"""

Created on 2017/8/03

machine-learning-course

@author: DSG
"""

LagrangeDuality

原始问题

假设 $f(x)$, $c_i(x)$, $h_j(x)$ 是定义在 R^n 上的连续可微函数，考虑约束条件下最优化问题：

$$\min_{x \in R^n} f(x)$$

$$s.t.\ \ c_i(x) \le 0,\ \ i = 1,2,...,k$$
$$h_j(x) = 0,\ \ j = 1,2,...,l$$

称为约束最优化问题的**原始问题**。
现在**如果不考虑约束条件**，原始问题就是：

$$\min_{x \in R^n} f(x)$$

没有约束条件下的问题很简单，高中生就可以搞定；在约束条件下引入广义拉格朗日函数（generalized Lagrange function）：

$$L(x,\alpha,\beta) = f(x) + \sum_{i=1}^{k} \alpha_i c_i(x) + \sum_{j=1}^{l} \beta_j h_j(x)$$

其中 $x = (x^{(1)}, x^{(2)}, ..., x^{(n)})^T \in R^n$，$\alpha_i, \beta_j$ 是拉格朗日乘子（某种意义上的参数），特别要求 $\alpha_i \ge 0$

现在，如果把 $L(x,\alpha,\beta)$ 看作是关于 α_i, β_j 的函数，要求其最大值，即：

$$\max_{\alpha,\beta:\alpha_j \ge 0} L(x,\alpha,\beta)$$

现在 $L(x,\alpha,\beta)$ 看作是关于向量 α, β 的函数，经过各种优化和处理，只要得到向量 α, β 的值，使得 $L(x,\alpha,\beta)$ 取最大值（当中把 x 当做是常量），确定了 α, β 的值，就可以得到 $L(x,\alpha,\beta)$ 的最大值，因为 α, β 已经确定，所以 $L(x,\alpha,\beta)$ 的最大值就是只和 x 有关的函数，定义这个函数为:

$$\theta_P(x) = \max_{\alpha,\beta:\alpha_j \ge 0} L(x,\alpha,\beta)$$

其中 $L(x,\alpha,\beta)=f(x)+\sum_{i=1}^{k}\alpha_i c_i(x)+\sum_{j=1}^{l}\beta_j h_j(x)$，特别强调的是这个最大值是针对$\alpha,\beta$而言的。

下面通过x是否满足约束条件来分析函数$\theta_P(x)$.

1、考虑某个x违反了原始的约束，即$c_i(x)\geq 0$或$h_j(x)\neq 0$，那么对于函数$\theta_P(x)$他的结果就是$+\infty$

$$\theta_P(x)=\max_{\alpha,\beta:\alpha_j\geq 0}[f(x)+\sum_{i=1}^{k}\alpha_i c_i(x)+\sum_{j=1}^{l}\beta_j h_j(x)]$$

因为是求的最大值，当$c_i(x)\geq 0$时为了取得最大值我们可以令$\alpha_i\to+\infty$，即可使得$\sum_{i=1}^{k}\alpha_i c_i(x)\to+\infty$；如果

$h_j(x)\neq 0$，则很容易取得β_j使得$\sum_{j=1}^{l}\beta_j h_j(x)\to+\infty$。

2、当x满足原始条件的时候，因为$c_i(x)\leq 0$，所以我们可以令$\alpha_i=0$即可保证$f(x)+\sum_{i=1}^{k}\alpha_i c_i(x)+\sum_{j=1}^{l}\beta_j h_j(x)$

有最大值，此时$h_j(x)=0$不用考虑，而此时的最大值就是$f(x)$。

综上两种情况我么你可以得到：

$$\theta_P(x)=\begin{cases}f(x), & x\text{ 满足原始问题约束}\\ +\infty, & \text{其他}\end{cases}$$

那么在满足条件的情况下$\min_x\theta_P(x)=\min_x\max_{\alpha,\beta:\alpha_j\geq 0}L(x,\alpha,\beta)=\min_x f(x)$，即$\min_x\theta_P(x)$与原始优化问题等价，

所以常用$\min_x\theta_P(x)$代表原始问题，下表P表示原始问题，定义运势问题的最优值为p^*，则

$$p^*=\min_x\theta_P(x)$$

通过拉格朗日函数将约束条件下最优化定义为一个无约束问题，这个无约束问题等价于原来的约束优化问题，从而将约束问题无约束化！

对偶问题

定义关于α,β的函数：

$$\theta_D(\alpha,\beta)=\min_x L(x,\alpha,\beta)$$

注意等式右边是关于x的函数的最小化，x确定以后，最小值就只与α,β有关，所以是一个关于α,β的函数。

考虑函数 $\theta_D(\alpha,\beta)$ 极大化，即 $\max\limits_{\alpha,\beta:\alpha_j\geq 0}\theta_D(\alpha,\beta)=\max\limits_{\alpha,\beta:\alpha_j\geq 0}\min\limits_{x}L(x,\alpha,\beta)$,这就是原始问题的对偶问题，

原始问题为：

$$\min_{x}\theta_P(x)=\min_{x}\max_{\alpha,\beta:\alpha_j\geq 0}L(x,\alpha,\beta)$$

对偶问题为：

$$\max_{\alpha,\beta:\alpha_j\geq 0}\theta_D(\alpha,\beta)=\max_{\alpha,\beta:\alpha_j\geq 0}\min_{x}L(x,\alpha,\beta)$$

