HUAWEI

# 4.1 机器学习中的分类问题（下）

# 目录

Contents

# 分类模型基本评估指标

分类模型评估指标可以定量的对模型的效果进行分析，对模型“准确率”进行准确的评估。

定义：假定一个二分类问题，标签是<0，1>两种。如果是一个多分类问题，则可以站在某一个类别标签的角度看待其它所有的类别标签都可以归为“其它类”的范畴，将多分类转换为二分类。各种评估指标如表所示。

| 评估指标 | 计算公式 |
| --- | --- |
| 准确率（accuracy） | $TP + TN / P + N$ |
| 错误率（error rate） | $FP + FN / P + N$ |
| 召回率（recall） | $TP / P$ |
| 真负例率（specificity） | $TN / N$ |
| 精度（precision） | $TP / TP + FP$ |
| F分数 | $2 \times \text{precision} \times \text{recall} / \text{precision} + \text{recall}$ |
| $F_\beta$分数 | $(1 + \beta^2) \times \text{precision} \times \text{recall} / \beta^2 \times \text{precision} + \text{recall}$ |

说明：假定站在标签0的角度。

P：标签为0的样本个数

N：标签为1的样本个数

TP：标签为0且模型判定为0的样本个数

TN：标签为1且模型判定为1的样本个数

FP：标签为1且模型判定为0的样本个数

FN：标签为0且模型判定为1的样本个数

$\beta$：非负实数，为了赋予precision和recall不同的权重，一般常用的$\beta$值是2和0.5。

# 分类基本评估指标的使用

分类模型评估指标的使用跟具体的业务场景相关，而且因为可以站在不同类别标签的角度去计算指标，这样就出产生（7 x 类别个数）个基本评估指标，需要在具体的业务场景中进行分析。

业务场景举例1：银行判定贷款申请者有无风险。此业务场景下，银行既希望能够扩大贷款业务、又希望降低坏账风险，所以对于“有风险”和“无风险”两种类别的判定评估都需要考虑，各项指标（错误率除外）都要尽可能的高。

业务场景举例2：某制造产线判定某生产设备有无故障。此业务场景下，首先要考虑的是故障的判定要准，所以以故障类别计算的召回率要接近100%，在此前提下，specificity要尽可能的高，越高越好。

评估指标的使用原则：

1、确立标签重要性。要确定哪一个类别是要非常关注的类别。

2、重要类别的召回率、精度需要制定一个高的标准。比如必须达到xx%。

3、非重要类别的指标尽可能的高。

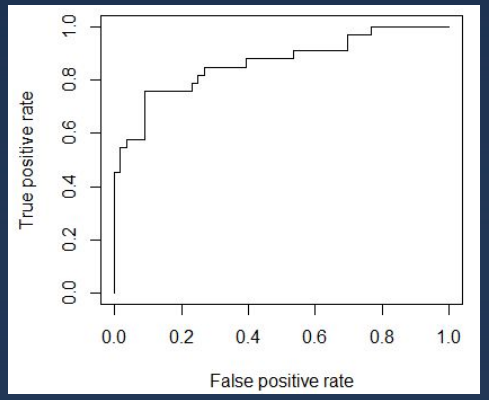
# 分类模型其它评估指标

1、混淆矩阵：当面临一个多分类问题，且每个类别的权重几乎等同时，利用混淆矩阵进行模型评估。

| 单位为样本个数 | 预测类别 | | | |
|---|---|---|---|---|
| 实际类别 | | 类别1 | 类别2 | 类别3 |
| | 类别1 | 500 | 10 | 5 |
| | 类别2 | 2 | 300 | 1 |
| | 类别3 | 3 | 4 | 200 |

注：矩阵对角线上的值越大越好，其它位置的值越小越好

2、ROC曲线：当一些分类器（比如基于神经网络）给出的判定值并不是类别标签，而是一些数值，那么就需要给这些数值一个阈值去产生类别，阈值不同，则类别判定结果不同，这样就产生了基于阈值的recall和（1-specificity），每一对这样的两个值看做平面上的一个点，多个阈值产生的点相连就产生了ROC曲线。ROC曲线越凸，表明模型效果越好。




