Constrained Optimization and Support Vector Machines

http://freemind.pluskid.org/machine-learning/constrained-optimization-and-support-vector-machines

posted on Free Mind on June 22, 2014
generated with pandoc on December 3, 2015
category: Machine Learning

tags: Optimization

之前我们一直在讨论无约束的优化问题，定义域就是整个欧氏空间，不过在实际问题中我们经常需要解决带约束的优化问题。今天我们就来对带约束的优化问题做一个初步的介绍。在这里我们仍然假设所涉及的函数都是可导的，非光滑的优化问题将留到后面的篇幅来介绍。

带约束的优化问题的一般形式是

$$\min_{x \in X} f(x)$$

其中 X 是我们允许的的取值范围。一般情况下我们会要求 X 是一个 Convex Set，这样的情况下优化通常会容易一点，但是需要注意的是在实际问题中碰到 X 是 non-Convex 的情况也是非常多的，并且有些 non-Convex 的情况也有很简单的解。其中一个比较典型的例子是我们在 Principal Component Analysis (PCA) 里碰到的一个优化问题：

$$\min_{X_r} \|X - X_r\|_F^2, \quad s.t. \text{rank}(X_r) \leq r$$

也就是要求一个矩阵 X 的 low-rank 近似。其中 $\|X\|_F = \sqrt{\text{Tr}(X^T X)} = \sqrt{\sum_{i,j} X_{ij}^2}$ 是矩阵的 Frobenius Norm。首先这个优化问题的约束集合 $\{X_r : \text{rank}(X_r) \leq r\}$ 显然是 non-Convex 的，然而我们知道该问题的解可以通过 Singular Value Decomposition (SVD) 来求得。具体来说，令 $X = U\Sigma V^T$ 是 X 的 SVD 分解，其中 Σ 是对角矩阵，其对角线上的元素称为 X 的奇异值 (Singular Value)，而 U 和 V 分别是正交矩阵，令 Σ_r 是将 Σ 对角线上的最大的 r 个值（奇异值）以外的元素置为零得到的矩阵，则 $X_r^* = U\Sigma_r V^T$ 是该问题的最优解。并且当第 r 个奇异值和第 $r+1$ 个奇异值不相等时，该最优解是唯一的。

这里可以简单解释一下，注意到乘以一个正交矩阵只是进行基的旋转，并不改变矩阵的 Frobenius Norm，于是我们有

$$\begin{aligned} \|X - X_r\|_F^2 &= \|U\Sigma V^T - X_r\|_F^2 \\ &= \|U^T U\Sigma V^T V - U^T X_r V\|_F^2 \\ &= \|\Sigma - U^T X_r V\|_F^2 \end{aligned}$$

注意到 $U^T X_r V$ 仍然是一个 r 阶矩阵，现在问题变成用一个 r 阶矩阵去近似一个对角矩阵，而 Frobenius Norm 又是简单的所有元素的平方

和，所以保留最大的 r 个对角元素是最优的解。对严格的证明感兴趣的同学可以参见 [Stewart, 1998] 的定理 4.32 (Schmidt-Mirsky)。

总而言之这里这个例子要说明的是，虽然我们会将精力主要集中在 convex set constrained 优化问题上，但是并不代表 non-convex set constrained 优化问题在实际中不会出现或者不重要，更不代表这样的问题就是没有全局最优解或者是不能很有效地求解的。

接下来让我们再回到优化问题上，一般情况下，$x \in X$ 这个约束条件本身也是由一些函数方程和不等式来描述的，此时优化问题一般写作如下形式：

$$\begin{aligned} \min_x \ & f(x) \\ s.t. \ & g_i(x) \leq 0, \quad i = 1, \dots, m \\ & h_i(x) = 0, \quad i = 1, \dots, l \end{aligned}$$

