1.
(1) 可分意味着 $\exists \beta \in \mathbb{R}^{p+1}$ s.t. 对于 $\forall x_i$ 对 $y_i = -1$ 满足 $\beta^T x_i < 0$

对于 $\forall x_i$ $y_i = 1$ 满足 $\beta^T x_i > 0$

将 y_i 乘上得 $y_i \beta^T x_i > 0$ 由于 β 可以任意放缩倍数 从几何上表示平面 $\beta^T x = 0$ 不变

$\therefore$ 一定可取 $\beta_{sep} = 2\beta$ s.t. $y_i \beta_{sep}^T \frac{x_i}{\|x_i\|} \geq 1$ 其中 2 为放缩倍数

可由 $\{x_i, y_i\}$ 中 $y_i \beta^T \frac{x_i}{\|x_i\|}$ 最小的 sample 决定

(2) 已知 $\begin{cases} y_i \beta_{sep}^T z_i \geq 1 \\ y_i \beta_{old}^T z_i \leq 0 \end{cases}$ 求证 $\|\beta_{new} - \beta_{sep}\|^2 \leq \|\beta_{old} - \beta_{sep}\|^2 - 1$

有 $\|\beta_{new} - \beta_{sep}\|^2 = \|(\beta_{old} - \beta_{sep}) + y_i z_i\|^2 = \|\beta_{old} - \beta_{sep}\|^2 + y_i^2 z_i^T z_i + 2 y_i (\beta_{old} - \beta_{sep})^T z_i$

由于 $y_i = \pm 1$, $\|z_i\| = 1$ $y_i (\beta_{old} - \beta_{sep})^T z_i \leq 0 - 1 = -1$

$\therefore \|\beta_{new} - \beta_{sep}\|^2 = \|\beta_{old} - \beta_{sep}\|^2 + 1 + 2 y_i (\beta_{old} - \beta_{sep})^T z_i \leq \|\beta_{old} - \beta_{sep}\|^2 - 1$ □

2.
若两组 vector linear seperable. 设这两组 vector 的 convex hull 相交.

由两组 vector $\{x_i\}$ $\{y_i\}$ linear seperable 知. 它存在 $\beta \in \mathbb{R}^{p+1}$ s.t. (x_i 为以原点为起点的向量)

$\forall x_i$. $\begin{cases} \beta^T x_i > 0 \text{ ①} \\ \forall y_i \quad \beta^T y_i < 0 \text{ ②} \end{cases}$ 若 $z \in Conv(\{x_i\}) \cap conv(\{y_i\})$

有 $\exists \alpha_i$ s.t. $z = \sum_{i=1}^{\#\{x_i\}} \alpha_i x_i$

$\alpha_i \in [0,1]$ 且 $\sum \alpha_i = 1$

$\exists \beta_i$ 同上 s.t. $z = \sum_{i=1}^{\#\{y_i\}} \beta_i y_i$

且 $\sum \beta_i = 1$

那么由 ① 有 $\beta^T z = \beta^T (\sum_{i=1}^{\#\{x_i\}} \alpha_i x_i) = \sum_{i=1}^{\#\{x_i\}} \alpha_i \beta^T x_i > 0$

由 ② 有 $\beta^T z = \beta^T (\sum_{i=1}^{\#\{y_i\}} \beta_i x_i) = \sum_{i=1}^{\#\{y_i\}} \beta_i \beta^T y_i < 0$ // 矛盾!

$\therefore$ 若两组 vector linear separable, 那么 convex hull 一定不相交

其逆否命题: 若两组 vector 的 convex hull 相交, 则两组 vector 一定不 linear separable 得证.

再证明: 若两组 vector 的 convex hull 不相交. 那么 两组 vector linear separable.

这就是 seperating plane theorem. 可通过两个 convex hull 上距离最近的两点做垂直平分线证明 (easy)

3.

a) Lagrangian: $L(\{\alpha_i\}, w) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (w^T x^{(i)} - 1)$

primal problem: $\min_{w} \max_{\substack{\{\alpha_i\} \\ s.t. \alpha_i \geq 0}} L(\{\alpha_i\}, w)$

dual problem: $\max_{\substack{\{\alpha_i\} \\ s.t. \alpha_i \geq 0}} \min_{w} \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i (w^T x^{(i)} - 1)$

