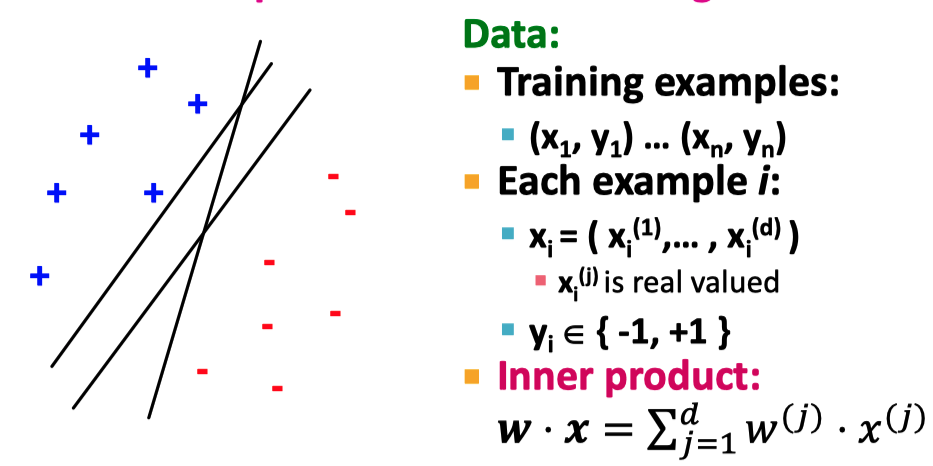
2 支持向量机介绍

2.1二分类问题和二分类模型

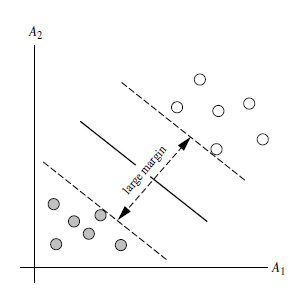
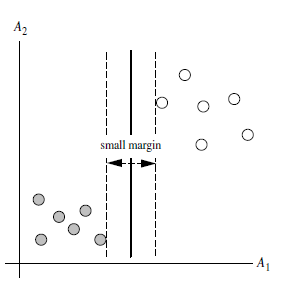
支持向量机（SVM）是一种二分类模型。二分类模型要解决的问题是：已知一组数据以及每条数据所属的类别（一共有两类），对于一个新的数据，需要判断它属于哪个类别。例如，假设已知人的情绪与特定位置的脑电信号有关，简单将人的情绪分为积极/消极两类，分别测量人在积极和消极情绪下的脑电信号，就得到了一个训练集。二分类模型需要完成的，是依据训练集，对于一个新的脑电信号，判断它属于积极还是消极的一类。这样，就可以通过脑电信号识别人的情绪了。

2.2线性支持向量机的基本思想

数据在经过预处理后得到若干特征值，可以表示为在多维空间中的一个点。如果能够找到一个超平面将点分成两部分，两类点分别位于平面两侧，那么对于新的一个点，只需判断它在平面的哪一侧，就可以判断出它属于哪一类了。例如在下图中，可以找到直线将蓝色“+”点与红色“-”点分割开来，对于新的黑色圆点，可以判断出它位于分割线右下方的位置，即属于红色“-”类。如果能够找到用于分割的超平面（二维空间中为直线），就称数据集线性可分，用于线性可分数据集的支持向量机称为线性支持向量机。



对于线性可分数据集，分割面是不唯一的。分割面的选取会影响分类效果。例如，如果像下面左图选择分割线，对于新的蓝色点，它处于分割线（实线）左侧，被模型归为灰色一类。但实际上，它距离白色点更近，这就产生了误差。解决的方法是引入所有的点到直线距离中的最小值，称为间隔（margin），最好的分割线使间隔最大，即在两类点之间有尽可能大的隔离带，如下面右图所示。在上下边界线上的点称为支持向量，它们确定了平面的位置。训练好的模型仅保存超平面的参数和支持向量。