当 f 和 g_i 是 convex 且 h_i 是线性函数时这是一个 convex optimization 问题——也就是说目标函数和约束集都是 convex 的。注意当 g_i 是 convex 的时候，sub-levelset $\{x : g_i(x) \leq 0\}$ 是一个凸集，但是 h_i 是 convex 的时候并不能保证 $\{x : h_i(x) = 0\}$ 是凸集，所以需要更强的条件，要求 h_i 是线性的，此时 $\{x : h_i(x) = 0\}$ 是一个线性子空间，显然是凸的。

作为一个具体的机器学习中的例子，让我们来考虑支持向量机 Support Vector Machines (SVM)。我在之前写过一个比较详细的支持向量机系列的介绍文章，不过这里让我们不妨来简单回顾一下。

问题的背景是有一系列数据点 $x_1, \dots, x_N$，每个点有一个二元类别 $y_i \in \{+1, -1\}$，我们想要寻找一个超平面 $w^T x = b$ 将这两类点分开。除此之外，我们还希望 margin 最大化，也就是说让所有数据点 $\{x_i\}_{i=1}^N$ 到超平面的距离最大化。注意到当 $y_i = 1$ 时数据点 x_i 到超平面 $w^T x - b = 0$ 的距离为

$$\frac{w^T x_i - b}{\|w\|}$$

而 $y_i = -1$ 时距离是上式的相反数。再注意到我们可以同时对 w 和 b 乘以一个正常数并不会改变超平面的位置和方向，于是我们可以直接限制 $\|w\| = 1$，于是优化问题可以完整地写为

$$\begin{aligned} \max_{w,b,\delta} \ & \delta \\ s.t. \ & w^T x_i - b \geq \delta, \quad y_i = +1 \\ & w^T x_i - b \leq -\delta, \quad y_i = -1 \\ & \|w\| = 1 \end{aligned}$$

这已经是一个标准的带约束优化问题了[1]。不过我们注意到 $\|w\| = 1$ 这个约束并不是线性的，所以并不能一下子看出来这个问题的凸性。并且这个也和我们常见的 SVM 的目标函数差别有点大，于是我们做一下变量代换，令 $\omega = w/\delta, \beta = b/\delta$，则上面的问题变成

$$\begin{aligned} \max_{\omega,\beta,\delta} \ & \delta \\ s.t. \ & \omega^T x_i - \beta \geq 1, \quad y_i = +1 \\ & \omega^T x_i - \beta \leq -1, \quad y_i = -1 \\ & \|\omega\| = \frac{1}{\delta} \end{aligned}$$

再做简单变化消掉变量 δ，即可得到

$$\begin{aligned} \min_{\omega,\beta} \ & \|\omega\| \\ s.t. \ & \omega^T x_i - \beta \geq 1, \quad y_i = +1 \\ & \omega^T x_i - \beta \leq -1, \quad y_i = -1 \end{aligned}$$

到这里我们看到了熟悉的 SVM 的目标函数了（比较简单的没有带 slack variable 的版本），并且此时所有约束都是线性的，我们可以很明显地看出这是一个凸优化问题。通常情况下为了求导方便目标函数会写成 $\|\omega\|^2$ 而不是 $\|\omega\|$，显然这两者的最优解是等价的。

现在让我们回到抽象的带约束优化问题上。回忆一下，在 unconstrained 优化问题中，我们证明了目标函数的 gradient 等于零是（局部）最优解的必要条件，并由此得出了 gradient descent 算法，通过不动点迭代来寻找 gradient 的零点。但是在 constrained 优化问题中，有可能出现的情况是目标函数 gradient 等于零的点是不能达到的，也就是说在约束集之外的，此时一般最优解会在约束集的边界上取到。总而言之 gradient 的零点不再是最优解的必要条件了，接下来我们将简要介绍一下在带约束的情况下如何来刻画最优解。

为了记号上方便，我们采用向量记法 $g(x) = (g_1(x), \ldots, g_m(x))$、$h(x) = (h_1(x), \ldots, h_l(x))$。此时带约束优化问题写作

$$\begin{aligned} \min_{x} \ & f(x) \\ s.t. \ & g(x) \preceq 0 \\ & h(x) = 0 \end{aligned}$$

