

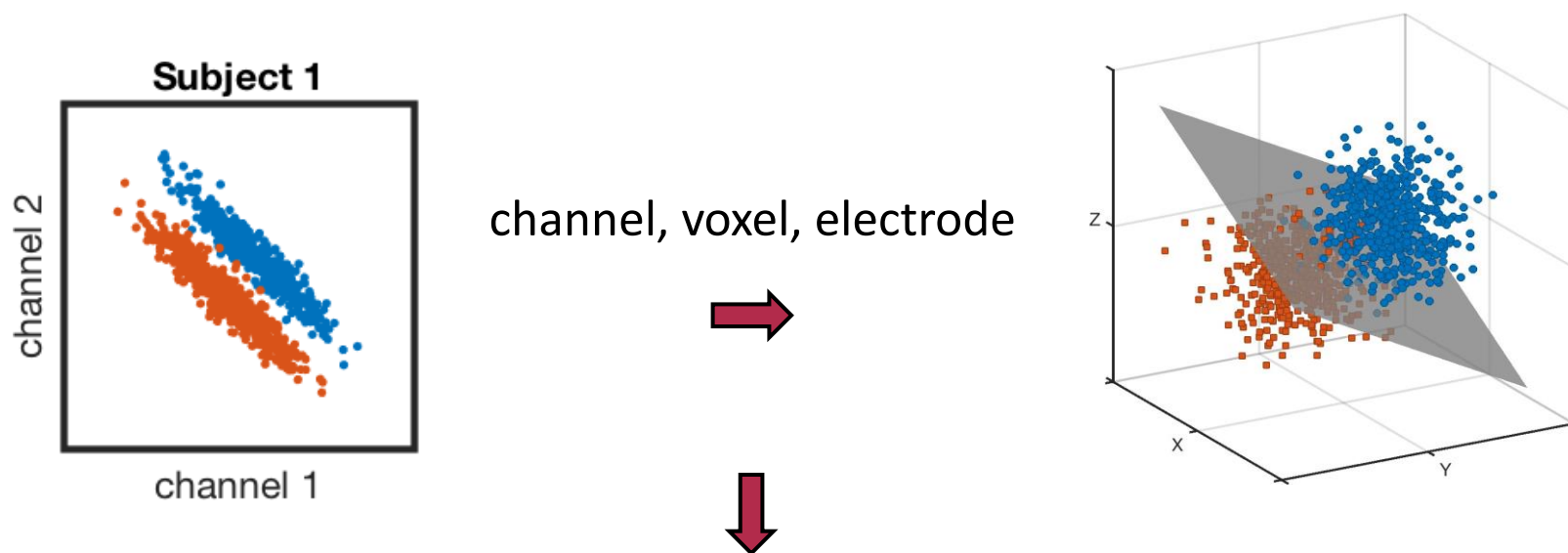
RacLab

Feature Extraction & Classifier

Yang Ziyang

2024.08.23

Feature Extraction



从一个 n 样本 \times m 维度的数据中 找到一个 $m-1$ 维度的超平面区分两组数据

$$\begin{matrix} A1 \\ A2 \\ A3 \\ A4 \\ \vdots \\ An \end{matrix} \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \dots & x_{1m} \\ x_{21} & x_{22} & x_{23} & x_{24} & \dots & x_{2m} \\ x_{31} & x_{32} & x_{33} & x_{34} & \dots & x_{3m} \\ x_{41} & x_{42} & x_{43} & x_{44} & \dots & x_{4m} \\ \vdots & & & & & \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \dots & x_{nm} \end{bmatrix}$$

$$W_1X_1+W_2X_2+\dots+W_mX_m+B=0$$

权重

维度数

决策超平面

Feature Extraction

n样本 x m维度

标签 Label

$$\begin{matrix} A1 \\ A2 \\ A3 \\ A4 \\ \vdots \\ An \end{matrix} \begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} & \dots & x_{1m} \\ x_{21} & x_{22} & x_{23} & x_{24} & \dots & x_{2m} \\ x_{31} & x_{32} & x_{33} & x_{34} & \dots & x_{3m} \\ x_{41} & x_{42} & x_{43} & x_{44} & \dots & x_{4m} \\ \vdots & & & & & \\ x_{n1} & x_{n2} & x_{n3} & x_{n4} & \dots & x_{nm} \end{pmatrix}$$