2.3线性支持向量机的数学表示和推导

首先，超平面使用表示。向量和常数b为需要求的参数，向量为点在空间中的坐标。坐标为向量的点到平面距离为

平移超平面，直到遇到两侧的点。记此时的超平面为 ，代表“隔离带”的上下边界。在上侧边界的点满足，在下侧边界的点满足 。为了统一表述，对每个点定义y值，上侧的点的y值为1，下侧为-1，则边界上的点满足

由解析几何和线性代数知识，边界上的点到中间的分割线的距离为

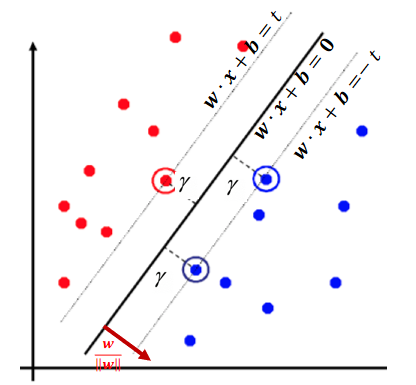
优化目标为

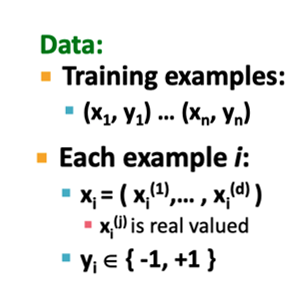
这组公式的意义为在保证所有点位于隔离带两侧的前提下，最大化间距。

如果将公式两侧同时除以t，用新的b，w代替和，就可以得到

如此在不改变公式意义的前提下，减少了未知量的个数。

最后，为了便于求导，最小化代替最大化，得到线性支持向量机的标准形式



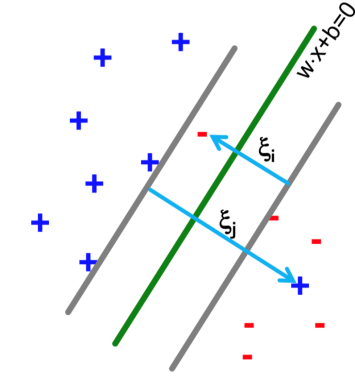


2.4非线性支持向量机

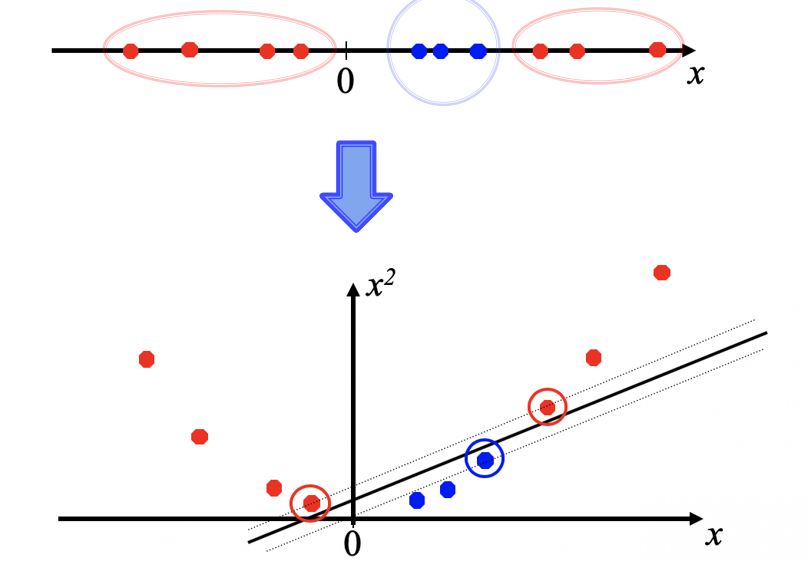
实际上，相当多的数据集都不满足线性可分的性质。对于近似可分的数据集，即如果排除少部分点就满足线性可分性质的数据集，可以允许部分点位于错误的位置，包括隔离带内和对侧区域。对于每个未处于正确位置的点赋予一个惩罚项，大小为点到其所属的一侧的边界线的距离，即