形式上可以看出很对称，只不过原始问题是先固定 $L(x,\alpha,\beta)$ 中的 x，求出优化参数 α,β，然后在求解问题。而对偶问题刚好相反，先固定 α,β 求出最优化的 x，然后在确定 α,β。定义**对偶问题的最优值：**

$$d^*=\max_{\alpha,\beta:\alpha_j\geq 0}\theta_D(\alpha,\beta)$$

在 SVM 算法中，我们已知一组样本，从样本中找到支撑向量，也就是在对偶问题中找到最优化 x，然后再根据 x 确定超平面的参数 α,β。

原始问题与对偶问题的关系

定理：若原始问题与对偶问题都有最优值，则：

$$d^*=\max_{\alpha,\beta:\alpha_j\geq 0}\min_{x}L(x,\alpha,\beta)\leq\min_{x}\max_{\alpha,\beta:\alpha_j\geq 0}L(x,\alpha,\beta)=p^*$$

证明，对任意的 α,β 和 x ：

$$\theta_D(\alpha,\beta)=\min_{x}L(x,\alpha,\beta)\leq L(x,\alpha,\beta)\leq\max_{\alpha,\beta:\alpha_j\geq 0}L(x,\alpha,\beta)=\theta_P(x)$$

即：

$$\theta_D(\alpha,\beta)\leq\theta_P(x)$$

由于原始问题与对偶问题都有最优值，所以：

$$d^*=\max_{\alpha,\beta:\alpha_j\geq 0}\min_{x}L(x,\alpha,\beta)\leq\min_{x}\max_{\alpha,\beta:\alpha_j\geq 0}L(x,\alpha,\beta)=p^*$$

也就是说原始问题的最优值不小于对偶问题的最优值，但是我们要通过对偶问题来求解原始问题，就必须使得原始问题的最优值与对偶问题的最优值相等，于是可以得出下面的推论：

推论：设 x^*,α^*,β^* 分别是原始问题和对偶问题的可行解，如果 $d^*=p^*$，那么 x^*,α^*,β^* 分别是原始问题和对偶问题的最优解。

所以，当原始问题和对偶问题的最优值相等：$d^*=p^*$ 时，可以用**求解对偶问题来求解原始问题（当然是对偶问题求解比直接求解原始问题简单的情况下）**，但是到底满足什么样的条件才能使的 $d^*=p^*$ 呢，这就是下面要阐述的 KKT 条件

调换对偶问题中对拉格朗日函数取最大化、最小化的顺序即可得到与原问题等价的优化问题。即，对偶问题是对拉格朗日函数先取最小化，再取最大化；而原问题则是对拉格朗日函数先取最大化，再取最小化。

为了对比两优化问题之间的对偶性，我先列出对偶问题的形式：

Karush–Kuhn–Tucker conditions ， KKT 条件

定理：对于原始问题和对偶问题，假设函数 $f(x)$ 和 $c_i(x)$ 是凸函数， $h_j(x)$ 是仿射函数（即由一阶多项式构成的函数， $f(x)=Ax+b$， A 是矩阵， x , b 是向量）；并且假设不等式约束 $c_i(x)$ 是严格可行的，即存在 x ，对所有 i 有 $c_i(x)\le 0$ ，则存在 x^*,α^*,β^* ，使得 x^* 是原始问题的最优解， α^*,β^* 是对偶问题的最优解，并且 $d^*=p^*=L(x,\alpha,\beta)$

定理：对于原始问题和对偶问题，假设函数 $f(x)$ 和 $c_i(x)$ 是凸函数， $h_j(x)$ 是仿射函数（即由一阶多项式构成的函数， $f(x)=Ax+b$， A 是矩阵， x , b 是向量）；并且假设不等式约束 $c_i(x)$ 是严格可行的，即存在 x ，对所有 i 有 $c_i(x)\le 0$ ，则存在 x^*,α^*,β^* ，使得 x^* 是原始问题的最优解， α^*,β^* 是对偶问题的最优解的充要条件是 x^*,α^*,β^* 满足下面的 Karush-Kuhn-Tucker（KKT）条件：

1. $c_i(x)\le 0,\ i=1,2,...,k$
2. $h_j(x)=0,\ j=1,2,...,l$
3. $\nabla f(x^*)+\sum_{i=1}^{k}\alpha_i\nabla c_i(x^*)+\sum_{j=1}^{l}\beta_j\nabla h_j(x^*)=0$ ，其中 ∇ 为梯度算子
4. $\beta_j\ne 0$ 且不等式约束条件满足 $\alpha_j\ge 0$ ， $\alpha_i c_i(x)=0,\ i=1,2,...,k$ （对偶互补条件）

关于 KKT 条件的理解：前面三个条件是由解析函数的知识，对于各个变量的偏导数为 0（这就解释了一开始为什么假设三个函数连续可微，如果不连续可微的话，这里的偏导数存不存在就不能保证），后面第四个条件就是原始问题的约束条件以及拉格朗日乘子需要满足的约束。

KKT 条件第一二项是说最优解 x^* 必须满足所有等式及不等式限制条件， 也就是说最优解必须是一个可行解， 这一点自然是毋庸置疑的． 第三项表明在最优点 x^* ， ∇f 必须是 ∇c_i 和 ∇h_j 的线性組合， α_i,β_i 都叫作拉格朗日乘子． 所不同的是不等式限制条件有方向性， 所以每一个 α_i 都必须大于或等于零， 而等式限制条件没有方向性，所以 β_i 没有符号的限制，其符号要视等式限制条件的写法而定．

KKT 条件的意义：它是一个非线性规划（Nonlinear Programming）问题能有最优化解法的必要和充分条件。

特别注意当$\alpha_j \geq 0$时，由 KKT 对偶互补条件可知：$\alpha_i c_i(x) = 0$，这个知识点会在 SVM 的推导中用到.