性别

是否患病

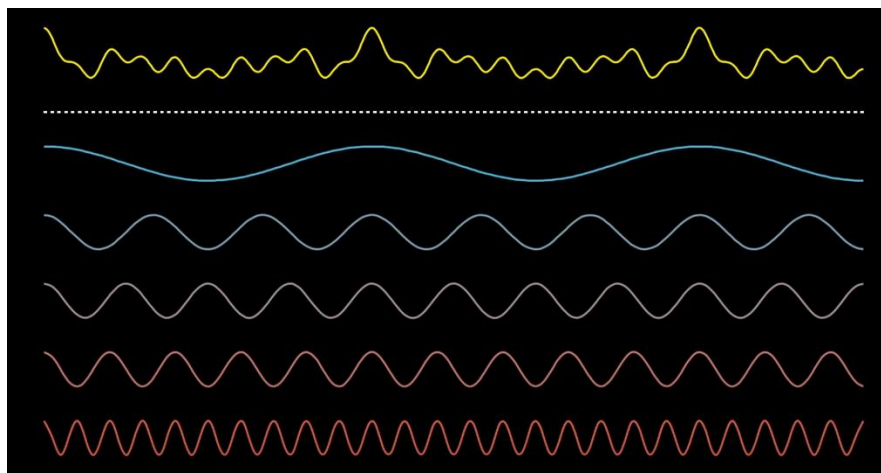
实验操纵

- 功率谱特征：脑电各频段的绝对/相对频谱能量
- 脑网络特征：通道（或脑区）之间的功能连通性，以及更高阶的图论指标。
- ERP特征：ERP幅值。

Feature Extraction

- 功率谱特征

- 功率谱：也称功率谱密度（Power Spectral Density, PSD），单位是功率/Hz。表现的是单位频带内信号功率随频率的变换情况
- 常用方法有FFT法，Welch法



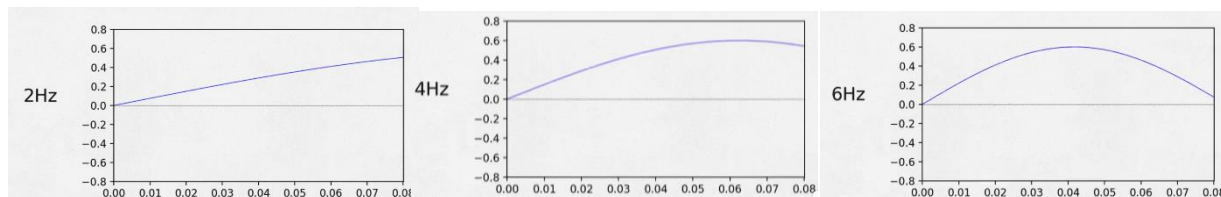
Feature Extraction

FFT(快速傅里叶变换)

- 功率谱特征

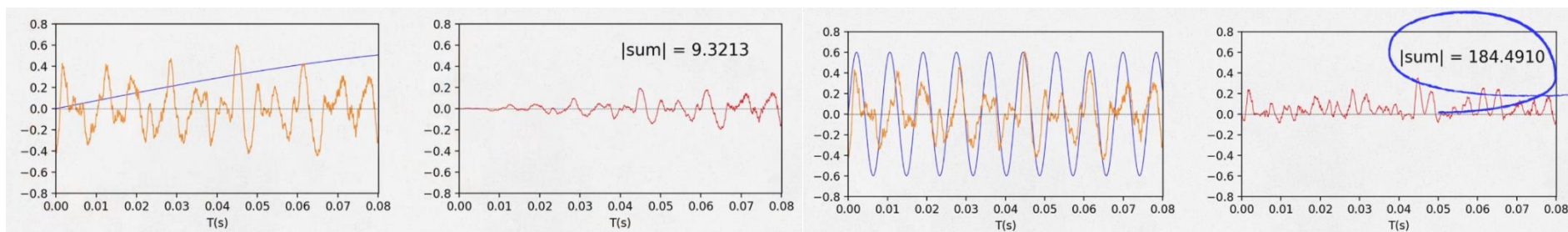
每一段脑电信号都是由约100个不同频率的**正弦波**叠加

傅里叶变换准备好了所有不同频率的正弦波



三角函数的正交不变性

将数据中每一个采样点(采样率决定)与之**相乘**



看看同向振动的部分多不多(**相乘再相加**, 同向正值异向负值, 再**取和的绝对值**)

