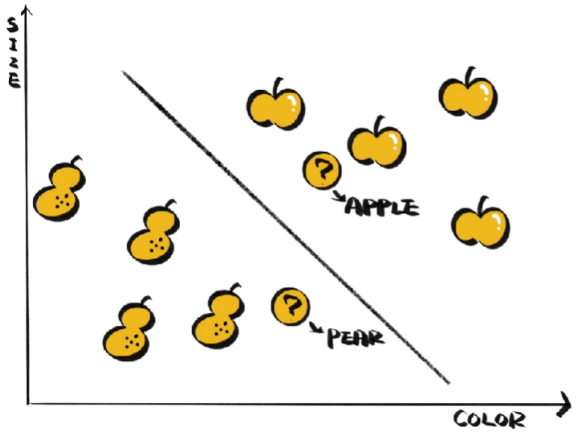
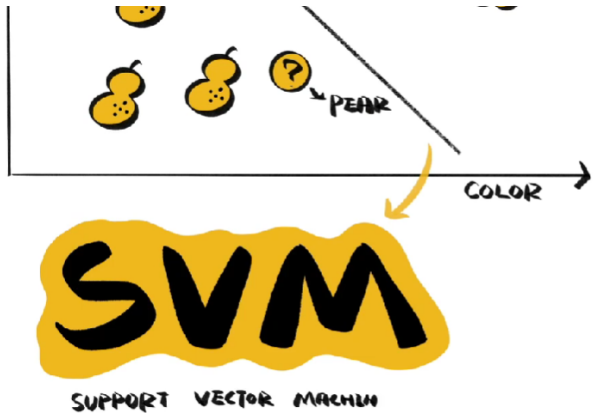
什么是SVM(支持向量机)

想要知道新拿到的水果是梨还是苹果，除了用KNN画个圈，还有什么好办法?

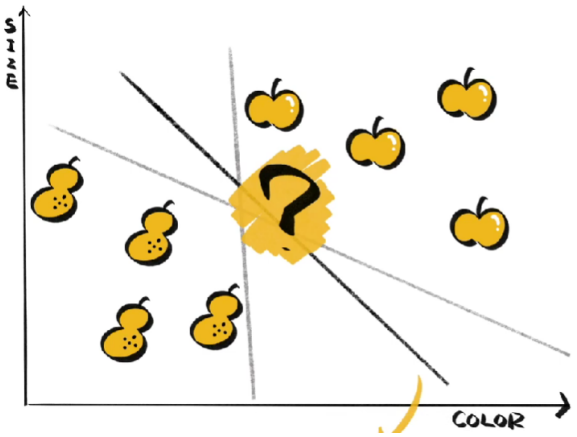
画条线好像也不错，通过将两者所在的空间做出区分。当新样本落在苹果一侧时，我们就认为它是苹果，反之就认为它是梨。



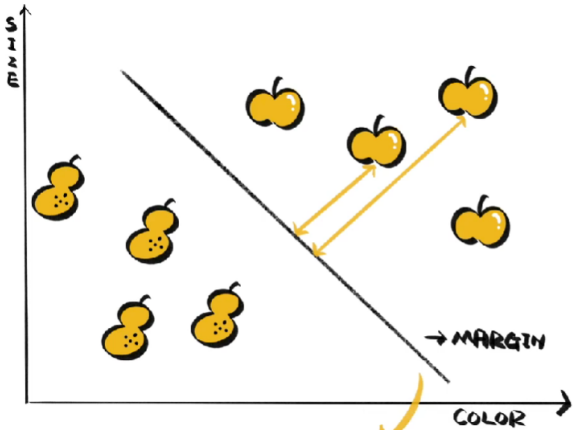
这条线就是**SVM-支持向量机**。



不过这条线有多种画法，但哪条线才是最合适的?



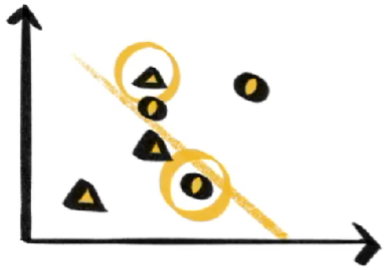
除了界限，样本与线的距离同样有意义，它代表样本分类的可信程度。



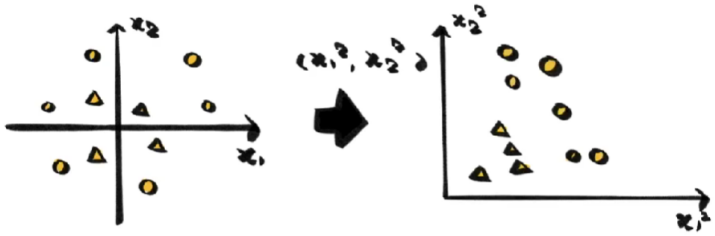
以苹果这一侧为例，与线的距离最远的苹果，是苹果的可能性最高，离得越近，是苹果的可能性越低。我们的目标是在两种样本间，找到能让所有样本的分类可信度最高的那条线。

不必计算所有的距离，只要找到线附近的样本，让它们与线的距离越远越好，这个距离被称为分类间隔，决定了线的样本被称为支持向量，这也是支持向量机名字的由来。

如果样本的分布有交叉怎么办？那么就关注这些无法被线正确分类的样本与线之间的距离，找到能最小化这个距离的线。



如果样本的分布并不理想，无法用直线区分怎么办？那就通过一定变换，将它们映射到一个能用直线区分的空间，再寻找分类线。



在深度学习出现前，随机森林和SVM是最好用的分类方法，SVM对样本依赖小，不会过拟合，小样本也能取得不错的效果。文本分类、垃圾邮件识别、图像分类、甚至分类蛋白质，SVM应用广泛，至今仍未褪色。