并将该惩罚项加入最终优化目标中，得到

如此，模型会在令间距尽可能大的同时令误差尽可能小。



对于线性不可分的数据集，一般利用核函数将其投射到高维空间中。在高维空间中，数据集可能展现出线性可分或近似线性可分的性质。例如下图中的点在一维空间中不满足线性可分，但投射到二维空间中就可以用一条直线将其分割开。



2.5支持向量机的应用实例

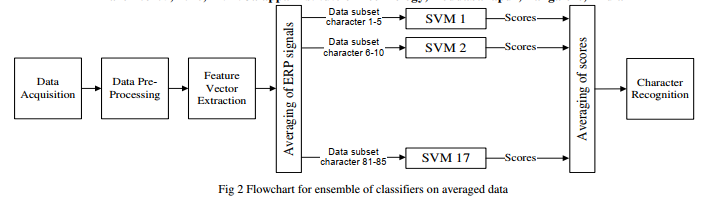
这里介绍基于P300 ERPs的“意念输入”。P300是事件相关电位的一种成分，是指给予神经系统特定的刺激，在该系统和脑的相应部位产生的可以检出的、与刺激有相对固定时间间隔和特定位相的生物电反应。该研究使用了P300 speller的范式。P300 speller是一个数字和字母组成的6x6矩阵。矩阵的每一行、列会随机高亮，一次实验中6行6列均被高亮一次，一共12次高亮刺激。受试者必须将注意力集中在矩阵中的字符上，以此来选择组成单词的每个字母。如果受试者想输入的字符被高亮，会产生P300。因此，可以通过检测P300的存在与否来确定受试者想输入的字符，实现“意念输入”：如果在某一行或列高亮的时候出现了P300，就说明该行或列含有受试者想输入的字符。每行每列均高亮一次后，即可定位出想要输入的字符。[1][2]



该研究利用支持向量机来检测P300的存在与否。首先，取得传感器收集到的数据，进行预处理、特征提取、取平均等操作，形成训练集。然后，将数据集分成多个子集，训练多个SVM分类器，完成模型搭建。

对于要分类的数据，同时使用多个支持向量机进行分类，综合多个分类器的结果，最后判断其是否含有P300。

该研究使用多个支持向量机的原因是，对于大数据集，支持向量机的训练时间长，同时使用多个小数据集训练可以提高速度。同时也可以提高准确性。[3]



2.5支持向量机的优缺点

优点：支持向量机由严格的数学推导得来，具有可解释性；它对于数据的特征没有要求，可以处理线性不可分的数据。

缺点：最大的缺点是运行速度慢，算法复杂度高，为。其次在升维处理线性不可分的数据时，采用哪种核函数需要人工猜测，可能难以达到好的效果。该模型是比较早期的模型，在神经网络出现后使用率就很低了。

参考文献

[1]S. Yan, H. Zhao, C. Liu and H. Wang, "Brain-computer interface design based on wavelet packet transform and SVM," 2012 International Conference on Systems and Informatics (ICSAI2012), 2012, pp. 1054-1056, doi: 10.1109/ICSAI.2012.6223215.

[2]A. Rakotomamonjy and V. Guigue, "BCI Competition III: Dataset II- Ensemble of SVMs for BCI P300 Speller," in IEEE Transactions on Biomedical Engineering, vol. 55, no. 3, pp. 1147-1154, March 2008, doi: 10.1109/TBME.2008.915728.

[3]V. Bhatnagar, N. Yede, R. S. Keram and R. K. Chaurasiya, "A modified approach to ensemble of SVM for P300 based brain computer interface," 2016 International Conference on Advances in Human Machine Interaction (HMI), 2016, pp. 1-5, doi: 10.1109/HMI.2016.7449163.