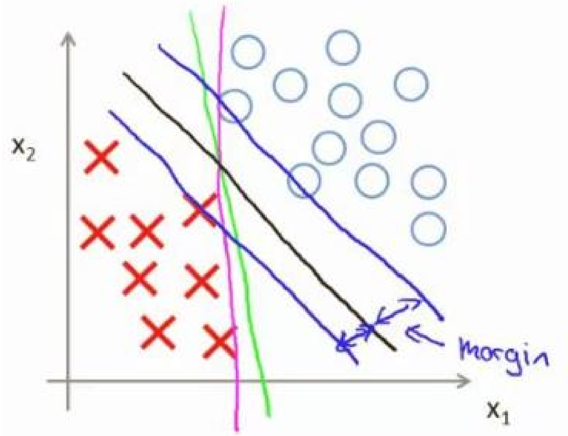
17.5.1（SVM浅析）

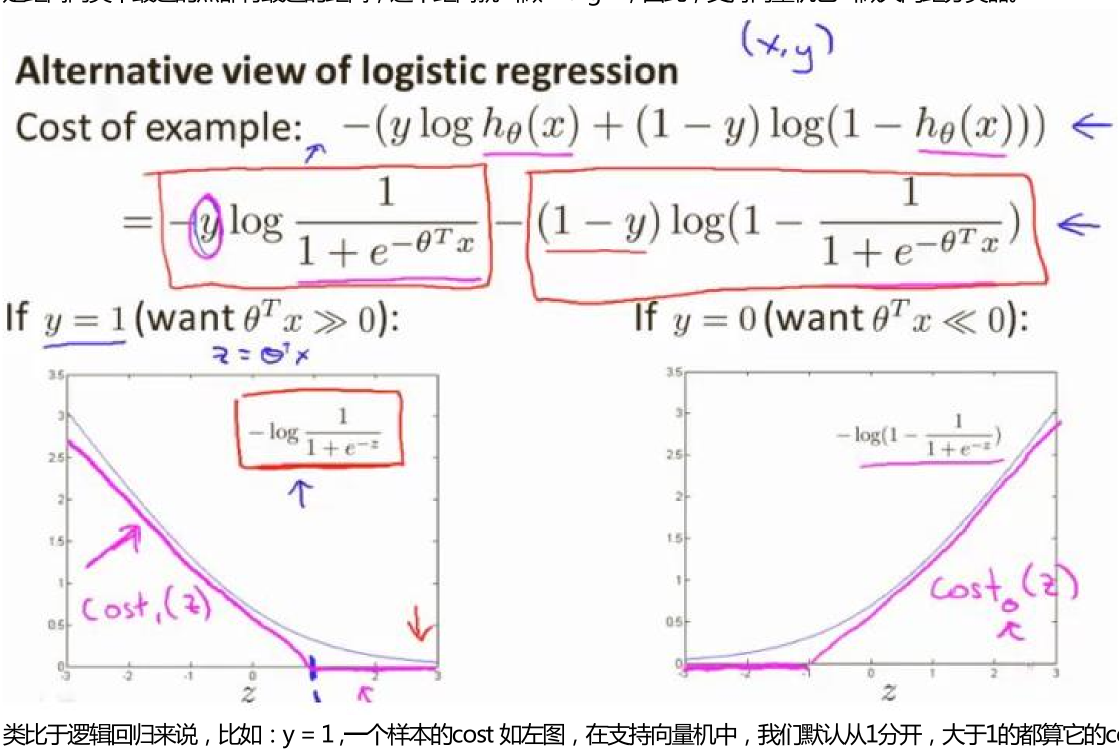
SVM 支持向量机（Support Vector Machines）

直观理解SVM：https://www.zhihu.com/question/21094489

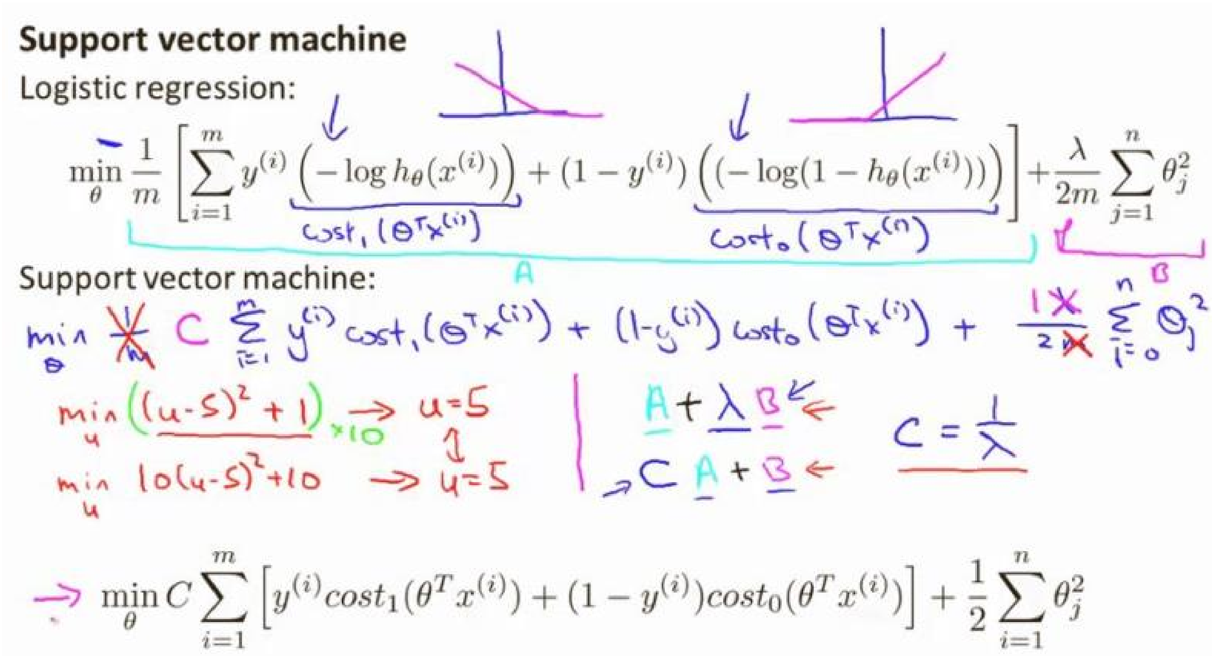
分类就是建一个分类函数或分类模型，对未知的进行预测分类。支持向量机就是把要分类的问题用超平面进行分类。支持向量机是监督学习算法中的分类算法，比如：把二维平面上的两类点分成两类，我们就可以把这些点投影到三维空间，然后用一个超平面把他们切成两类，然后把这个切面再投影到二维平面上就是一个很好地决策分界线。