3、AUC值：AUC值为ROC曲线所覆盖的区域面积，AUC越大，分类器分类效果越好。
AUC = 1：是完美分类器，采用这个预测模型时，不管设定什么阈值都能得出完美预测。绝大多数预测的场合，不存在完美分类器。
0.5 < AUC < 1：优于随机猜测。这个分类器（模型）妥善设定阈值的话，能有预测价值。
AUC = 0.5：跟随机猜测一样（例：丢铜板），模型没有预测价值。
AUC < 0.5：比随机猜测还差；但只要总是反预测而行，就优于随机猜测。

# 目录

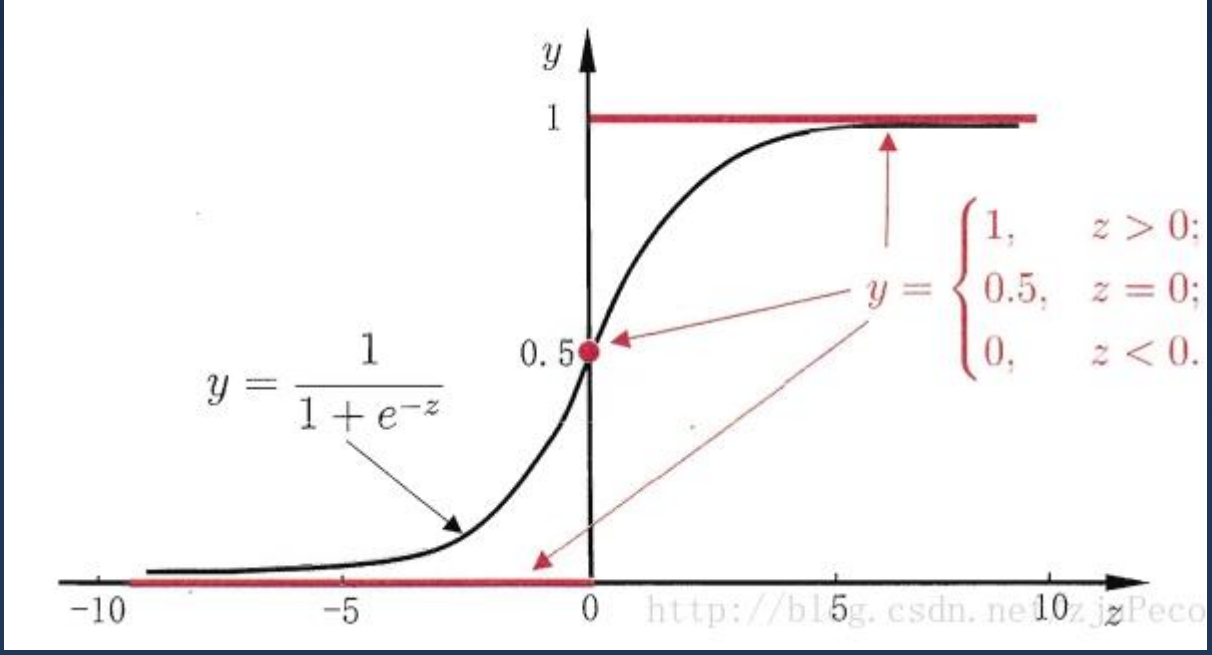
Contents

# 逻辑回归概念

逻辑回归是指Logistic Regression，简称LR。LR是一种回归分析方法，但常用在分类问题中。

在二分类问题中，假定类别与类别之间差异巨大，将类别标签定义为<0，1>，赋予0和1数值意义，希望存在一个函数f，能够根据特征计算出位于区间[0，1]的值，然后根据阈值赋予类别标签。

