Support Vector Machine

1. Notation

在讨论 SVM 的时候，出于简化的目的，我们需要引进新的符号来表示分类。假设我们要对一个二分类问题建立一个线性分类器，其中，标签(label)为 y，特征(feature)为 x。从现在开始，我们使用 $y \in \{-1, 1\}$ 来表示类别(而不是之前的 $y \in \{0, 1\}$)。同时，我们也不再使用向量 θ 来表示线性分类器的参数，而是使用参数 w 和 b，所以，我们的线性分类器为

$$h_{w,b}(x) = g(w^T x + b)$$

其中，当 $z \geq 0$ 时，$g(z) = 1$；当 $z < 0$ 时，$g(z) = -1$。这里的参数 "w, b" 可以让我们把截距项(intercept term) b 和其他参数分开。（此外，我们不需要像之前那样设定 $x_0 = 1$。）因此，这里的参数 b 就相当于之前的 θ_0，而参数 w 则相当于 $[\theta_1, \cdots, \theta_n]^T$。

需要注意的是，根据我们对函数 g 的定义，我们的分类器会直接预测类别是1或者-1（参考感知机算法 perceptron algorithm ），这样也就不需要经过中间步骤来估计 y 为1的概率。（这里的中间步骤指的是逻辑回归的步骤）

2. Functional and geometric margins

给定一个训练样本 $(x^{(i)}, y^{(i)})$，我们定义该训练样本的函数间隔(functional margin)为

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$$

注意到，为了使函数间隔尽可能大（也就是说，为了让我们的预测是正确的且有着高的确信度），当 $y^{(i)} = 1$ 时，我们需要尽可能地使 $w^T x^{(i)} + b$ 是一个大的正数。相反，当 $y^{(i)} = -1$ 时，我们需要让 $w^T x^{(i)} + b$ 尽可能是一个大的负数。而且，如果 $y^{(i)}(w^T x + b) > 0$，则说明我们对这个训练样本的预测是正确的。因此，一个大的函数间隔代表着一个正确且有着高确信度的预测。

给定一个训练集 $S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, m\}$，我们定义该训练集的函数间隔为所有训练样本的函数间隔的最小值。记为 $\hat{\gamma}$，即

$$\hat{\gamma} = \min_{i=1,\cdots,m} \hat{\gamma}^{(i)}$$

对于一个线性分类器，选择上面给定的函数 g (取值范围为 $\{-1, 1\}$)。我们注意到函数间隔有一个性质，使得它不能够很好地衡量确信度。如果我们将 w 和 b 换成 $2w$ 和 $2b$，那么由于 $g(w^T x + b) = g(2w^T x + 2b)$，这将不会改变 $h_{w,b}(x)$。也就是说，函数 g 和 $h_{w,b}(x)$ 只取决于 $w^T x + b$ 的正负符号(sign)，但不受其数值大小(magnitude)的影响。但是，把 (w, b) 替换成 $(2w, 2b)$ 会导致函数间隔被放大了两倍。换句话说，成比例地缩放 w 和 b 后，超平面(hyperplane)并没有变化，而函数间隔却变化了。直观地看，这导致我们需要引入某种归一化条件，例如，$\|w\| = 1$。

因此，给定一个训练样本 $(x^{(i)}, y^{(i)})$ 我们定义该训练样本的几何间隔(geometric margins)为

$$\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{\|w\|} \right)^T x^{(i)} + \frac{b}{\|w\|} \right)$$

这样，当我们成比例地缩放 w 和 b 的时候，几何间隔的大小是不变的。值得注意的是，正是这种对参数缩放的不变性，当我们试图对某个训练集拟合 w 和 b 的时候，我们可以对 w 添加任意的缩放约束。例如，我们可以要求 $\|w\| = 1$ ，$|w_1| = 5$ ，$|w_1 + b| + |w_2| = 2$ 等等，这些只要对 w 和 b 进行成比例的缩放就可以满足。而且由于是成比例地缩放，所以并不改变几何间隔的大小。

然后，给定一个训练集 $S = \{(x^{(i)}, y^{(i)}), i = 1, \cdots, m\}$，我们也可以定义该训练集的几何间隔为所有训练样本的几何间隔的最小值。记为 γ，即

$$\gamma = \min_{i=1,\cdots,m} \gamma^{(i)}$$

最后，根据函数间隔和几何间隔的定义可知，函数间隔和几何间隔有如下关系：

$$\gamma^{(i)} = \frac{\hat{\gamma}^{(i)}}{\|w\|}$$
$$\gamma = \frac{\hat{\gamma}}{\|w\|}$$

