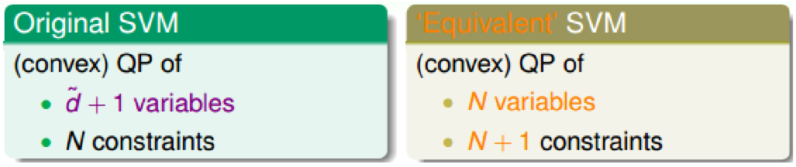
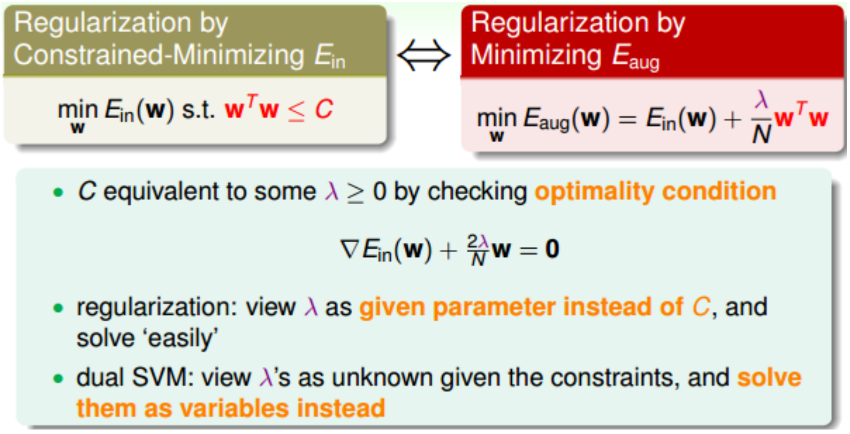
Dual Support Vector Machine

1.Motivation of Dual SVM



Original SVM二次规划问题的变量个数是$\hat{d}+1$，有N个限制条件；
Equivalent SVM二次规划变量个数为N个，有N+1个限制条件。
这种对偶SVM的好处就是问题只跟N有关，与$\hat{d}$无关，这样就不会出现当$\hat{d}$无限大时难以求解的情况。



Regularization中，在最小化E_{in}的过程中，也添加了限制条件：$w^Tw \le C$。我们的求解方法是引入拉格朗日因子λ，将有条件的最小化问题转换为无条件的最小化问题：$min\ E_{aug}(w) = E_{in}(w) + \frac{\lambda}{N}w^Tw$，最终得到的w的最优化解为：

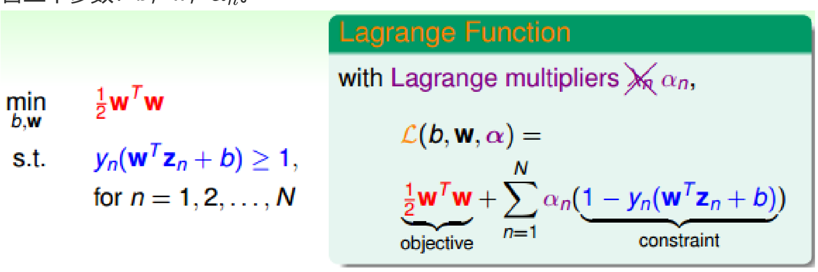
$$\nabla E_{in}(w) + \frac{2\lambda}{N}w = 0$$

所以，在regularization问题中，λ是已知常量，求解过程变得容易。那么，对于dual SVM问题，同样可以引入λ，将条件问题转换为非条件问题，只不过λ是未知参数，且个数是N，需要对其进行求解。