一个看起来非常完美的函数：Sigmoid function $\phi(z) = \frac{1}{1+e^{-z}}$



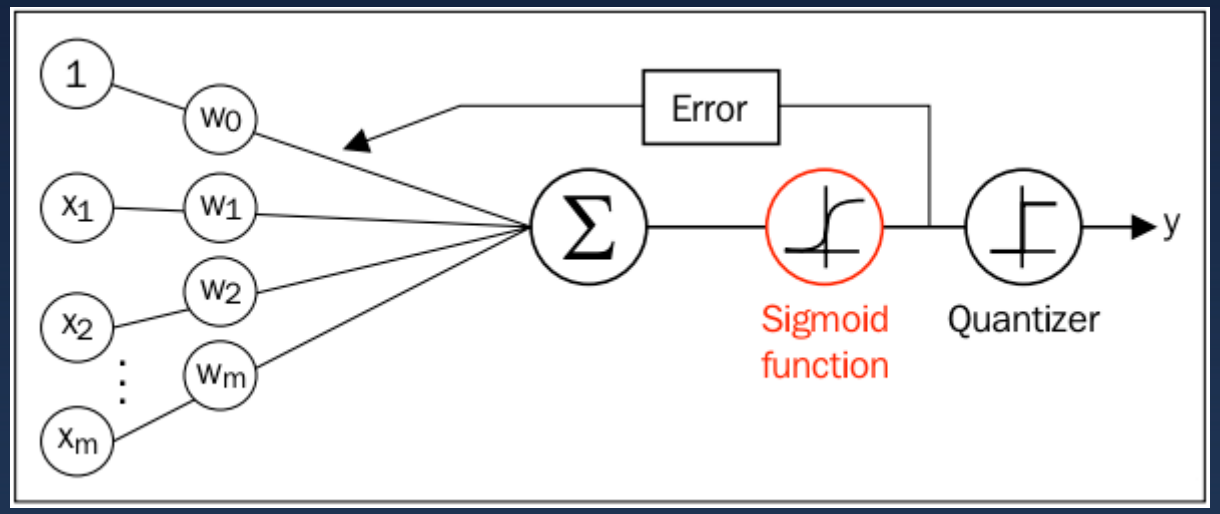

1、通过Sigmoid function，将样本代入其中进行计算。

2、将Sigmoid function的计算结果看做是这个样本属于类别1的概率大小。

3、将Sigmoid function的计算结果大于等于0.5的归为类别1，小于0.5的归为类别0.

# 逻辑回归计算

机器学习中样本的特征是很多的，特征数n通常是一个比较大的数字，LR计算的本质是在n+1维空间中使用Sigmoid function去对训练样本集进行拟合，在拟合的过程中对于每个特征赋予不同的权重，这个过程如图所示。




假定将拟合（训练）过程中的代价函数设置为真实值与预测值之间的误差：

$$J(w) = \sum_i \frac{1}{2}(\phi(z^{(i)}) - y^{(i)})^2$$

将Sigmoid function $\phi(z^{(i)}) = \frac{1}{1+e^{-z^{(i)}}}$ 代入计算，会得到一个非凸函数，存在着许多的局部最小值，无法求解。

需要为LR进行代价函数的设定！

# 逻辑回归代价函数计算

1、由于 Sigmoid function的计算结果可以看做是样本属于类别1的概率大小，则有：

$$p(y=1|x;w)=\phi(w^Tx+b)=\phi(z) \qquad p(y=0|x;w)=1-\phi(z)$$

2、上面两式转换为一般形式，则有：

$$p(y|x;w)=\phi(z)^y(1-\phi(z))^{(1-y)}$$

3、采用对$\omega$的极大似然估计，则有：

$$L(w)=\prod_{i=1}^{n}p(y^{(i)}|x^{(i)};w)=\prod_{i=1}^{n}(\phi(z^{(i)}))^{y^{(i)}}(1-\phi(z^{(i)}))^{1-y^{(i)}}$$

4、上式两边取自然对数，则有：

$$l(w)=lnL(w)=\sum_{i=1}^{n}y^{(i)}ln(\phi(z^{(i)}))+(1-y^{(i)})ln(1-\phi(z^{(i)}))$$

5、上式是求使得$l(\omega)$最大的$\omega$，而$l(\omega)$表示的预测准确需要付出的代价，则$LR$的代价函数可以定义为如下形式：

$$J(w)=-l(w)=-\sum_{i=1}^{n}y^{(i)}ln(\phi(z^{(i)}))+(1-y^{(i)})ln(1-\phi(z^{(i)}))$$