图中的三条线都能把平面内的两类给分开，但是明显的黑色的线分类效果更好，因为黑色的线是在两类中离得最近的点的中间位置，也就是距离两类中最近的点都有最远的距离，这个距离就叫做 margin，因此，支持向量机也叫做大间距分类器。



类比于逻辑回归来说，比如：y = 1 ,一个样本的cost 如左图，在支持向量机中，我们默认从1分开，大于1的都算它的cost 为0，小于1的部分就画一条接近逻辑回归图像的直线，这样就把SVM中y=1情况下的cost写出来了。当y=0.时，cost 如右图。



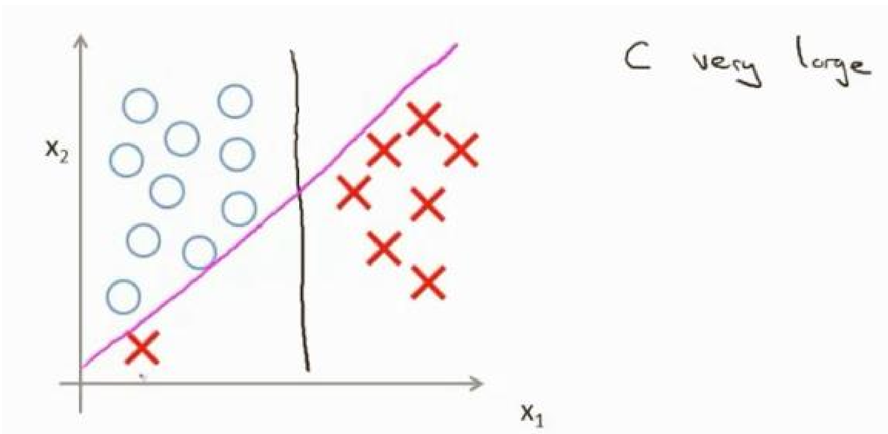
说到底，就是最小化cost 问题，原来的逻辑回归函数前面的1/m变成C，后面的正则项的lambda去掉，就变成SVM的cost function。C就相当于1/lambda。C的作用就是调节两项的权重问题。