内部优化 $\min_w$ 取到最小值时，一定满足 $\frac{\partial L(\{\alpha_i\}, w)}{\partial w} = 0$

$\Rightarrow w - \sum_{i=1}^{n} \alpha_i x^{(i)} = 0 \Rightarrow w = \sum_{i=1}^{n} \alpha_i x^{(i)}$

dual problem 代入 $w = \sum_{i=1}^{n} \alpha_i x^{(i)}$ 有转化为:

$$\max_{\substack{\{\alpha_i\} \\ \alpha_i \geq 0}} \frac{1}{2} \sum_{i=1}^{n} \alpha_i \sum_{j=1}^{n} \alpha_i \alpha_j x^{(i)T} x^{(j)} - \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{n} \alpha_j x^{(j)T} x^{(i)} - 1 \right)$$

$$\Rightarrow \max_{\substack{\{\alpha_i\} \\ \alpha_i \geq 0}} -\frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j x^{(i)T} x^{(j)} + \sum_{i=1}^{n} \alpha_i$$

(b) 可以、由于训练中的所有 $x^{(i)}, x^{(j)}$ 的相似度都是内积形式，很容易可推广到高维空间的内积 $x^{(i)} \Rightarrow \phi(x^{(i)})$ 都是成对出现的

$x^{(j)} \Rightarrow \phi(x^{(j)})$ $\quad \phi(x^{(i)})^T \phi(x^{(j)}) = K(x^{(i)}, x^{(j)})$

可用一个原 feature 空间 $\mathcal{X} \times \mathcal{X}$ 上的 kernel 函数计算

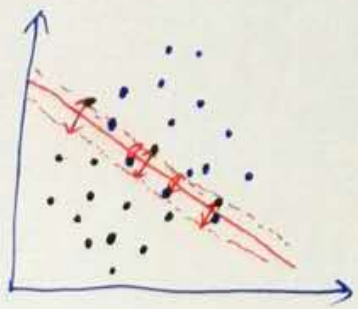
在 testing 时，求 $w^T x = \sum_{i=1}^{n} \alpha_i x^{(i)T} x$ 也可推广到高维空间的内积

可写作 $w^T \phi(x) = \sum_{i=1}^{n} \alpha_i K(x^{(i)}, x)$

4. 无约束形式为 $\min_w \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \max(0, 1-y_i w^T x_i)$

其中 $\max(0, 1-y_i w^T x_i) = \xi_i$ 为第 i 个训练数据的惩罚项.

一个直观解释如右图. 蓝色与黑色为两类样本, 红线是考虑分割个 w 确定的线性判决面. 所有不满足 $y_i w^T x_i \geq 1$ 的训练样本有个非0的正的惩罚项, 其值为该 sample 到相对应自己类的支撑面 $w^T x_i = +1$ 或 $w^T x_i = -1$ 的垂直距离(但没有除 $\|w\|$ normalize)



但优化时, 由于 max 函数不光滑. 一般将此问题转化成等价形式求解.

5. (1) 不直接存 $\theta^{(i)}$, 而是存 $\theta^{(i)} = \sum_{i=1}^{i} \beta_i \phi(x^{(i)})$ 中的 β_i 系数.

这个系数是容易计算的. 为 $\alpha[y^{(i+1)} - h_{\theta^{(i)}}(x^{(i)})]$

(2) 其中 $h_{\theta^{(i-1)}}(x^{(i)})$ 可用 kernel trick 求值

$$h_{\theta^{(i-1)}}(x^{(i)}) = \text{sgn}[\theta^{(i-1)T}\phi(x^{(i)})] = \text{sgn}[\sum_{j=1}^{i-1} \beta_j \phi(x^{(j)})^T \phi(x^{(i)})]$$

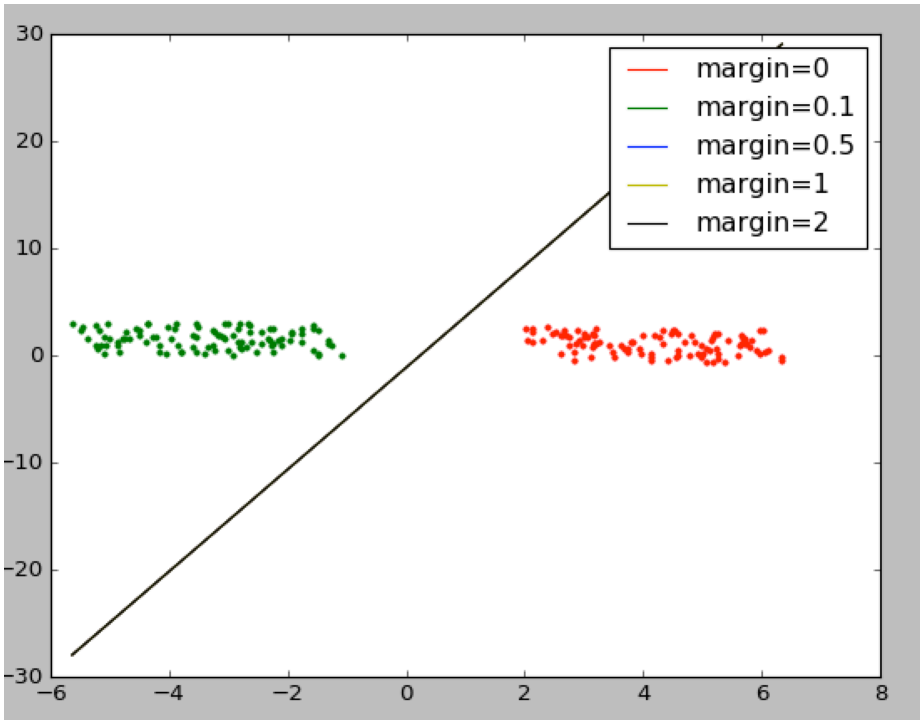
$$= \text{sgn}[\sum_{j=1}^{i-1} \beta_j K(x^{(j)}, x^{(i)})]$$ 即只需用 (1) 中存下的 β_j 乘上用 kernel 函数计算得的 $K(x^{(j)}, x^{(i)})_j$ 即可算出 $h_{\theta^{(i-1)}}(x^{(i)})$