小结：

我们的目的是要建立一个最大间隔分类器。具体地说，我们希望最大化整个训练集的间隔，而整个训练集的间隔被定义为所有训练样本间隔的最小值。而之所以不采用函数间隔来衡量训练样本到超平面的距离，是因为对于任意一个成功划分训练样本的超平面来说，我们只要成比例地放大 w 和 b 就可以让函数间隔任意大。这样我们就无法找到最优的超平面。因此，我们使用几何间隔来衡量训练样本到超平面的距离。而几何间隔其实就是高中学的点到直线距离的推广。而由于几何间隔对于参数缩放的不变性，使得我们下面可以使用这个性质简化我们的问题。

3. The optimal margin classifier

到目前为止，我们假设给定的训练集是线性可分的(linearly separable)。也就是说，能够在正负训练样本之间用某种分离超平面(separating hyperplane)进行划分。我们的希望是能够找到使得几何间隔最大的分离超平面。因此，我们可以提出如下的优化问题：

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

即我们希望最大化超平面关于训练集的几何间隔 γ，约束条件表示的是超平面与每个训练样本之间的几何间隔至少是 γ。因为几何间隔和函数间隔可以通过 $\gamma = \frac{\hat{\gamma}}{\|w\|}$ 联系起来，所以可以将问题改写为：

$$\begin{aligned} \max_{w,b} \quad & \frac{\hat{\gamma}}{\|w\|} \\ s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \cdots, m \end{aligned}$$

但注意到，此时我们的目标函数 $\frac{\hat{\gamma}}{\|w\|}$ 是非凸的(non-convex)。

上一节，我们提到参数的成比例缩放不改变几何间隔的大小，这使得我们可以对 w 和 b 的缩放增加约束。基于这个性质，我们可以令函数间隔 $\hat{\gamma} = 1$。也就是说，我们对 w 和 b 的成比例缩放要使得函数间隔为1。因此

$$\begin{aligned} \max_{w,b} \quad & \frac{1}{\|w\|} \\ s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \cdots, m \end{aligned}$$

由于最大化 $\frac{1}{\|w\|}$ 和最小化 $\frac{1}{2}\|w\|^2$ 等价，因此，可得

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\ s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \cdots, m \end{aligned}$$

通过这样的转化，这个问题就比较容易解决了。上面的问题是一个凸二次规划(convex quadratic programming)问题。此时，我们可以调用成熟的商业QP软件包对此进行求解。

接下来，我们先岔开话题，介绍一下拉格朗日对偶(Lagrange duality)，这样会引出我们这个优化问题的对偶形式。通过这种对偶形式，我们可以推出一种非常有效的算法，来解决上面这个优化问题，而且通常这个算法比通用的QP软件更好用。

4. Lagrange duality

对于任意一个带约束的优化问题，我们可以将其写成这样的形式

$$\begin{aligned} \min_{x} \quad & f_0(x) \\ s.t. \quad & f_i(x) \leq 0, \quad i = 1, \cdots, m \\ & h_i(x) = 0, \quad i = 1, \cdots, p \end{aligned}$$

假如 $f_0, f_1, \cdots, f_m$ 全都是凸函数，并且 $h_1, \cdots, h_p$ 全都是仿射函数，那么这个问题就叫做凸优化(convex optimization)问题。凸优化问题有许多优良的性质。不过，这里我们并没有假定我们要处理的问题是凸优化问题。

我们希望把带约束的优化问题转化为无约束的优化问题。为此，我们定义 Lagrangian 如下：

$$L(x, \lambda, v) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)$$

这里，λ_i 和 v_i 称为拉格朗日乘数(Lagrange multipliers)。接下来，我们让 Lagrangian 针对 λ 和 v 最大化，令

$$\theta_P(x) = \max_{\lambda,v\,:\,\lambda_i\geq 0} L(x,\lambda,v)$$

上式中的“P”是对“primal”的简写。可以看到，如果 x 不满足原始问题的约束条件（例如，对于某些 i，存在 $f_i(x) > 0$ 或者 $h_i(x) \neq 0$），那么

$$\begin{aligned}\theta_P(x) &= f_0(x) + \max_{\lambda,v\,:\,\lambda_i\geq 0}\left(\sum_{i=1}^{m}\lambda_i f_i(x) + \sum_{i=1}^{p} v_i h_i(x)\right)\\ &= \infty\end{aligned}$$