6、上式的意义在于求解使得预测准确代价最小的$\omega$。使用随机梯度下降的方法进行求解（$\eta$是学习率）：

$$w_j:=w_j+\eta(y^{(i)}-\phi(z^{(i)}))x_j^{(i)},for\ i\ in\ range(n)$$

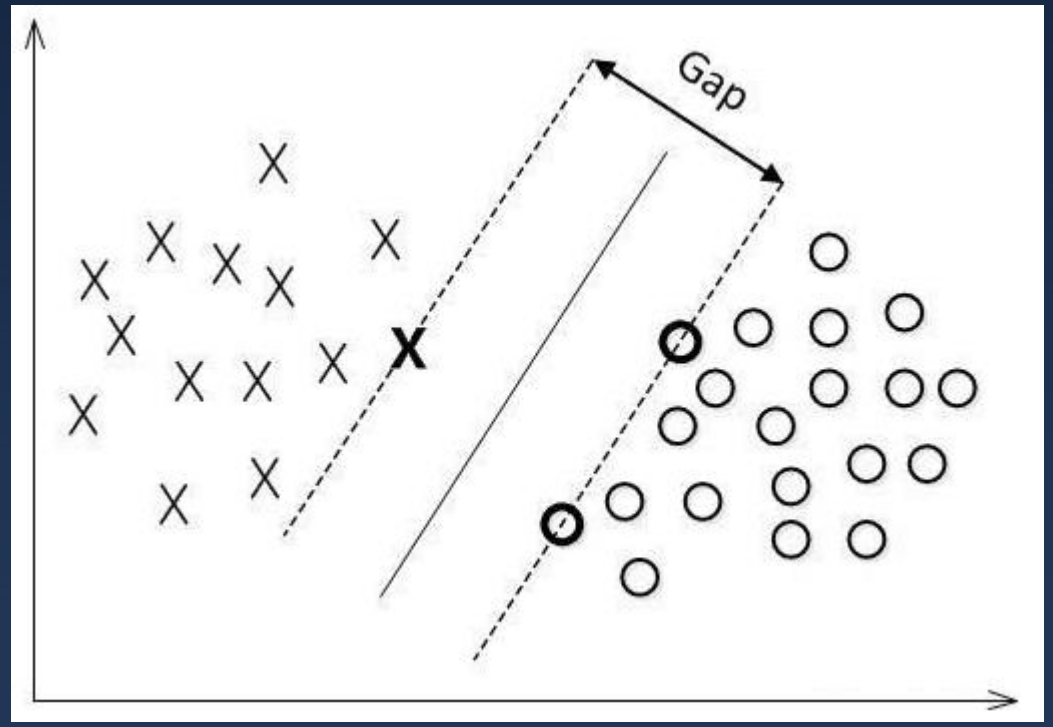
# 目录

Contents

# 从最简单的问题开始

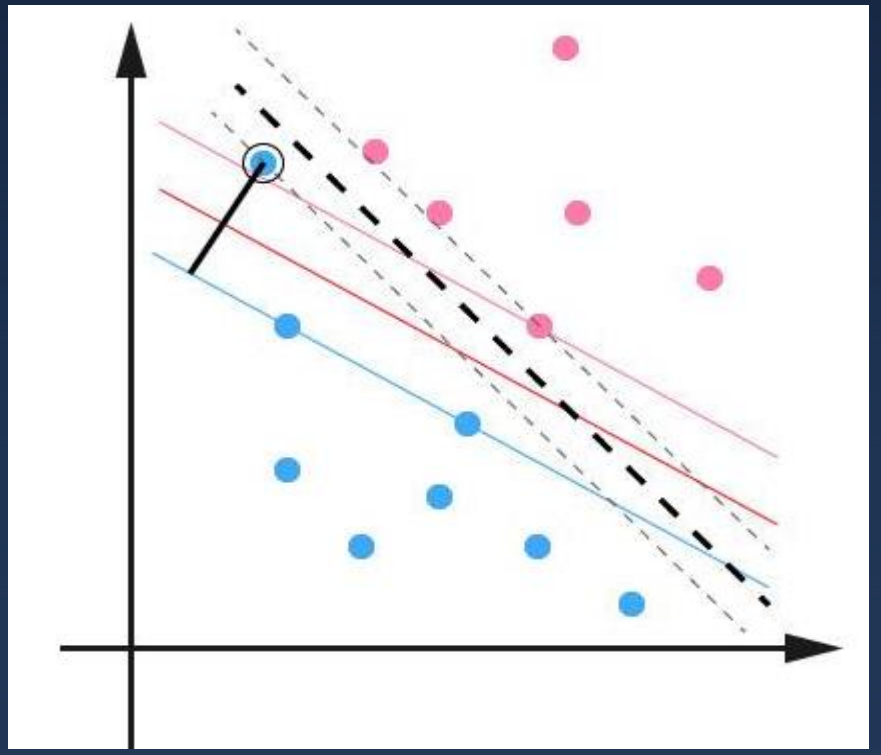
考虑一个最简单二分类问题，样本的特征只有两个连续值型维度，那么这个样本就可以看做二维平面上的一个点，假设训练样本集中样本是比较均衡的，且不同类别之间区别非常大，那么所有的样本点呈现在平面上时，可以清晰的看到明显的分隔。

