斯坦福 ML 公开课 7

本篇笔记针对 ML 公开课的第七个视频，主要内容包括最优间隔分类器（Optimal Margin Classifier）、原始/对偶问题（Primal/Dual Problem）、svm 的对偶问题，都是 svm（support vector machine，支持向量机）的内容。

在上篇笔记中，我们提到了函数间隔与几何间隔，这两个定义是 svm 的基本定义，因为 svm 是比较复杂的模型，公开课横跨了三个视频才将其介绍完。这里先简要说明一下理解 svm 的必要的几个部分，使读者有个宏观的概念。首先是函数间隔与几何间隔，由它们引出最优间隔分类器；为了多快好的解决最优间隔分类器问题，使用了拉格朗日对偶性性质，于是，先要理解原始优化问题与对偶问题及它们在什么条件（KKT 条件）下最优解等价，然后写出最优间隔分类器的对偶形式；通过对最有间隔分类器对偶问题求解，发现求解时目标函数中存在内积形式的计算，据此引入了核技法；引入核技法后就得到了完完全全的 svm 求解问题，使用序列最小化算法（SMO）进行求解，这就是公开课对 svm 介绍的全部内容，读者按照先后顺序一一理解即可快速理解 svm。

最优间隔分类器

在开始之前，仍然要强调一下本篇所讲的内容仍然是假设数据集是线性可分的。

首先，回顾一下讲述函数间隔时对目标函数的表示方法所做的变化：

类别 y 可取值由{0,1}变为{-1,1}，假设函数变为：

$$g(z)=\begin{cases}-1 & z\le 0\\ 1 & z>0\end{cases} \tag{1}$$

$$h_{w,b}(x)=g(w^Tx+b) \tag{2}$$

由公式 2，我们得知，w,b 可以唯一的确定一个超平面。

回顾一下上篇笔记中介绍的函数间隔的缺点，只要成倍的增大 w,b，就可以使函数间隔变大。而几何间隔不会遇到这个问题，究其原因，是成倍增大 w,b 后，决策面的位置不会发生改变。本节会利用这个性质，对 w,b 进行缩放，从而简化问题。

最优间隔分类器（optimal margin classifier），是指在对数据分类时，得到的决策面的一个性质，即决策面距离数据点的几何间隔最大。可以使用置信度对它来进行解释，对于线性可分数据，我们可以得到无数个决策面，直观上看，数据点距离决策面越远，决策面对数据点的预测可信度就越高。最优间隔分类器即是寻找一个决策面，使之对数据点的预测的置信度达到最高。

使用数学语言对最优间隔分类器进行表示，即 #1：

$$\begin{aligned}max_{\gamma,w,b}\quad & \gamma\\ \text{s.t.}\quad & y^{(i)}\left(w^Tx^{(i)}+b\right)\ge\gamma,\ \ i=1,2,\dots,m\\ & ||w||=1\end{aligned}$$

其中，||w||=1 保证了目标值是几何间隔。#1 的含义是通过改变 w,b，寻找一个最大的γ值，使得对于训练集中所有的点，点到决策面的几何距离都大于γ。

该问题不易解决，因为约束是非凸性约束，最优解容易达到局部最优。于是，我们对该问题进行转换，得到 #2：

$$\begin{aligned}max_{\gamma,w,b}\quad & \frac{\gamma}{||w||}\\ \text{s.t.}\quad & y^{(i)}\left(\frac{w^T}{||w||}x^{(i)}+\frac{b}{||w||}\right)\ge\frac{\gamma}{||w||},\ \ i=1,2,\dots,m\\ \text{s.t.}\quad & y^{(i)}\left(w^Tx^{(i)}+b\right)\ge\gamma,\ \ i=1,2,\dots,m\end{aligned}$$