而对于那些满足约束条件的 x 来说，$\theta_P(x) = f_0(x)$。因为满足约束条件的 x 会使得 $h_i(x) = 0$，所以最后一项消掉了。而 $f_i(x) \leq 0$，并且我们要求 $\lambda_i \geq 0$，因此，$\lambda_i f_i(x) \leq 0$，所以最大值就只能在它们都为零的时候得到，这个时候就只剩下 $f_0(x)$ 了。因此，对于那些满足约束条件的 x 来说，$\theta_P(x) = f_0(x)$。这样一来，原始的带约束优化问题就等价于如下的无约束优化问题：

$$\min_x\ \theta_P(x)$$

到这里，我们成功地把带约束优化问题转化为无约束优化问题，不过这其实只是一个形式上的重写，并没有什么本质上的改变。我们只是把原来的问题通过 Lagrangian 写作了如下形式：

$$\min_x \theta_P(x) = \min_x \max_{\lambda,v\,:\,\lambda_i\geq 0} L(x,\lambda,v)$$

这个问题（或者说原始的带约束的形式）称做 primal problem。接下来，我们定义

$$\theta_D(\lambda,v) = \min_x\ L(x,\lambda,v)$$

上式中的“D”是“dual”的缩写。这里要注意的是，之前对 θ_P 的定义中，是对 λ 和 v 进行优化（找最大值），这里则是找 x 的最小值。

现在，我们就可以定义 dual problem，即

$$\max_{\lambda,v\,:\,\lambda_i\geq 0} \theta_D(\lambda,v) = \max_{\lambda,v\,:\,\lambda_i\geq 0}\ \min_x\ L(x,\lambda,v)$$

$\theta_D(\lambda,v)$ 有一个很好的性质就是它是 primal problem 的一个下界。换句话说，如果 primal problem 的最小值记为 p^*，那么对于所有的 $\lambda_i \geq 0$ 和 v，我们有：

$$\theta_D(\lambda,v) \leq p^*$$

因为对于极值点（实际上包括所有满足约束条件的点）x^*，注意到 $\lambda_i \geq 0$，我们总是有

$$\sum_{i=1}^{m}\lambda_i f_i(x^*) + \sum_{i=1}^{p} v_i h_i(x^*) \leq 0$$

因此

$$L(x^*, \lambda, v) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i f_i(x^*) + \sum_{i=1}^{p} v_i h_i(x^*) \leq f_0(x^*)$$

于是

$$\theta_D(\lambda, v) = \min_x L(x, \lambda, v) \leq L(x^*, \lambda, v) \leq f_0(x^*) = p^*$$

这样一来，就确定了 $\theta_D(\lambda, v)$ 的下界性质，于是

$$\max_{\lambda, v\,:\,\lambda_i \geq 0} \theta_D(\lambda, v)$$

实际上就是最大的下界。这是很自然的，因为在得到下界之后，我们自然就希望得到最大的下界，因为它离我们要逼近的值最近。记 dual problem 的最优值为 d^* 的话，根据上面的推导，我们就得到了如下性质：

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

这个性质称为 weak duality，对于所有的优化问题都成立。其中 $p^* - d^*$ 称为 duality gap。需要注意的是，无论 primal problem 是什么形式，dual problem 总是一个凸优化问题，即它的极值是唯一的（如果存在的话）。

既然有 weak duality，显然就会有 strong duality。所谓的 strong duality，就是

$$d^* = p^*$$

这是一个很好的性质，在 strong duality 成立的条件下，我们可以通过求解 dual problem 来优化 primal problem。这里我们简要地介绍一下 strong duality 成立的条件。

如果 primal problem 是凸优化问题，strong duality 通常（但不总是）成立。而如果 primal problem 是凸优化问题并且满足 Slater 条件的话，那么 strong duality 成立。Slater 条件是指：**存在**一点 $x \in$ **relient** D 使得下式成立：

$$\begin{aligned} f_i(x) &< 0 \quad i = 1, \cdots, m \\ h_i(x) &= 0 \quad i = 1, \cdots, p \end{aligned}$$