若|SUM|值大, 则代表该频率在脑电信号中的构成越重要

FFT(X,[],DIM) or FFT(X,N,DIM)

Feature Extraction

- 功率谱特征

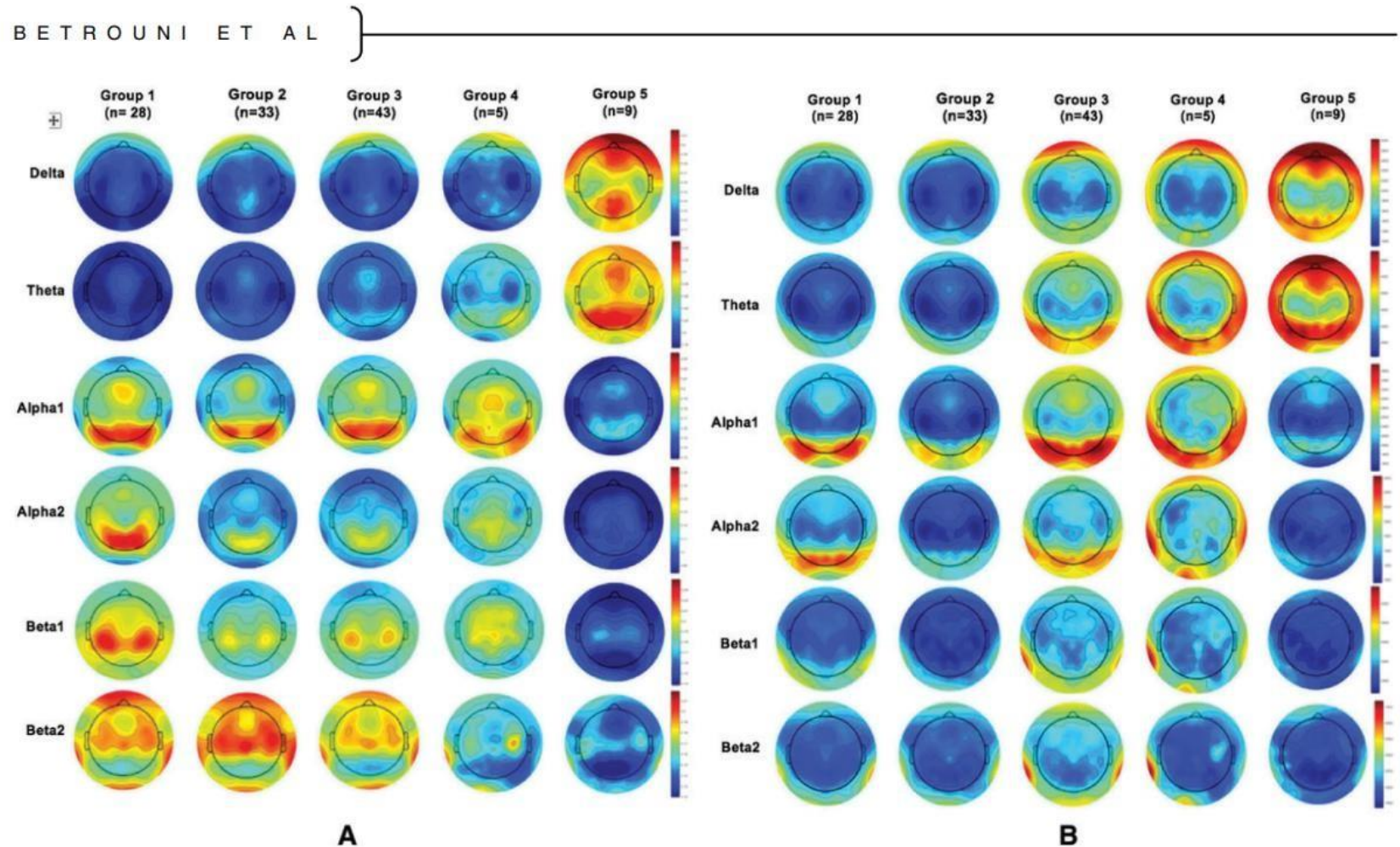


FIG. 1. Head plots of the distribution of mean power per frequency band. The warmer colors indicate higher relative power (scaled from minimum to maximum values of the total group): (A) relative powers; (B) absolute powers.

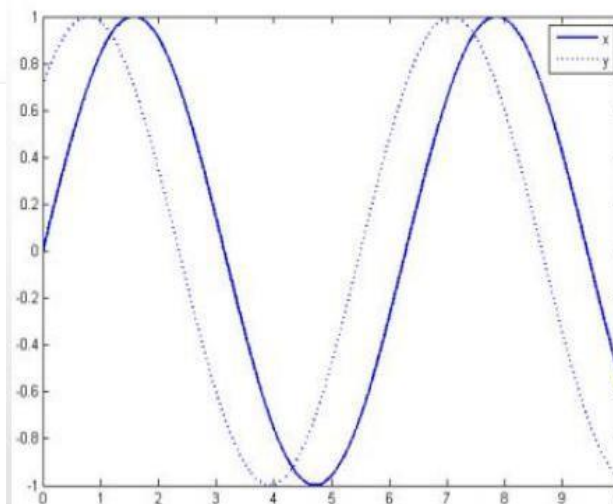
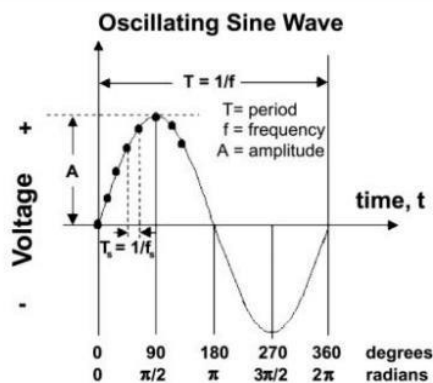
Feature Extraction

- 功能连接特征

- 二、基于相位同步的指标
- 什么是相位同步

相位同步 (phase synchronization , PS) 指的是, 两个相互耦合的神经振荡活动的相位 (phase) 同步化 (即两个活动的相位差不随着时间的变化而变化, 有一个固定的相位差)