(3) 此题其实上面已经写 … 记录下 $x^{(i+1)}$, 并记录下 $\beta^{(i+1)} = \alpha[y^{(i+1)} - \text{sgn}[(\sum_{j=1}^{i} \beta_j K(x^{(j)}, x^{(i+1)})]]$

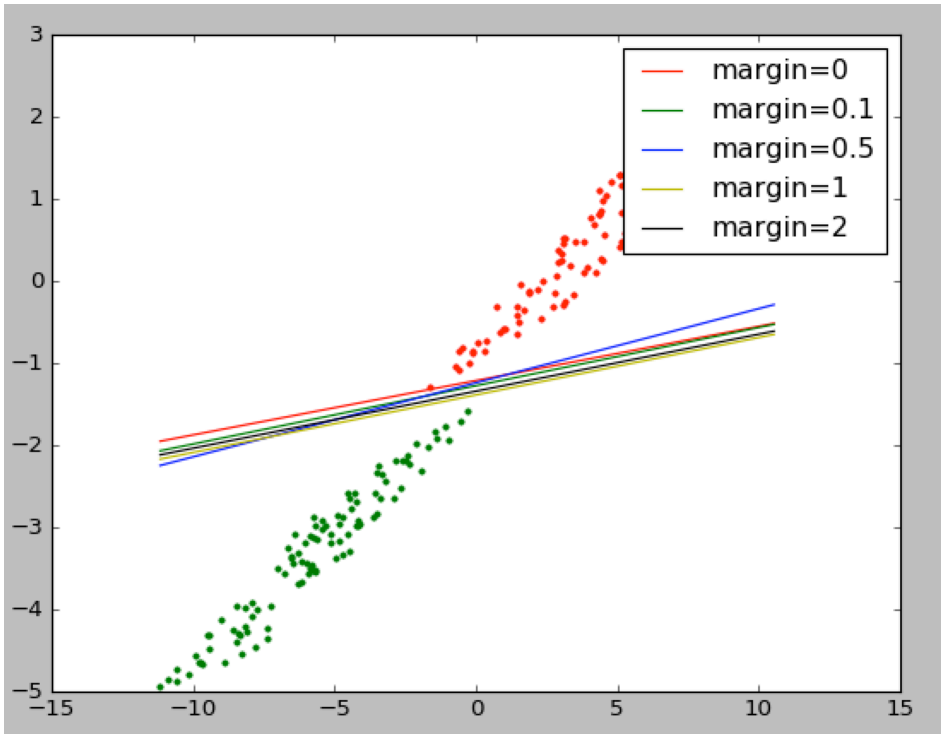
即可!