满足上述条件的点有时称为严格可行，因为不等式约束严格成立。需要注意的是，这里只是指出了 strong duality 成立的一种情况，并不是唯一的情况。例如，对于某些非凸优化问题，strong duality 也成立。

接下来，我们来看看 strong duality 成立的时候的一些性质。假设 x^* 和 (λ^*, v^*) 分别是 primal problem 和 dual problem 的极值点，相应的极值为 p^* 和 d^*。此时，我们可以得到

$$
\begin{aligned}
p^* = f_0(x^*) = d^* &= \theta_D(\lambda^*, v^*) \\
&= \min_x \left(f_0(x) + \sum_{i=1}^{m} \lambda_i^* f_i(x) + \sum_{i=1}^{p} v_i^* h_i(x) \right) \\
&\le f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*) + \sum_{i=1}^{p} v_i^* h_i(x^*) \\
&\le f_0(x^*)
\end{aligned}
$$

由于两头是相等的，所以上式中的不等号都可以换成等号。根据第一个不等号，我们可以得到 x^* 是 $L(x, \lambda^*, v^*)$ 的一个极值点，由此可以知道 $L(x, \lambda^*, v^*)$ 在 x^* 的梯度应该等于0，即

$$
\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(x^*) = 0
$$

此外，由第二个不等号，我们可以得到

$$
\lambda_i^* f_i(x^*) = 0 \quad i = 1, \cdots, m
$$

上式也叫做 KKT 对偶互补条件(dual complementarity condition)。显然，如果 $\lambda^* > 0$，那么必有 $f_i(x^*) = 0$；反过来，如果 $f_i(x^*) < 0$，那么可以得到 $\lambda_i^* = 0$。（也就是说，$f_i(x) \le 0$ 存在的话，应该是相等关系，而不是不等关系）。在实际情况中，通常（不总是） $\lambda^* \ne 0$，当且仅当 $f_i(x^*) = 0$。在之后的学习中，这个等式很重要，尤其对于表明 SVM 只有少数的支持向量(support vectors)；在学习 SMO 算法的时候，还可以用 KKT 对偶互补条件来进行收敛性检测(convergence test)。

之后，再其他显而易见的条件写在一起，就是 KKT 条件(Karush-Kuhn-Tucker):

$$
\begin{aligned}
f_i(x^*) &\le 0 \quad i = 1, \cdots, m \quad (\nabla_\lambda L(x^*, \lambda, v) \le 0) \\
h_i(x^*) &= 0 \quad i = 1, \cdots, p \quad (\nabla_v L(x^*, \lambda, v) = 0) \\
\lambda_i^* &\ge 0 \quad i = 1, \cdots, m \\
\lambda_i^* f_i(x^*) &= 0 \quad i = 1, \cdots, m \\
\nabla f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla f_i(x^*) + \sum_{i=1}^{p} v_i^* \nabla h_i(x^*) &= 0 \quad (\nabla_x L(x^*, \lambda^*, v^*) = 0)
\end{aligned}
$$

对于目标函数和约束函数都可微的任意优化问题，如果 strong duality 成立（不一定要求是通过 Slater 条件得到，也不一定要求是凸优化问题），那么任何一对原问题最优解和对偶问题最优解必须满足 KKT 条件。换句话说，这是 strong duality 的一个必要条件。

不过，当 primal problem 是凸优化问题时（当然还要求目标函数和约束函数可微），KKT 就变为充要条件。也就是说，当 primal problem 是一个凸优化问题，且存在 $\tilde{x}$ 和 $(\tilde{\lambda}, \tilde{v})$ 满足 KKT 条件，那么它们分别是 primal problem 和 dual problem 的极值点并且 strong duality成立。

其证明也比较简单。前两个条件说明了，$\tilde{x}$ 是原问题的可行解。因为 $\tilde{\lambda}_i \ge 0$，$L(x, \tilde{\lambda}, \tilde{v})$ 是 x 的凸函数。最后一个 KKT 条件说明，$L(x, \tilde{\lambda}, \tilde{v})$ 关于 x 求极小在 $\tilde{x}$ 处取得最小值。我们得出结论

$$
\begin{aligned}
\theta_D(\tilde{\lambda}, \tilde{v}) &= \min_x L(x, \tilde{\lambda}, \tilde{v}) \\
&= L(\tilde{x}, \tilde{\lambda}, \tilde{v}) \\
&= f_0(\tilde{x}) + \sum_{i=1}^{m} \tilde{\lambda}_i^* f_i(\tilde{x}) + \sum_{i=1}^{P} \tilde{v}_i^* h_i(\tilde{x}) \\
&= f_0(\tilde{x})
\end{aligned}
$$