图1




从图1上可以看出，可以使用一条直线将两个类别的样本完全隔开，并且它和两条虚线（类别的边界）的垂直距离相等。

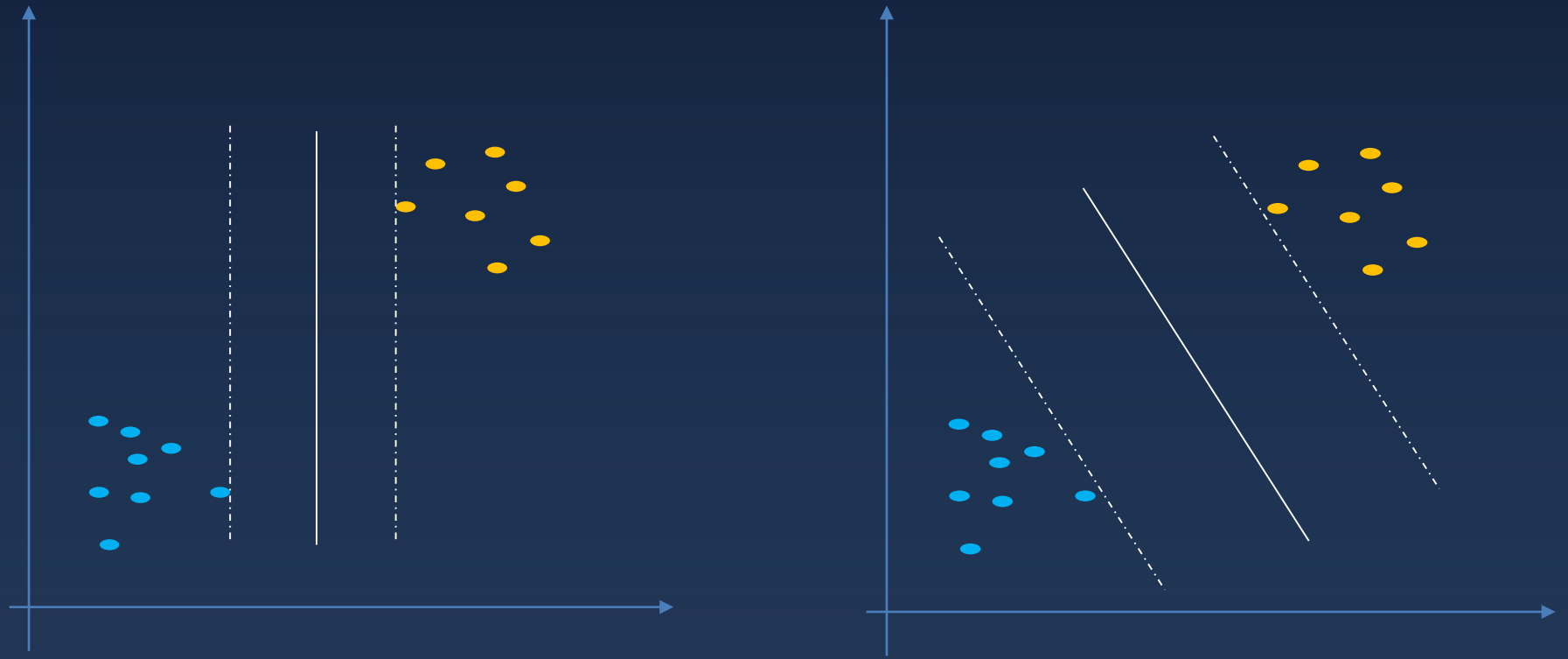
图2



如图2所示的样本被称为线性可分的样本，这样的分隔线理论上存在着无限多个（图2）

# SVM的构想

能够进行样本分隔的直线只和位于类别边界上的样本相关，SVM希冀通过一些方法找到一条最佳的直线。当把数据推广到超过3维的空间，SVM希冀寻找的就是最佳超平面，这个平面被称为MMH（Maximum Marginal Hyperplane，最大边缘超平面）



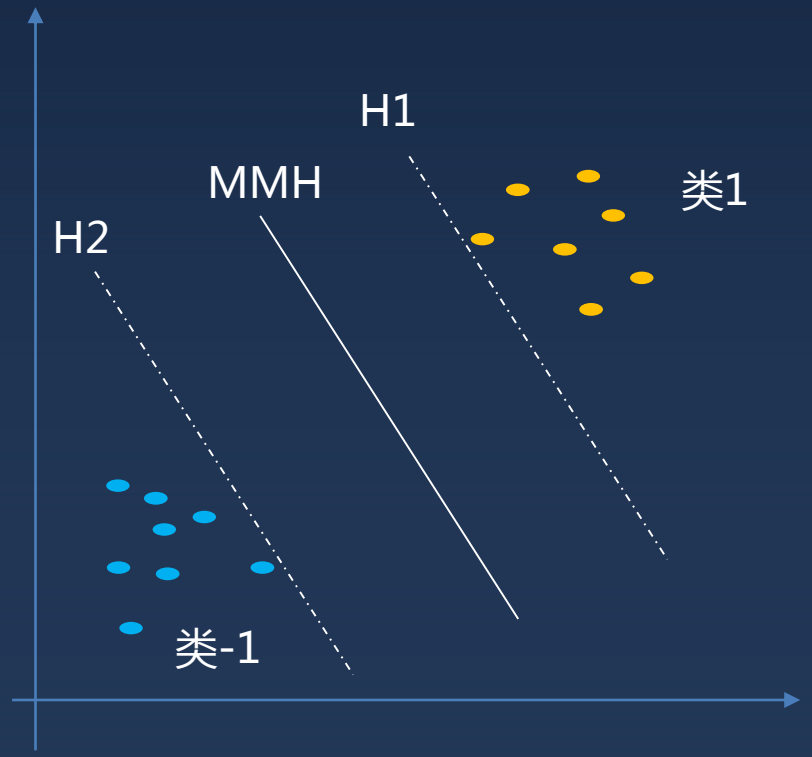
上面两图的实直线都正确的对样本进行了分类，不同之处在于左图的直线和虚线之间的距离小，SVM认为这个距离越大，说明实直线在未来可以分隔的样本越多，所以需要寻找的是MMH。

# SVM的计算

还是依然从最简单的问题进行考虑，样本的特征只有两个连续值型维度< $x_1$ , $x_2$>。

1、设分隔直线为$w_0 + w_1x_1 + w_2x_2 = 0$，则一类样本满足的条件为$w_0 + w_1x_1 + w_2x_2 > 0$,另外一类满足$w_0 + w_1x_1 + w_2x_2 < 0$

2、设类别标签为<1,-1>,赋予类别标签数值意义，设W=< $w_0$ , $w_1$ , $w_2$>，通过调整W，可以很容易的使得虚线上和虚线外的样本满足H1: $w_0 + w_1x_1 + w_2x_2 \geq 1$和H2: $w_0 + w_1x_1 + w_2x_2 \leq -1$，融合两式，可得： $y_i(w_0 + w_1x_1 + w_2x_2) \geq 1$