回顾 C=1/λ，因此：

C 较大时，相当于 λ 较小，可能会导致过拟合，高方差。

C 较小时，相当于 λ 较大，可能会导致低拟合，高偏差。

一个好的C值可以帮助SVM很好地处理异常点问题。当C非常大时就是下图中的紫红线，当C取得恰到好处时就是图中的黑线，如下图：



SVM决策边界

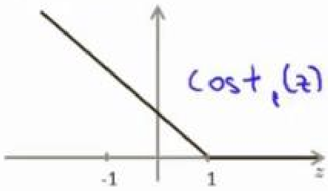
求cost 最优解的两个约束。

SVM Decision Boundary

$$\min_{\theta} \underset{\nearrow}{C} \underbrace{\sum_{i=1}^{m}\left[y^{(i)}cost_1(\theta^T x^{(i)}) + (1-y^{(i)})cost_0(\theta^T x^{(i)})\right]}_{=0} + \frac{1}{2}\sum_{i=1}^{n}\theta_j^2$$

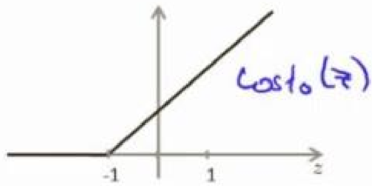
Whenever $y^{(i)} = 1$:

$\theta^T x^{(i)} \geq 1$



Whenever $y^{(i)} = 0$:

$\theta^T x^{(i)} \leq -1$



在最小化cost function的时候，我们想要让SVM的cost function的第一项为0，以便求出最优解。在此情况成立，上图中给出了两种解释。也就 是在上面画出的两个cost 的图像解释。由此，把SVM的cost function简化成下图：

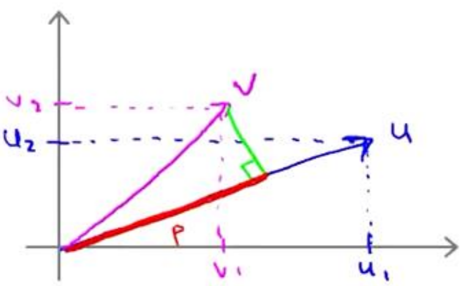
SVM Decision Boundary

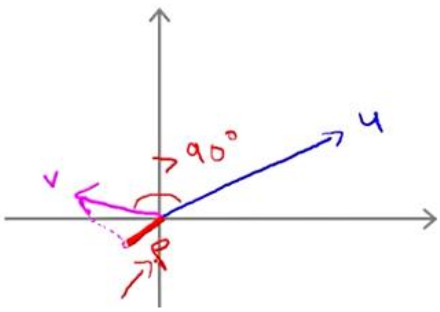
$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n}\theta_j^2$$

s.t. $\theta^T x^{(i)} \geq 1$ if $y^{(i)} = 1$

$\theta^T x^{(i)} \leq -1$ if $y^{(i)} = 0$

Vector Inner Product





$\rightarrow u = \begin{bmatrix}u_1\\u_2\end{bmatrix}$ $\rightarrow v = \begin{bmatrix}v_1\\v_2\end{bmatrix}$

$u^T v = ?$ $[u_1 \ \ u_2]\begin{bmatrix}v_1\\v_2\end{bmatrix}$

$\|u\|$ = length of vector u

$= \sqrt{u_1^2 + u_2^2} \in \mathbb{R}$

p = length of projection of v onto u.

Signed

$u^T v = \underline{p} \cdot \underline{\|u\|} \leftarrow \qquad = v^T u$

$= u_1 v_1 + u_2 v_2 \leftarrow \qquad p \in \mathbb{R}$

$u^T v = p \cdot \|u\|$

$p < 0$

向量内积的两种计算方式，1.用v在u上的投影长度p（p是有方向的，看v 和u 的夹角）乘u的范数（长度），或者反过来，结果是一个实数。2.是v和u元素对元素相乘再相加。

SVM被称为大间距分类器的原因：首先由上面两图可以把theta的转置乘x，换成p乘theta的范数。

求cost最小，就是求theta的范数最小，所以，要首先满足y的两个限制，也就是让p最大，才能让theta的范数最小。假设theta0=0，也就是决策线过原点，下图左图的决策边界线所示，这个情况下，离决策线最近的样本投影到theta向量上的p很小，所以为了满足上面两个限制，theta要很大。下图右图的决策边界线所示，这个情况下，离决策线最近的样本投影到theta向量上的p相对很大，所以为了满足上面两个限制，theta要很小，也就求出了最小的cost。大边距就是让图中的margin最大。

SVM Decision Boundary

$$\min_{\theta} \frac{1}{2}\sum_{j=1}^{n} \theta_j^2 = \frac{1}{2}\|\theta\|^2$$

s.t. $\boxed{p^{(i)} \cdot \|\theta\| \geq 1}$ if $y^{(i)} = 1$

$p^{(i)} \cdot \|\theta\| \leq -1$ if $y^{(i)} = 1$

C very large

where $p^{(i)}$ is the projection of $x^{(i)}$ onto the vector θ.

Simplification: $\theta_0 = 0$