\mathbf{b}$，使得：

$$\mathbf{h}(\mathbf{w}) = \mathbf{A}\mathbf{w} - \mathbf{b},$$

在此条件下，对偶间隙为零。

参考文献

C.-C. Chang and C.-J. Lin.
LIBSVM: a library for support vector machines.
http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001.

Nello Cristianini and John Shawe-Taylor.
An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.
Cambridge University Press, 2000.

谢谢！