#2 与#1 描述的是同一个问题，即寻找一个最大的值，使得训练集中所有的点到决策面

的几何距离都大于该值。#2 表述中删除部分是省略掉的推导过程，即不等式两边都乘以||w||。

#2 通过将非凸性的约束条件转移到目标函数中，是问题变成凸性问题。

对于#2 来说，还可以再做一次变换，使之更为简单。我们知道，等比例对 w,b 进行缩放，不会改变决策面的位置。假设已经得到 w,b，那么就能求出γ的值；那么我们可以通过缩放 w,b(同时除以γ)，使γ值变为 1；这样得到的决策面与开始时就将γ设为 1 是一样一样的；于是，得到更简单的问题，#3：

$$max_{\gamma,w,b} \quad \frac{1}{||w||}$$

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

$$\text{s.t.} \quad y^{(i)}\left(w^T x^{(i)} + b\right) \geq 1, \quad i = 1,2,\dots,m$$

同#2 一样，删除部分为推导过程，对$\frac{1}{||w||}$求极大与对$\frac{1}{2}||w||^2$求极小是等同的。

#3 的表述就已经是凸性问题了。为了更好的解决该问题，需要使用它的对偶问题。下面首先介绍原始问题与对偶问题的概念。

原始/对偶优化问题

回想当年高数课上的拉格朗日（Lagrangian）方程，它用于求解这样的问题：

$$min_w \quad f(w)$$
$$s.t. \ h_i(w) = 0, i = 1,2,\dots,l$$

我们构造拉格朗日方程：

$$L(w,\beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w) \tag{3}$$

公式 3 中，β是拉格朗日乘子。

构造拉格朗日方程后，我们对其求偏导数，将偏导数设为 0，如公式 4，求得的值就是原问题的解了。

$$\frac{\partial L}{\partial w_i} = 0; \frac{\partial L}{\partial \beta_i} = 0 \tag{4}$$

这是比较标准的拉格朗日方程的应用，下面我们将通过扩展约束条件，介绍更为广义的拉格朗日方程。在约束条件中添加不等式约束后，我们得到如下问题，称之为原始优化问题，即#4：

$$min_w \quad f(w)$$
$$s.t. \ h_i(w) = 0, i = 1,2,\dots,l$$
$$g_i(w) \leq 0, i = 1,2,\dots,k$$

该问题对应的广义拉格朗日方程是：

$$L(w,\alpha,\beta) = f(w) + \sum_{i=1}^{l} \beta_i h_i(w) + \sum_{i=1}^{k} \alpha_i g_i(w) \tag{5}$$

公式 5 中，α,β是拉格朗日乘子。

考虑公式 6 的形式：

$$\theta_p(w) = \max_{\alpha,\beta:\alpha_i>0} f(w) + \sum_{i=1}^{l} \beta_i h_i(w) + \sum_{i=1}^{k} \alpha_i g_i(w) \tag{6}$$

其中，θ下标 p 代表原始问题。可以发现，在给定 w 时，对 L(w,α,β)求极大时，当 w

不满足所有约束条件，比如$h_i(w) \neq 0$，或者$g_i(w) > 0$，总可以找到相应的α, β，使$\theta_p(w) = \infty$。因而就有了如下的结论：

$$\theta_p(w) = \begin{cases} f(w) & \text{所有约束都满足} \\ \infty & \text{否则} \end{cases} \tag{7}$$

因而，我们可以认为，$\theta_p(w)$即是将约束条件与目标函数融合在一起的表述方法，考虑#4 中的目标函数优化，我们得到如下公式：

$$\min_{\mathrm{w}} \theta_p(w) = \min_{\mathrm{w}} \max_{\alpha,\beta:\alpha_i>0} L(w,\alpha,\beta) \tag{8}$$

公式 8 即是原始问题的最终表述方法，与#4 的表述完全等价。令：

$$p^* = min_w \theta_p(w) \tag{9}$$

p*为原始问题取得最优解时的函数值。

定义：

$$\theta_D(\alpha,\beta) = \min_w L(w,\alpha,\beta) \tag{10}$$

其中，d 代表对偶问题。这样，可以得到原始问题的对偶问题的定义：

$$\max_{\alpha,\beta:\alpha_i>0} \theta_D(\alpha,\beta) = \max_{\alpha,\beta:\alpha_i>0} \min_w L(w,\alpha,\beta) \tag{11}$$