注意这里的讨论中我们并没有要求这是一个 convex 优化问题，但是为了简单起见，我们假设 $h(x)$ 是线性约束，但是即使 $h(x)$ 是非线性的，结论也会在给 $h(x)$ 的 gradient 满足一定条件的时候成立，只是论证的过

[1] 只要把 max 通过一个负号转换成 min 即可。

程需要使用隐函数定理之类的工具来进行局部讨论，有点繁琐。于是我们索性将问题写作

$$\begin{aligned} \min_x \ & f(x) \\ s.t. \ & g(x) \preceq 0 \\ & Ax - b = 0 \end{aligned}$$

接下来我们考虑一个比较直观的几何条件。假设我们在一个 feasible point $\bar{x}$，如果想要移动到一个更优的点，那么首先我们肯定希望向着 descent 方向移动；另一方面，我们希望移动的方向不会使得我们跑到 feasible region 外面去。

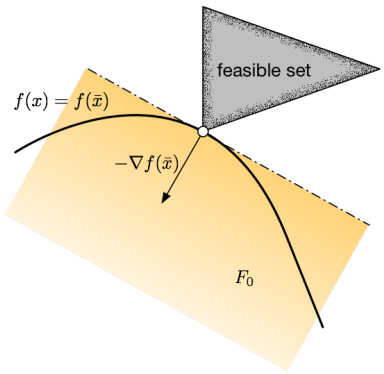

Figure 1: **几何条件的简单示例。**局部最优的必要条件是目标函数的下降方向和 active constraints 的下降方向没有交集。

定理 1 (Geometric First-order Necessary Conditions) *令 $F_0 = \{d \neq 0 : \nabla f(\bar{x})^T d < 0\}$ 为 descent 方向的集合，I 表示当前所有 active 的不等式 constraints 的下标集合：$I = \{i : g_i(\bar{x}) = 0\}$，而 $G_0 = \{d \neq 0 : \nabla g_i(\bar{x})^T d < 0, \forall i \in I\}$ 是所有当前 active 的不等式 constraints 的 descent 方向的交集，$H_0 = \{d \neq 0 : Ad = 0\}$ 是保持 equality constraints 继续满足的方向集合。*

则 $\bar{x}$ 是局部最优解的必要条件是 $F_0 \cap G_0 \cap H_0 = \emptyset$。

该定理其实就是把我们刚才的 intuition 用数学语言陈述了一下。证明也非常简单，如果交集非空的话，我们任取交集中的一个方向 d，只要往该方向移动足够小的距离，可以保证同时减小目标函数的值并维持 feasible 状态，于是 $\bar{x}$ 就不可能是局部最优解了。不过这个几何角度的定理虽然直观，但是却不怎么有用，接下来我们把这里的必要条件转化为可以具体运算的代数描述。

为此我们需要一些凸分析方面的工具，所以这里暂时 “离题” 介绍一些结论，因为可能以后也会用到这些结论，所以这里明确地给出来。

性质 1 *设 $S \subset \mathbb{R}^n$ 是一个非空闭凸集，$y \notin S$，则*

$$\min_{x \in S} \|x - y\|$$

存在唯一解 x^*。且 x^* 是最优解的充要条件是

$$(y - x^*)^T (x - x^*) \leq 0, \quad \forall x \in S$$

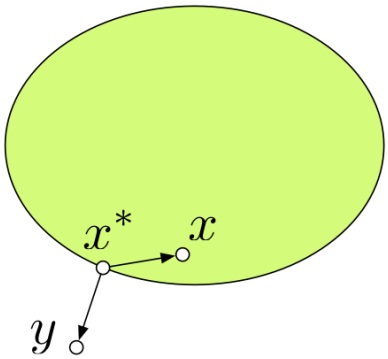

Figure 2: **凸集投影示例。**向量 $y - x^*$ 与 $x - x^*$ 成钝角。其中 x^* 是 y 到凸集中的投影，而 x 是凸集中任意其他的点。