最后一行成立是因为 $h_i(\tilde{x}) = 0$ 以及 $\tilde{\lambda}_i f_i(\tilde{x}) = 0$。这说明 primal problem 的解 $\tilde{x}$ 和 dual problem 的解 $(\tilde{\lambda}, \tilde{v})$ 之间的 duality gap 为零，因此分别是 primal problem 和 dual problem 的解。

总之，对于目标函数和约束函数可微的任意**凸优化**问题，任意满足 KKT 条件的点，strong duality 成立。

小结：

1. 一个优化问题（不一定是凸优化问题），通过求出它的 dual problem，在只有 weak duality 成立的情况下，我们至少可以得到 primal problem 的一个下界。注意到，不管 primal problem 是一个什么形式，dual problem 始终是一个凸优化问题。
2. 在 strong duality 成立的情况下，可以通过求解 dual problem 来解决 primal problem。如果一个问题是凸优化问题，strong duality 通常（不总是）成立。而如果 primal problem 是凸优化问题，并且满足 Slater 条件，strong duality 成立。
3. 对于目标函数和约束函数可微的任意优化问题，由 strong duality 可以推出 KKT 条件。对于目标函数和约束函数可微的任意**凸**优化问题，任意满足 KKT 条件的点，它们分别是 primal problem 和 dual problem 的极值点并且 strong duality 成立。

5. Optimal margin classifiers

接下来的讨论用到的符号与上一节有些不同，在这里先进行说明。首先，上一节我们使用 λ_i 和 v_i 表示拉格朗日乘数，这里我们分别使用 α_i 和 β_i，由于 SVM 只有不等式约束，所以下面只用了 α_i。其次，上面我们使用 x 表示 primal problem 的参数，但在 SVM 中，我们有 w 和 b 两个参数。

回到之前我们对 SVM 的讨论，我们提出了如下凸优化问题：

$$
\begin{aligned}
\min_{w,b} \quad & \frac{1}{2}\|w\|^2 \\
s.t. \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \cdots, m
\end{aligned}
$$

我们可以把约束条件改写为：

$$
g_i(w, b) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0
$$

其中，对于训练集中每一个样本，我们都有一个这样的约束条件。

接下来，我们构造 Lagrangian ，得到：

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

因为该问题只有不等式约束，没有等式约束，所以这里的拉格朗日乘数只有 α_i，没有 β_i。

根据对偶问题的定义，我们首先要固定 α，让 $L(w,b,\alpha)$ 关于 w 和 b 求极小，从而得到 θ_D。具体方法就是令 L 关于 w 和 b 的导数为零。即

$$\nabla_w L(w,b,\alpha)=w-\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}=0$$
$$\frac{\partial}{\partial b}L(w,b,\alpha)=\sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

由上式我们可得：

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

将其代入 Lagrangian 可得：

$$\begin{aligned}
\min_{w,b}L(w,b,\alpha)&=\frac{1}{2}\|w\|^2-\sum_{i=1}^{m}\alpha_i[y^{(i)}(w^Tx^{(i)}+b)-1]\\
&=\frac{1}{2}w^Tw-w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i\\
&=\frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i\\
&=-\frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i\\
&=-\frac{1}{2}\left(\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}\right)^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i\\
&=-\frac{1}{2}\sum_{i=1}^{m}\alpha_i y^{(i)}(x^{(i)})^T\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i\\
&=-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i y^{(i)}(x^{(i)})^T\alpha_j y^{(j)}x^{(j)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i\\
&=-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j(x^{(i)})^Tx^{(j)}-b\sum_{i=1}^{m}\alpha_i y^{(i)}+\sum_{i=1}^{m}\alpha_i
\end{aligned}$$

由 $\frac{\partial}{\partial b}L(w,b,\alpha)=\sum_{i=1}^{m}\alpha_i y^{(i)}=0$ 可得：

$$\min_{w,b} L(w,b,\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}$$

最后，得到我们的对偶优化问题：

$$\begin{aligned} \max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j < x^{(i)}, x^{(j)} > \\ s.t. \quad & \alpha_i \geq 0, \quad i = 1, \cdots, m \\ & \sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \end{aligned}$$