现在要将条件问题转化成非条件问题。
SVM中，目标是：$min\ \frac{1}{2}w^Tw$，条件是：$y_n(w^Tz_n+b) \ge 1,\ for\ n=1,2,\cdots,N$。首先，我们令拉格朗日因子为$\alpha_n$（区别于regularization），构造一个函数：

$$L(b,w,\alpha) = \frac{1}{2}w^Tw + \sum_{n=1}^{N}\alpha_n(1-y_n(w^Tz_n+b))$$

这个函数右边第一项是SVM的目标，第二项是SVM的条件和拉格朗日因子α_n的乘积。我们把这个函数称为拉格朗日函数，其中包含三个参数：b，w，α_n。



再利用拉格朗日函数，把SVM构成一个非条件问题。

Claim

$$\mathrm{SVM} \equiv \min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) = \min_{b,\mathbf{w}} \left(\infty \text{ if violate ; } \tfrac{1}{2}\mathbf{w}^T\mathbf{w} \text{ if feasible} \right)$$

- any 'violating' $(b, \mathbf{w})$: $\max_{\text{all } \alpha_n \geq 0} \left(\square + \sum_n \alpha_n (\text{some positive}) \right) \to \infty$
- any 'feasible' $(b, \mathbf{w})$: $\max_{\text{all } \alpha_n \geq 0} \left(\square + \sum_n \alpha_n (\text{all non-positive}) \right) = \square$

首先我们规定拉格朗日因子$\alpha_n \geq 0$，根据SVM的限定条件可得：$(1 - y_n(w^T z_n + b)) \leq 0$
如果没有达到最优解，即有不满足$(1 - y_n(w^T z_n + b)) \leq 0$的情况，因为$\alpha_n \geq 0$，那么必然有
$\sum_n \alpha_n (1 - y_n(w^T z_n + b)) \geq 0$。
对于这种大于零的情况，其最大值是无解的。

如果对于所有的点，均满足$(1 - y_n(w^T z_n + b)) \leq 0$，那么必然有$\sum_n \alpha_n (1 - y_n(w^T z_n + b)) \leq 0$
则当$\sum_n \alpha_n (1 - y_n(w^T z_n + b)) = 0$时，其有最大值，最大值就是我们SVM的目标：$\frac{1}{2} w^T w$。
因此，这种转化为非条件的SVM构造函数的形式是可行的。

2.Lagrange Dual SVM

现在SVM问题已经转化为与拉格朗日因子α_n有关的最大最小值形式。已知$\alpha_n \geq 0$，那么对于任何固定的α'，且$\alpha'_n \geq 0$，一定有如下不等式成立：

for any fixed $\boldsymbol{\alpha}'$ with all $\alpha'_n \geq 0$,

$$\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) \geq \min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}')$$

对上述不等式右边取max，不等式依然成立。

for best $\boldsymbol{\alpha}' \geq \mathbf{0}$ on RHS,

$$\min_{b,\mathbf{w}} \left(\max_{\text{all } \alpha_n \geq 0} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}) \right) \geq \underbrace{\max_{\text{all } \alpha'_n \geq 0} \min_{b,\mathbf{w}} \mathcal{L}(b, \mathbf{w}, \boldsymbol{\alpha}')}_{\text{Lagrange dual problem}}$$

上述不等式表明，我们对SVM的min和max做了对调，满足这样的关系，这叫做Lagrange dual problem。不等式右边是SVM问题的下界，我们接下来的目的就是求出这个下界。

已知≥是一种弱对偶关系，在二次规划QP问题中，如果满足以下三个条件:
1.函数是凸的（convex primal）
2.函数有解（feasible primal）
3.条件是线性的（linear constraints）

那么，上述不等式关系就变成强对偶关系，≥变成=，即一定存在满足条件的解(b, w, α)，使等式左边和右边都成立，SVM的解就转化为右边的形式。

经过推导，SVM对偶问题的解已经转化为无条件形式：

$$\max_{\text{all } \alpha_n \geq 0} \left(\underbrace{\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \sum_{n=1}^{N} \alpha_n (1 - y_n(\mathbf{w}^T\mathbf{z}_n + b))}_{\mathcal{L}(b,\mathbf{w},\boldsymbol{\alpha})} \right)$$