从图 2 中来看这个结论所说的事情还是比较容易直观地记住的，证明并不复杂，我们就不在这里详细给出了。通过这个结论我们可以得出

一系列的关于凸集的 separating hyperplane 的定理，根据凸集的开闭性、紧致性以及是凸集与点之间还是两个凸集之间会有各种不同的结论和变种。下面介绍一个我们马上会用到的版本。

定理 2 (Strong Separation) 设 $S \subset \mathbb{R}^n$ 是一个非空闭凸集，且 $y \notin S$，则存在超平面 $w^T x - b = 0$ 以及 $\epsilon > 0$，使得 $w^T y - b \geq \epsilon$ 且对任意 $x \in S$ 有 $w^T x - b \leq -\epsilon$。此时我们称超平面 *strongly separate* 凸集 S 和点 y。

利用前面的命题中得到的结论，令 x^* 是 y 到 S 的投影，则容易验证实际上 $(y - x^*)^T x - (y - x^*)^T x^* - \epsilon = 0$ 当 $\epsilon \leq 0.5\|y - x^*\|^2$ 时，就是满足条件的 separating hyper plane。

定理 3 (Farkas' Lemma) 设 $A \in \mathbb{R}^{m \times n}$，$c \in \mathbb{R}^n$，则下面的两个线性系统中一定有且只有一个存在解：

1. $Ax \preceq 0, c^T x > 0$ 2. $A^T y = c, y \geq 0$

这个定理初看有些莫名其妙，简单解释一下。考虑 A 的行所对应的向量 $A_1, \ldots, A_m$，第一条有解说的是存在一个向量 x，与所有的 $A_1, \ldots, A_m$ 成钝角（非锐角），而和 c 成锐角。而该向量以法向量实际上定义了一个超平面将 A 的行向量与 c 分开。

第二条有解说的是存在一个非负权重，使得 c 是 $A_1, \ldots, A_m$ 的加权平均。换句话说，$c \in \text{cone}(A_1, \ldots, A_m)$。

这样从几何的角度来解释之后 Farkas' Lemma 其实还是挺直观的，实际上我们正是要用 Farkas' Lemma 将代数和几何联系起来。我们这里省略证明，主要的麻烦的地方在于要证明一个构造出来的凸集是闭的，之后直接运用上面的 strong separation 结论就可以证明了。具体可以参见 [Bertsekas, 1999] 的附录 B.3 的内容。

定理 4 (Fritz John Necessary Conditions) 设 $\bar{x}$ 是局部最优解，则存在 u_0, u 和 v，使得

$$\begin{aligned} u_0 \nabla f(\bar{x}) + \nabla g(\bar{x})^T u + A^T v &= 0 \\ u_0, u &\geq 0, (u_0, u, v) \neq 0 \\ u_i g_i(\bar{x}) &= 0, i = 1, \ldots, m \end{aligned}$$

根据 1，$F_0 \cap G_0 \cap H_0 = \emptyset$，也就是说，不存在 d 使得

$$\begin{aligned}
&\nabla f(\bar{x})^T d < 0 \\
&\nabla g_i(\bar{x})^T d < 0, \quad i \in I \\
&Ad = 0
\end{aligned}$$

令 $\nabla g_I(\bar{x}) = (\nabla g_i(\bar{x}))_{i \in I}$，以及

$$B = \begin{bmatrix} \nabla f(\bar{x})^T \\ \nabla g_I(\bar{x})^T \end{bmatrix}$$

则如下 system 无解

$$\begin{aligned}
Bx + e\theta \preceq 0, \quad \theta > 0 \\
Ax \preceq 0 \\
-Ax \preceq 0
\end{aligned}$$