对比公式 11 与公式 8，我们发现这两个很相似，只是 max 和 min 调换了位置，这也是它们被称为对偶问题的原因吧，再令：

$$d^* = max_{\alpha,\beta:\alpha_i>0} \theta_D(\alpha,\beta) \tag{12}$$

即 d*是对偶问题取得最优解时的函数值。

写了那么多公式，它们是干什么用的呢？从公式 3 到公式 12，先介绍了基本的拉格朗日方程，然后再介绍广义的拉格朗日方程，在广义的拉格朗日方程中，定义了原始最优化问题与对偶最优化问题，在下面会讲到在符合某些条件的时候，原始最优化问题与对偶最优化问题可以取得相同的最优解，从而使得在原始问题上比较难求解的问题可以转移到对偶问题上去求，最有间隔分类器正是这样一种问题。下面，介绍一下原始问题与对偶问题的关系及何种条件下等价，所述皆是结论，我们只是使用结论，因而不做证明深究了。

一般情况下，d*与 p*的关系是：

$$d^* \leq p^*$$

在一些条件下，d*可以等于 p*，首先，先做一些假设，假设约束不等式 g_i 都是凸（convex）函数（线性函数都属于凸函数），约束等式 h_i 都是仿射（affine）函数（仿射函数定义：$h(w)=w^Tx+b$，仿射几乎和线性等价，只不过允许截距 b 的存在）。在假设不等式约束条件是严格可执行的，即存在 i 使得 $g_i(w)<0$。

在这些假设下，肯定存在w^*, α^*, β^*，使得w^*是原始问题的解，α^*, β^*是对偶问题的解，且$p^* = d^* = L(w^*, \alpha^*, \beta^*)$。这样的$w^*, \alpha^*, \beta^*$需要满足 KKT（Karush-Kuhn-Tucker）条件，KKT 条件如下：

$$\frac{\partial}{\partial w_i} L(w^*, \alpha^*, \beta^*) = 0, i = 1,2,\dots,n \tag{13}$$

$$\frac{\partial}{\partial \beta_i} L(w^*, \alpha^*, \beta^*) = 0, i = 1,2,\dots,l \tag{14}$$

$$\alpha_i^* g_i(w^*) = 0, i = 1,2,\dots,k \tag{15}$$

$$g_i(w^*) \leq 0, i = 1,2,\dots,k \tag{16}$$

$$\alpha^* \geq 0, i = 1,2,\dots,k \tag{17}$$

这里关注公式 15，它被称为 KKT 互补条件。即当α_i^*不为 0 时，$g_i(w^*)=0$，即该条件被激活（达到临界条件）。这个条件比较重要是因为在后面，它将展示出 svm 只有一些支持向量点会起作用，在 SMO 算法中会给出收敛测试。

最优间隔分类器的求解

上面讲述的原始/对偶优化问题（primal/dual optimal problem），其目的在于对在原始问题上不易求解的问题进行变换，使之更易求解。

下面介绍通过对最优间隔分类器的对偶问题进行求解，得到的简化后问题的过程。

根据最优间隔分类器#3 的表述，我们将其变换为广义拉格朗日格式的问题，即#5：

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

$$\text{s.t.} \quad g_i(w) = -y^{(i)}\left(w^T x^{(i)} + b\right) + 1 \le 0, \ i = 1,2,\dots,m$$

该问题对应的拉格朗日方程是：

$$\mathrm{L(w,b,\alpha)} = \frac{1}{2}||w||^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}\left(w^T x^{(i)} + b\right) - 1] \tag{18}$$

注意到，#5 问题只有不等式约束，没有等式约束，所以拉格朗日乘子只有α。由#5 知，该问题符合 d*=p*的假设，肯定存在w^*, α^*, β^*是原始问题和对偶问题共有的最优解。