上式括号里面的是对拉格朗日函数$L(b, w, \alpha)$计算最小值。
那么根据梯度下降算法思想:最小值位置满足梯度为零。
首先，令$L(b, w, \alpha)$对参数b的梯度为零:

$$\frac{\partial L(b,w,\alpha)}{\partial b}=0=-\sum_{n=1}^{N}\alpha_n y_n$$

那么，我们把这个条件代入计算max条件中（与$\alpha_n\geq 0$同为条件），并进行化简：

$$\max_{\text{all } \alpha_n\geq 0, \sum y_n\alpha_n=0}\left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w}+\sum_{n=1}^{N}\alpha_n(1-y_n(\mathbf{w}^T\mathbf{z}_n))\right)$$

这样，SVM表达式成功消去了b。
现在，令$L(b,w,\alpha)$对参数w的梯度为零:

$$\frac{\partial L(b,w,\alpha)}{\partial w}=0=w-\sum_{n=1}^{N}\alpha_n y_n z_n$$

同样把这个条件带入原式并化简：

$$\max_{\text{all } \alpha_n\geq 0, \sum y_n\alpha_n=0, \mathbf{w}=\sum\alpha_n y_n \mathbf{z}_n}\left(\min_{b,\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w}+\sum_{n=1}^{N}\alpha_n-\mathbf{w}^T\mathbf{w}\right)$$

$$\Longleftrightarrow \max_{\text{all } \alpha_n\geq 0, \sum y_n\alpha_n=0, \mathbf{w}=\sum\alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n y_n\mathbf{z}_n\right\|^2+\sum_{n=1}^{N}\alpha_n$$

这样，SVM表达式成功消去了w。问题更加简化，这时候的条件有三个：

1. all $\alpha_n\geq 0$
2. $\sum_{n=1}^{N}\alpha_n y_n=0$
3. $w=\sum_{n=1}^{N}\alpha_n y_n z_n$

SVM简化为只有α_n的最佳化问题，即计算满足上述三个条件下，函数$-\frac{1}{2}||\sum_{n=1}^{N}\alpha_n y_n z_n||^2+\sum_{n=1}^{N}\alpha_n$最小值时对应的$\alpha_n$是多少。

总结一下，SVM最佳化形式转化为只与α_n有关：

$$\max_{\text{all } \alpha_n\geq 0, \sum y_n\alpha_n=0, \mathbf{w}=\sum\alpha_n y_n \mathbf{z}_n} -\frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n y_n\mathbf{z}_n\right\|^2+\sum_{n=1}^{N}\alpha_n$$

其中，满足最佳化的条件称之为Karush-Kuhn-Tucker(KKT)：

if primal-dual optimal $(b,\mathbf{w},\alpha)$,

- primal feasible: $y_n(\mathbf{w}^T\mathbf{z}_n+b)\geq 1$
- dual feasible: $\alpha_n\geq 0$
- dual-inner optimal: $\sum y_n\alpha_n=0$; $\mathbf{w}=\sum\alpha_n y_n\mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1-y_n(\mathbf{w}^T\mathbf{z}_n+b))=0$$

—called Karush-Kuhn-Tucker (KKT) conditions, necessary for optimality [& sufficient here]

3.Solving Dual SVM

将max问题转化为min问题，再做一些条件整理和推导。

standard hard-margin SVM dual

$$\min_{\alpha} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} y_n\alpha_n = 0;$$

$$\alpha_n \geq 0, \text{for } n = 1, 2, \ldots, N$$

(convex) QP of N variables & $N+1$ constraints, as promised

optimal $\alpha = ?$

$$\min_{\alpha} \quad \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n\alpha_m y_n y_m \mathbf{z}_n^T\mathbf{z}_m - \sum_{n=1}^{N}\alpha_n$$

$$\text{subject to} \quad \sum_{n=1}^{N} y_n\alpha_n = 0; \quad \alpha_n \geq 0, \text{ for } n = 1, 2, \ldots, N$$

optimal $\alpha \leftarrow \mathrm{QP}(\mathrm{Q}, \mathbf{p}, \mathrm{A}, \mathbf{c})$

$$\min_{\alpha} \quad \frac{1}{2}\alpha^T\mathrm{Q}\alpha + \mathbf{p}^T\alpha$$

$$\text{subject to} \quad \mathbf{a}_i^T\alpha \geq c_i, \text{ for } i = 1, 2, \ldots$$