变换一下形式

$$\begin{bmatrix} B & e \\ A & 0 \\ -A & 0 \end{bmatrix} \begin{bmatrix} x \\ \theta \end{bmatrix} \preceq 0, \quad \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T \begin{bmatrix} x \\ \theta \end{bmatrix} > 0$$

于是根据刚才的 Farkas' Lemma，我们知道存在 $y = (u_0, u_I, v^+, v^-) \succeq 0$ 使得

$$\begin{aligned}
u_0 \nabla f(\bar{x}) + \nabla g_I(\bar{x})^T u_I + A(v^+ - v^-) = 0 \\
u_0 + e^T u = 1
\end{aligned}$$

令 $v = v^+ - v^-$，且定义 u 使得 active set I 以外的元素为零，即证。

不过 Frtiz John 条件在 $\nabla f(\bar{x})$ 前面有一个系数 u_0 比较讨厌，如果 $u_0 > 0$ 的话，我们可以在等式两边同时除以 u_0，得到

$$-\nabla f(\bar{x}) = \nabla g(\bar{x})^T u / u_0 + A^T v / u_0$$

此时可以理解成目标函数的 gradient 向量的相反方向由 constraints 的 gradient 线性张成，并且考虑不等式 constraints 的话，由于 $u \geq 0$，所以还是在张成的 cone 中的。并且这种形式也是我们平常经常看到的 Lagrangian 的 gradient 等于零的形式。

为了做到这一点，注意到 Fritz John 条件里已经有了 $u_0 \geq 0$，所以再只要排除 $u_0 = 0$ 这种情况就可以了。当 $u_0 = 0$ 时，我们可以得到：

$$\nabla g(\bar{x})^T u + A^T v = 0, \quad (u, v) \neq 0$$

也就是说等式和不等式的 constraints 的 gradient 是线性相关的。因此如果我们能保证它们线性无关的话，就可以排除这种情况。于是有如下结论。

定理 5 (Karush-Kuhn-Tucker (KKT) Necessary Conditions) *设 $\bar{x}$ 是局部最优解，如果 $\nabla g_i(\bar{x}), i \in I$ 和 A 的行所对应的向量是线性无关的，则存在 u, v 使得*

$$\begin{aligned} \nabla f(\bar{x}) + \nabla g(\bar{x})^T u + A^T v &= 0 \\ u &\succeq 0 \\ u_i g_i(\bar{x}) &= 0, \quad i = 1, \ldots, m \end{aligned}$$

这就是我们常见的 KKT 条件，成立的条件是 $\bar{x}$ 是局部最优解并且 "线性无关"，这里的 "线性无关" 也可以换成其他条件，通常称为 constraint qualification。除了刚才的定理之外，这里再列举两个比较常用的 constraint qualification。

定理 6 (Slater Condition) *如果 $g_i(x)$ 是 convex，且 A 的行线性无关，并且优化问题存在 Slater point，亦即存在一个 feasible 点 x 使得所有的不等式 constraints 严格满足：$g_i(x) < 0$。那么 KKT 条件是局部最优解的必要条件。*

定理 7 (Linear Constraints) *如果所有 constraints 都是线性的，那么 KKT 条件是局部最优解的必要条件。*

注意以上两个结论中都并没有要求目标函数 $f(x)$ 是 convex 或者是 linear 的。但是如果 $f(x)$ 和 $g_i(x)$ 全都是 convex 的话，KKT 同时也是充分条件。

定理 8 (KKT Sufficient Conditions for Convex Problems) *设 $\bar{x}$ 是一个 feasible point，如果存在 u, v 使得 KKT 条件满足，亦即*

$$\begin{aligned} \nabla f(\bar{x}) + \nabla g(\bar{x})^T u + A^T v &= 0 \\ u &\succeq 0 \\ u_i g_i(\bar{x}) &= 0, \quad i = 1, \ldots, m \end{aligned}$$