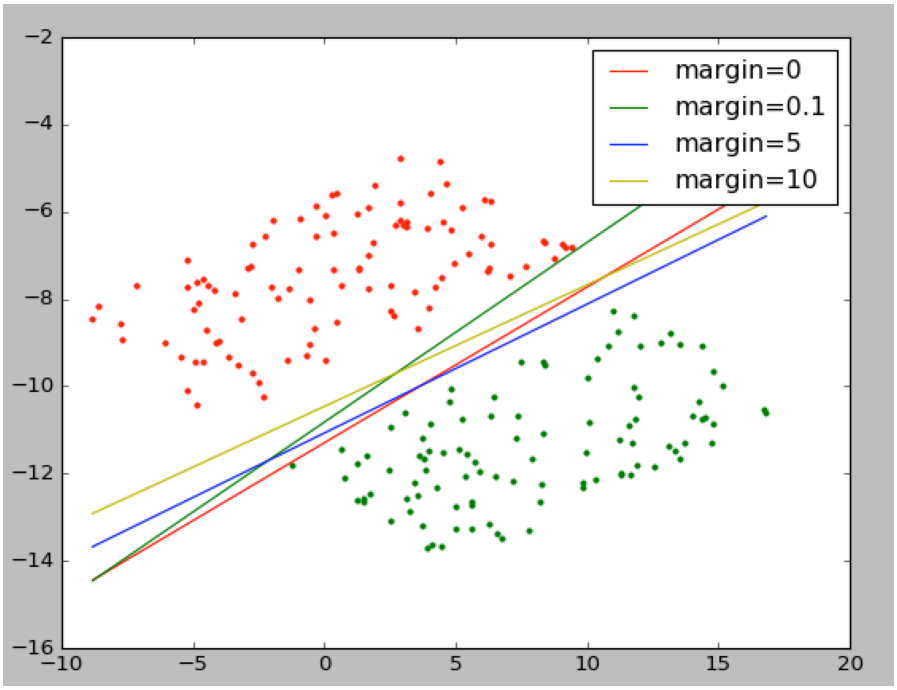
6. Perceptron 实验

当数据分布的比较开的时候，用不同的 margin 得到的结果十分相近



当数据分布比较集中的时候用不同的 margin 得到的结果会有差距，从夏眠两张图可以看出，margin=0 也就是无 margin 的 perceptron 得到的线(红线，经常会过红色或者绿色的点，而把 margin 设定大一点之后，线的质量会稍好一点，并且直观上也更接近于通过两类点簇的中心。





下表为上图不同 margin 训练时经过的 iter 数和训练结果：可以看出在 margin 较大时，虽然得到的分类面效果较好，但是需要更多的迭代次数才能收敛，而迭代次数越多，weight 的范数一般越大。

	迭代次数	w[0]	w[1]	b
Margin=0	1583	-2.9058	8.1443	92
Margin=0.1	1101	-2.8153	6.8373	74
Margin=5	2783	-4.2686	14.4502	160
Margin=10	4843	-6.1205	21.9846	230

7. SVM 实验

选用了 sklearn 库的 svm 实现
(参考 http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html)

由于这个库实现的是 soft SVM，为了模拟 hard SVM，把错误的 penalty 项设的十分大，在这里我设为 1e10，检查发现没有不可分的情况。对所有数据的实验结果如下：

Sample 个数	weight	b	margin
2	[0. 0.00050195 -0.00276071 -0.00250974 0.03119602 0.00406577]	-1.28164339	31.5613782737
4	[1.73472348e-18 -1.21109459e-02	-2.27402498	9.68457705142

	-7.38806252e-02 4.13510967e-02 2.21943909e-02 5.34272788e-02]		
6	[1.78893358e-18 -1.21103944e-02 -7.38800900e-02 4.13468195e-02 2.22001672e-02 5.34276434e-02]	-2.2741499	9.68464551975
8	[3.25260652e-19 -1.21106580e-02 -7.38849139e-02 4.13399760e-02 2.21863083e-02 5.34313513e-02]	-2.27348551	9.68467532232
10	[-5.42101086e-19 -1.21105557e-02 -7.38832668e-02 4.13440551e-02 2.21993150e-02 5.34303389e-02]	-2.2740744	9.68442075188
12	[1.30104261e-18 -1.11127510e-02 -9.88725208e-02 3.79044162e-02 2.39160883e-02 2.10434614e-02]	-2.04541467	8.99812672
14	[4.33680869e-19 -1.11119602e-02 -9.88633638e-02 3.79006775e-02 2.39078174e-02 2.10431619e-02]	-2.04492403	8.99904475727
16	[-8.67361738e-19 -1.11115698e-02 -9.88639095e-02 3.78986270e-02 2.39068700e-02 2.10395687e-02]	-2.04481736	8.99913684492
18	[1.30104261e-18 -1.11131653e-02 -9.88706481e-02 3.78990397e-02	-2.04474006	8.99848088711

	2.39027340e-02 2.10537960e-02]		
20	[-8.67361738e-19 -1.11120406e-02 -9.88635351e-02 3.78999287e-02 2.39074147e-02 2.10449994e-02]	-2.04490349	8.99903128959

由于转换后 feature 的第一维(也就是 1)的变化完全可以用 b 的变化来替代，而我们对||w||做最小化，所以第一维的 weight 的值十分接近 0。在转换后的空间中这组数据都是线性可分的。