- $q_{n,m} = y_n y_m \mathbf{z}_n^T\mathbf{z}_m$
- $\mathbf{p} = -\mathbf{1}_N$
- $\mathbf{a}_{\geq} = \mathbf{y}$, $\mathbf{a}_{\leq} = -\mathbf{y}$; $\mathbf{a}_n^T = n$-th unit direction
- $c_{\geq} = 0$, $c_{\leq} = 0$; $c_n = 0$

显然，这是一个convex的QP问题，且有N个变量α_n，限制条件有N+1个。用QP解法，找到Q，p，A，c对应的值，用软件工具包进行求解即可。

- $q_{n,m} = y_n y_m \mathbf{z}_n^T\mathbf{z}_m$, often non-zero
- if $N = 30,000$, dense $\mathrm{Q_D}$ (N by N symmetric) takes > 3G RAM
- need **special solver** for
  - not storing whole $\mathrm{Q_D}$
  - utilizing special constraints properly

  to scale up to large N

$q_{n,m} = y_n y_m z_n^T z_m$，大部分值是非零的，称为dense。
当N很大的时候，那么对应的Q_D的计算量将会很大，存储空间也很大。

KKT conditions

if primal-dual optimal $(b, \mathbf{w}, \alpha)$,

- primal feasible: $y_n(\mathbf{w}^T\mathbf{z}_n + b) \geq 1$
- dual feasible: $\alpha_n \geq 0$
- dual-inner optimal: $\sum y_n\alpha_n = 0$; $\mathbf{w} = \sum \alpha_n y_n \mathbf{z}_n$
- primal-inner optimal (at optimal all 'Lagrange terms' disappear):

$$\alpha_n(1 - y_n(\mathbf{w}^T\mathbf{z}_n + b)) = 0 \text{ (complementary slackness)}$$

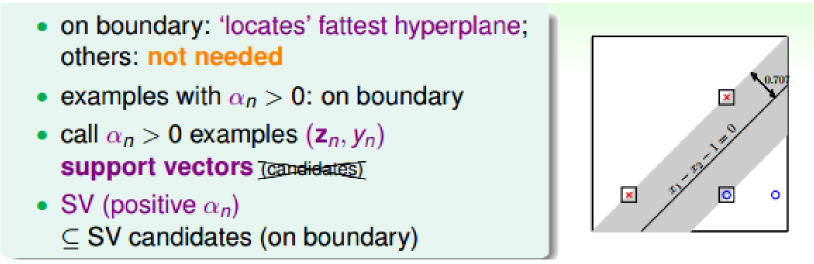
- optimal $\alpha \Longrightarrow$ optimal $\mathbf{w}$? easy above!
- optimal $\alpha \Longrightarrow$ optimal b? a range from primal feasible & equality from **comp. slackness** if one $\alpha_n > 0 \Rightarrow b = y_n - \mathbf{w}^T\mathbf{z}_n$

得到α_n之后，再根据之前的KKT条件，就可以计算出w和b了。
首先利用条件$w = \sum \alpha_n y_n z_n$得到w，然后利用条件$\alpha_n(1 - y_n(w^T z_n + b)) = 0$，取任一$\alpha_n \neq 0$即$\alpha_n$>0的点，得到

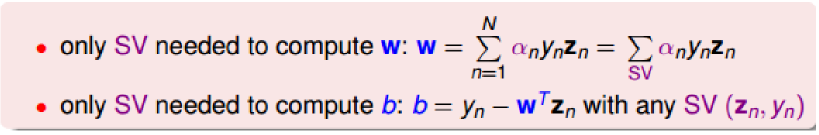
$1 - y_n(w^T z_n + b) = 0$
进而求得$b = y_n - w^T z_n$。