如果 $f(x)$ 和 $g_i(x), i = 1, \ldots, m$ 是 convex 的，那么 $\bar{x}$ 是该问题的全局最优解。

到这里为止几乎未加证明地介绍了许多结论了，让我们简单地带入 SVM 的问题中小结一下结束本篇文章。这里我们再对 SVM 问题做一些简单的变换，一个是用 $1/2\|\omega\|^2$ 代替 $\|\omega\|$，再就是将所有不等式 constraints 写成统一形式：

$$\begin{aligned}&\min_{\omega,\beta}\ \frac{1}{2}\|\omega\|^2\\&s.t.\ y_i(\omega^T x_i-\beta)\geq 1,\quad i=1,\dots,N\end{aligned}$$

首先该问题显然是 convex 的，因此 KKT 条件是充分条件。然后 Slater point 的存在性其实就是看数据是否严格线性可分[2]，也就是说是否存在一个超平面 $\omega^T x-\beta=0$ 将两类数据严格分开：$y_i(\omega^T x_i-\beta)>1$。如果 Slater point 存在的话，那么 KKT 条件会是刻画该问题的全局最优解的充分必要条件：

$$\begin{aligned}\omega-\sum_{i=1}^{N}u_i y_i x_i&=0\\u_i&\geq 0,\quad i=1,\dots,N\\u_i\left(1-y_i(\omega^T x_i-\beta)\right)&=0,\quad i=1,\dots,N\end{aligned}$$

如果将 u_i 和 SVM 的 dual 推导里的 dual 变量联系起来的话，其中第一条实际上是给出了关于结果的 separating hyperplane 的法向量 ω 的一个显式表达式，可以看到对于任意需要分类的点 $\tilde{x}$，有

$$\omega^T\tilde{x}-\beta=\sum_{i=1}^{N}u_i y_i\langle x_i,\tilde{x}\rangle-\beta$$

只涉及到 x_i 和 $\tilde{x}$ 之间的内积运算，因此可以使用 kernel tricks。然后 $u_i\geq 0$ 是每一个 training data point x_i 的权重，注意第三个式子（称为 complememtary slackness），如果 x_i 不在 separating hyperplane 的 margin 上，也就是 $y_i(\omega^T x_i-\beta)>1$ 的话，为了让 complememtary slackness 成立，对应的 u_i 必须为零。也就是说不在 margin 上的数据点，在最后进行分类的时候其对应的权重 $u_i=0$，也就是说只有 margin 上的数据点（称为 supporting vectors）才会起到作用，这也是 SVM 名字的由来。

最后小结一下，优化问题的最优解的充分条件和必要条件各自有什么用处。我们目前还没有涉及到任何实际的求解算法，但是必要条件一般可以用来寻找可能的最优解，因为最优解如果存在的话，一定会满足必要条件，因此一个比较暴力的办法就是直接枚举所有符合必要条件的解寻找最优的那一个。有些问题可以比较直接地解出满足必要条件的点，而令一些问题可能需要更多地步骤，例如在之前介绍的 unconstrained 优化问题中就是通过不动点迭代去寻找对应的必要条件（目标函数的 gradient 等于零）的点。而反过来，充分条件则有时候不是想象中的那么

[2] 注意到 SVM 的 constraints 全都是线性的，所以我们也可以利用 linear constraints 那个结论，不需要 Slater point 点存在性也可以得到 KKT 条件的必要性。

有用：如果你刚巧拿到了一个解并且这个解刚巧满足了充分条件的话，那么就皆大欢喜了：恭喜你找到了最优解。但是如果不满足充分条件的话，就说不清楚了，它既有可能是最优解也有可能不是。

References

[Bertsekas, 1999] Bertsekas, D. P. (1999). *Nonlinear programming*. Athena Scientific.

[Stewart, 1998] Stewart, G. W. (1998). *Matrix Algorithms: Volume 1, Basic Decompositions*. Matrix Algorithms. Society for Industrial and Applied Mathematics.