3、由上可得，MMH到H1的距离为$\frac{1}{||W||}$， MMH到H2的距离也同样，其中$||W|| = \sqrt{W \cdot W}$。所以SVM的目标是最大化$\frac{1}{||W||}$，即最小化$\frac{1}{2}||W||^2$

4、将目标转换为KKT条件求解：

$$\max_\alpha \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$s.t., \alpha_i \geq 0, i = 1, \cdots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

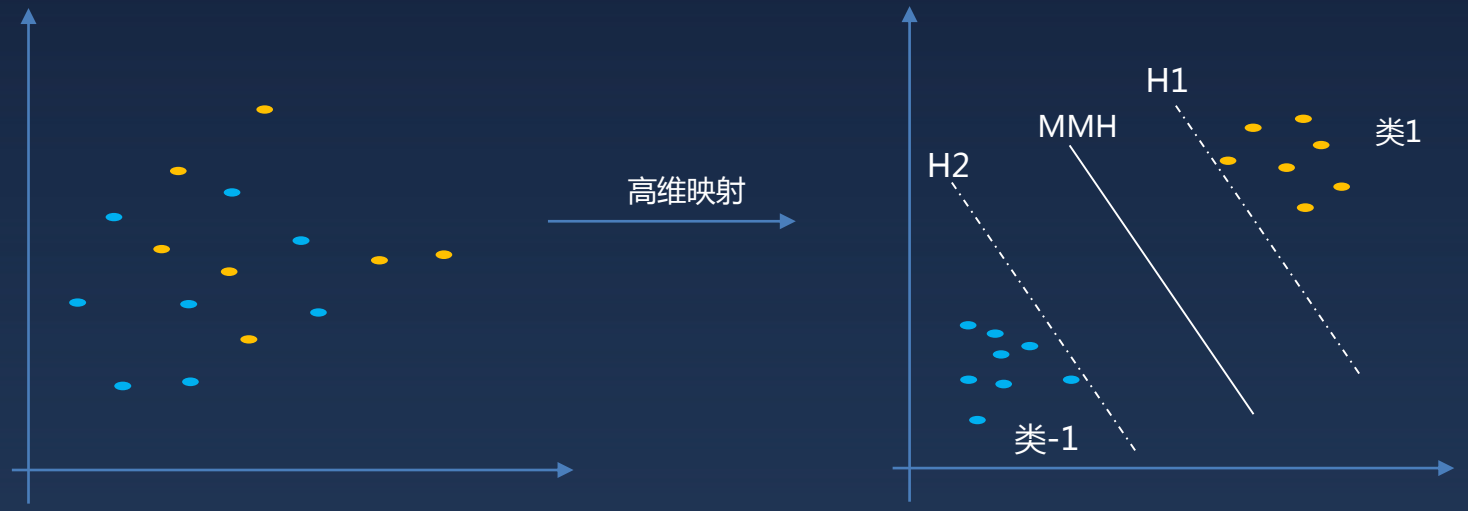
确定H1和H2的样本被称为支持向量

# 线性不可分时SVM的计算

线性不可分才是数据的常态，对于线性不可分的数据，SVM可以进行以下操作。

1、使用非线性的映射将原始的输入数据变换到更加高维的空间。

2、在新的高维空间进行MMH的求解。




在求解得到支持向量和相应的参数$\alpha_i$之后，假定有一个样本A需要进行进行预测，则使用$f(A) = \sum_{i=1}^{l} y_i \alpha_i X_i A^T$进行预测，计算结果为正，则属于类1，为负属于类-1。其中$y_i$和$X_i$就是支持向量的类别和支持向量。

在SVM求解过程中会涉及大量的向量点积运算，高维空间中进行求解会使得运算开销非常大，所以将点积运算用核函数代替，常用的核函数有：h次多项式核函数、高斯径向基函数核函数、Sigmoid型核函数、

# Thank You.



华为云机器学习服务MLS
**www.huaweicloud.com/product/mls.html**