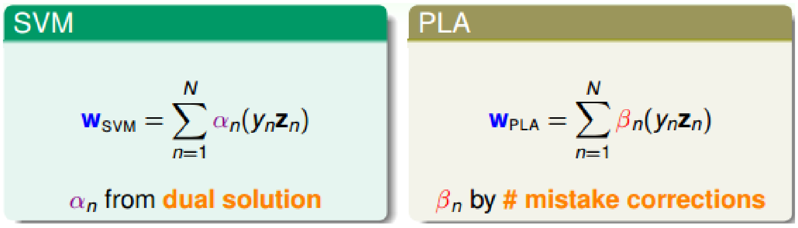
值得注意的是，计算b值，α_n>0时，有$y_n(w^T z_n + b) = 1$成立。$y_n(w^T z_n + b) = 1$正好表示的是该点在SVM分类线上，即fat boundary。也就是说，满足α_n>0的点一定落在fat boundary上，这些点就是Support Vector。这是一个非常有趣的特性。

4.Messages behind Dual SVM

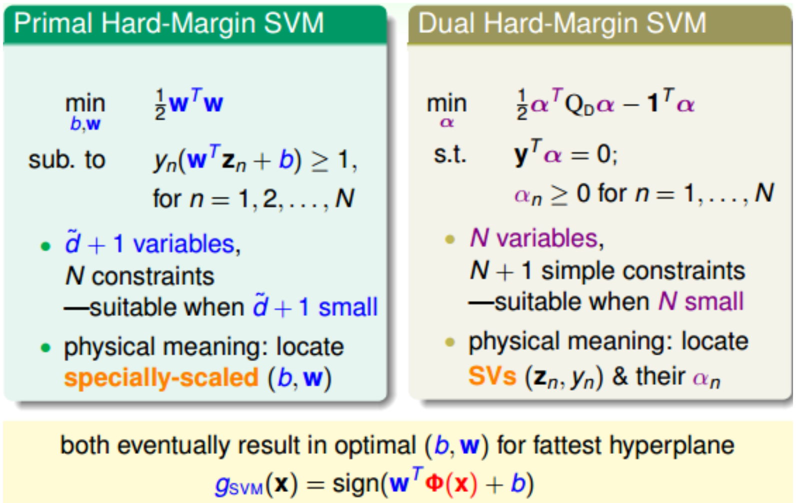


把位于分类线边界上的点称为support vector（candidates）。
α_n>0的点一定落在分类线边界上，这些点称之为support vector
也就是说分类线上的点不一定都是支持向量，但是满足α_n>0的点，一定是支持向量。





w_{SVM}由fattest hyperplane边界上所有的SV决定，w_{PLA}由所有当前分类错误的点决定。
w_{SVM}和w_{PLA}都是原始数据点$y_n z_n$的线性组合形式，是原始数据的代表。



总结一下，本节课和上节课主要介绍了两种形式的SVM
一种是Primal Hard-Margin SVM，另一种是Dual Hard_Margin SVM。Primal Hard-Margin SVM有$\hat{d}+1$个参数，有N个限制条件。当$\hat{d}+1$很大时，求解困难。
而Dual Hard_Margin SVM有N个参数，有N+1个限制条件。当数据量N很大时，也同样会增大计算难度。
两种形式都能得到w和b，求得fattest hyperplane。通常情况下，如果N不是很大，一般使用Dual SVM来解决问题。

总结

本节课主要介绍了SVM的另一种形式：Dual SVM。我们这样做的出发点是为了移除计算过程对$\hat{d}$的依赖。Dual SVM的推导过程是通过引入拉格朗日因子α，将SVM转化为新的非条件形式。然后，利用QP，得到最佳解的拉格朗日因子α。再通过KKT条件，计算得到对应的w和b。最终求得fattest hyperplane。下一节课，我们将解决Dual SVM计算过程中对$\hat{d}$的依赖问题。