求解对偶问题时，首先要固定α，以 w,b 为变量，最小化 L；最小化 L 时，求解 L 对 w 和 b 的偏导，并将导数设为 0，可以得到：

$$\nabla_w L(\mathrm{w,b,\alpha}) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0 \Longrightarrow w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \tag{19}$$

$$\frac{\partial}{\partial b} L(\mathrm{w,b,\alpha}) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \tag{20}$$

将公式 19 和公式 20 代入到公式 18，可以推导出一种更简单的形式，从该形式可以引入核技法，下面是推导过程：

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

$$= \frac{1}{2} w^T w - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} b + \sum_{i=1}^{m} \alpha_i$$

$$= \frac{1}{2} w^T \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} - \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)} + \sum_{i=1}^{m} \alpha_i$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i y^{(i)} w^T x^{(i)}$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \alpha_i y^{(i)} \left(\sum_{j=1}^{m} \alpha_j y^{(j)} (x^{(j)})^T \right) x^{(i)}$$

$$= \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(j)})^T x^{(i)} \quad (21)$$

其中，第一步是原问题，第二部将累加和展开，第三步代入公式 19 和 20；第四步合并系数，第五步代入公式 19，第六步展开。

原问题针对参数 w,b 上做了最小化操作后，就要针对参数α做最大化操作。将对偶问题中存在的$\alpha_i \geq 0$的约束条件和公式 20 作为最大化操作的约束条件，我们得到经过对偶化后的简化问题如下，即#6：

$$\max_{\alpha} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle x^{(j)}, x^{(i)} \rangle$$

$$\text{s.t.} \quad \alpha_i \geq 0, \mathrm{i} = 1,2,\dots,\mathrm{m}$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

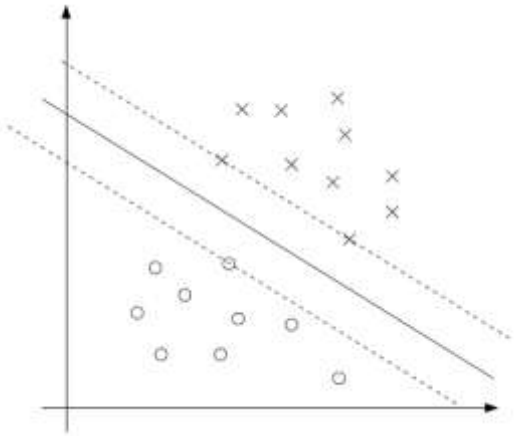
其中，目标函数中的$\langle x^{(j)}, x^{(i)} \rangle$为内积。

对于该优化问题，与原始优化问题对比，我们可以验证它符合 KKT 条件。可以逐一验证，对于 KKT 条件的五个要求来说，公式 13 在最小化操作时满足；没有出现β乘子，所以公式 14 可以忽略；公式 17 仍然是对偶优化问题的约束；可以用反证法证明公式 15 和公式 16 已经满足，如果公式 16 没有满足，那么对偶优化就会得到正无穷，如果公式 15 没有满足，那么必然有 $\alpha_i^* g_i(w^*) > 0$，会导致在极小的时候得到负无穷小。

在求得最优解 w*后，可以得到 b*的解：

$$b^* = -\frac{max_{i:y^{(i)}=-1}(w^*)^T x^{(i)} + min_{i:y^{(i)}=1}(w^*)^T x^{(i)}}{2} \quad (22)$$

公式 22 表明，这是确定w^*后，正例和负例中的支持向量所对应的截距的平均值。为了更直观的理解该问题，可以看下图：



公式 22 即是两条虚线与纵轴的截距的平均值的含义。

考虑公式 15，对于$g_i(w^*)$不为 0 时，即为虚线以外的点，此时α_i^*为 0，即此点在目标函数的计算中不作出贡献。反之，则在虚线上，α_i^*不为 0，为目标函数的计算做出贡献。

那些为目标函数的计算做出贡献的点，称为支持向量。

下一个视频中将引入核技法到#6 中的目标函数中，从而得到完全的支持向量机算法，然后介绍 SMO（序列最小化算法），该算法是优化问题的一种较快的解决方法。