这里我们把 $(x^{(i)})^T x^{(j)}$ 写成内积的形式 $< x^{(i)}, x^{(j)} >$。当使用核技巧 (kernel trick) 的时候，算法用内积的形式表达就非常重要了。

回到 SVM 的 priaml problem，这是一个凸优化问题并满足 Slater 条件。因为在这里，Slater 条件实际上是指存在一个超平面 (w, b) 将数据分隔开，也就是说“数据是线性可分的”。注意到 Slater 条件并不要求 (w, b) 是最优的，举例来说，如果线性数据是线性可分的，当我们找到最优的 (w^*, b^*) 后，成比例地放缩 w, b， 就找到使得不等式严格成立的 (w, b)。所以，SVM 满足 strong duality和 KKT 条件（注意，我们现在讨论的都是训练集线性可分的情况）。

这样，我们就可以通过解这个 dual problem 来解决 primal problem。具体来说，我们构建了一个以 α_i 为参数的取最大值问题。如果我们找到 α 使得 $W(\alpha)$取得最大值，并且满足约束条件，那么我们就可以利用等式 $w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$ 找到 w 的最优值。当我们得到 w^* 之后，再考虑主优化问题，就能直接找到截距项 b 的最优值：

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

假如我们已经使用了一个训练集对模型参数进行了拟合，接下来要对新输入的 x 进行预测。此时，我们需要计算 $w^T x + b$，如果结果大于0，那么就预测 $y = 1$。通过等式 $w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$，可以写成：

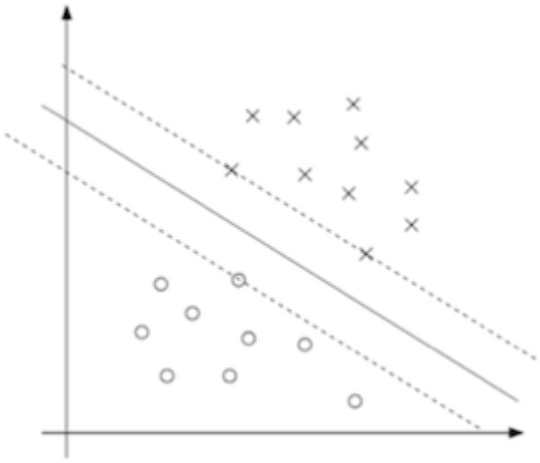
$$\begin{aligned} w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} \right)^T + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} < x^{(i)}, x > + b \end{aligned}$$

可以看到，上面的计算只依赖于新输入的 x 于训练集各点的内积。

最后，我们说说支持向量 (support vectors)。回忆我们之前提到的 KKT 对偶互补条件为：

$$\alpha_i^* g_i(w^*, b^*) = 0 \quad i = 1, \cdots, m$$

当 $\alpha_i^* \geq 0$时，$g(w^*, b^*) = 0$。也就是说，训练样本 $(x^{(i)}, y^{(i)})$ 函数间隔为1。如下图所示



假设上图是你的训练集和最优的超平面，所以函数间隔为1的训练样本 $(x^{(i)}, y^{(i)})$ 就是指上图那三个离超平面最近的训练样本。通常来说，这些函数间隔为1的训练样本所对应的 $\alpha_i \neq 0$。由上图可以看到，通常当我们找到最优化问题的解，只有少数的训练样本的函数间隔为1（图中只有三个）。我们称这些样本为支持向量 (support vectors)。由于支持向量很少，这也就意味着大多数情况下，$\alpha_i = 0$。

由于除了支持向量外，其他训练样本对应的 $\alpha_i = 0$。因此，$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$ 只依赖于很少一部分训练样本。当我们进行预测时，你只需要计算新输入的 x 与支持向量的内积，而它们只占整个训练集的一小部分。这是一个非常漂亮的性质。

当目前为止，我们都假设我们的训练集是线性可分的。接下来，我们会谈到核 (kernel) 以及针对训练集线性不可分的情况进行讨论。最后，我们会讲到 SMO 算法，该算法对于求解上述 SVM 的对偶问题十分有效。

参考资料：

1. 吴恩达 cs229 的上课视频以及讲义
2. pluskid的支持向量机系列 http://blog.pluskid.org/?page_id=683 (http://blog.pluskid.org/?page_id=683)
3. Convex Optimization, Stephen P. Boyd

In [ ]: