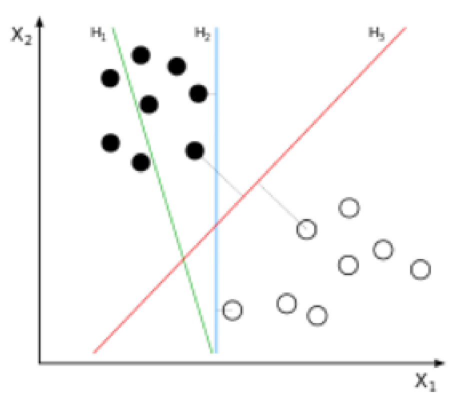
svm

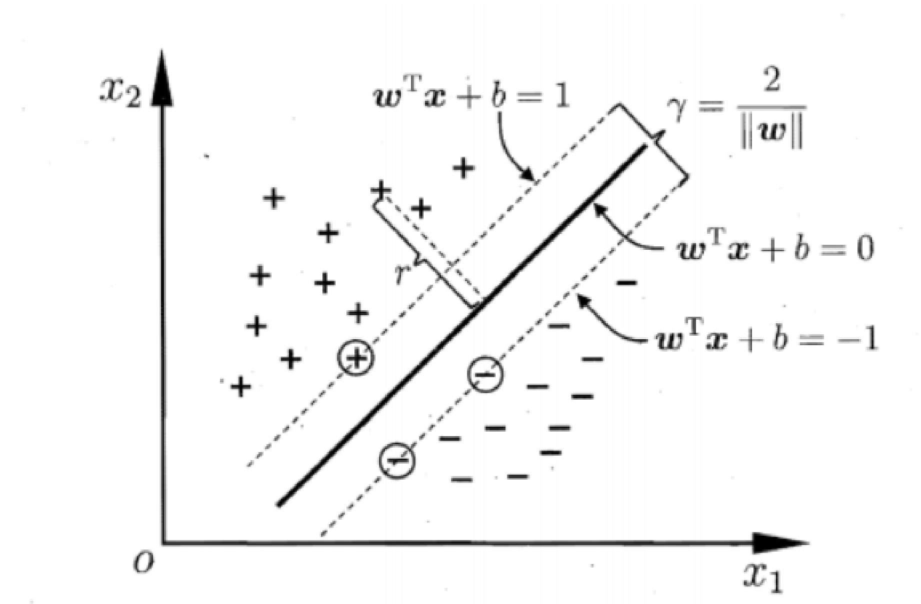
支持向量机（support vector machines，SVM）

一种二分类模型，它将实例的特征向量映射为空间中的一些点，SVM 的目的就是想要画出一条线，以“最好地” 区分这两类点，以至如果以后有了新的点，这条线也能做出很好的分类。SVM 适合中小型数据样本、非线性、高维的分类问题。

将实例的特征向量（以二维为例）映射为空间中的一些点，如下图的实心点和空心点，它们属于不同的两类。SVM 的目的就是想要画出一条线，以“最好地”区分这两类点，以至如果以后有了新的点，这条线也能做出很好的分类。



hard-margin SVM



划分超平面可以定义为一个线性方程：$\mathrm{w^TX+b=0}$，其中：

- $\mathrm{w=\{w_1;w_2;...;w_d\}}$ 是一个法向量，决定了超平面的方向，d 是特征值的个数
- X 为训练样本
- b 为位移项，决定了超平面与原点之间的距离

只要确定了法向量 w 和位移 b，就可以唯一地确定一个划分超平面。划分超平面和它两侧的边际超平面上任意一点的距离为 $\frac{1}{||\mathrm{w}||}$。

利用一些数学推导，公式 $\mathrm{y_i*(w_0+w_1x_1+w_2x_2)\geq 1, \forall i}$ 可变为有限制的凸优化问题(convex quadratic optimization)

利用 Karush-Kuhn-Tucker (KKT)条件和拉格朗日公式，可以推出 MMH 可以被表示为以下“决定边界 (decision boundary)”

$$\mathrm{d(X^T)=\sum_{i=1}^{l} y_i\alpha_iX_iX^T+b_0}$$

此方程就代表了边际最大化的划分超平面。

- l 是支持向量点的个数，因为大部分的点并不是支持向量点，只有个别在边际超平面上的点才是支持向量点。那么我们就只对属于支持向量点的进行求和；
- $\mathrm{X_i}$ 为支持向量点的特征值；
- $\mathrm{y_i}$ 是支持向量点$\mathrm{X_i}$的类别标记（class label），比如+1还是-1；
- $\mathrm{X^T}$ 是要测试的实例，想知道它应该属于哪一类，把它带入该方程
- α_i 和 $\mathrm{b_0}$ 都是单一数值型参数，由以上提到的最优算法得出，α_i 是拉格朗日乘数。

每当有新的测试样本 X，将它带入该方程，看看该方程的值是正还是负，根据符号进行归类。