PS的优点是 (理论上) 与两个神经振荡活动的波幅无关, 而只与相位有关

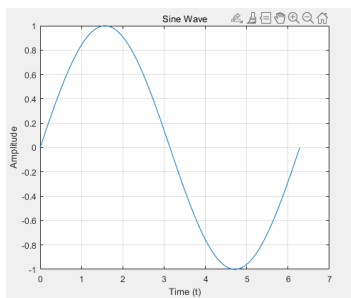


信号x和信号y之间的相位差恒定为 $\pi/4$, 即45度

Hilbert 变换

Feature Extraction

- 熵 - 熵这个概念最早是用于热力学中，它的物理意义是体系中混乱程度或者复杂程度的度量。关于熵的应用也在不断拓展，从热力学到生物学、物理学，以及在时间序列分析上都有应用。



$$r = 0.2 \text{std}(\text{signal})$$

对时间序列 $\{x_1, x_2, \dots, x_N\}$ 选择嵌入维数 m 和相似性容忍度 r 。

构造长度为 m 的向量序列 $X(i) = [x_i, x_{i+1}, \dots, x_{i+m-1}]$ ，其中 $i = 1, 2, \dots, N-m+1$

从 i 点开始的 m 个连续的 x 值

$$X(1) = [a1, a2]$$

$$X(2) = [a2, a3]$$

切比雪夫距离

$$\text{MAX}(d1, d2) = D \text{ 与 } r \text{ 作比较}$$

然后，我们将嵌入维数增加到 $m+1$ ，并计算 $m+1$ 维向量的匹配情况

所有序列下小于 r 的 D 值总数为 B

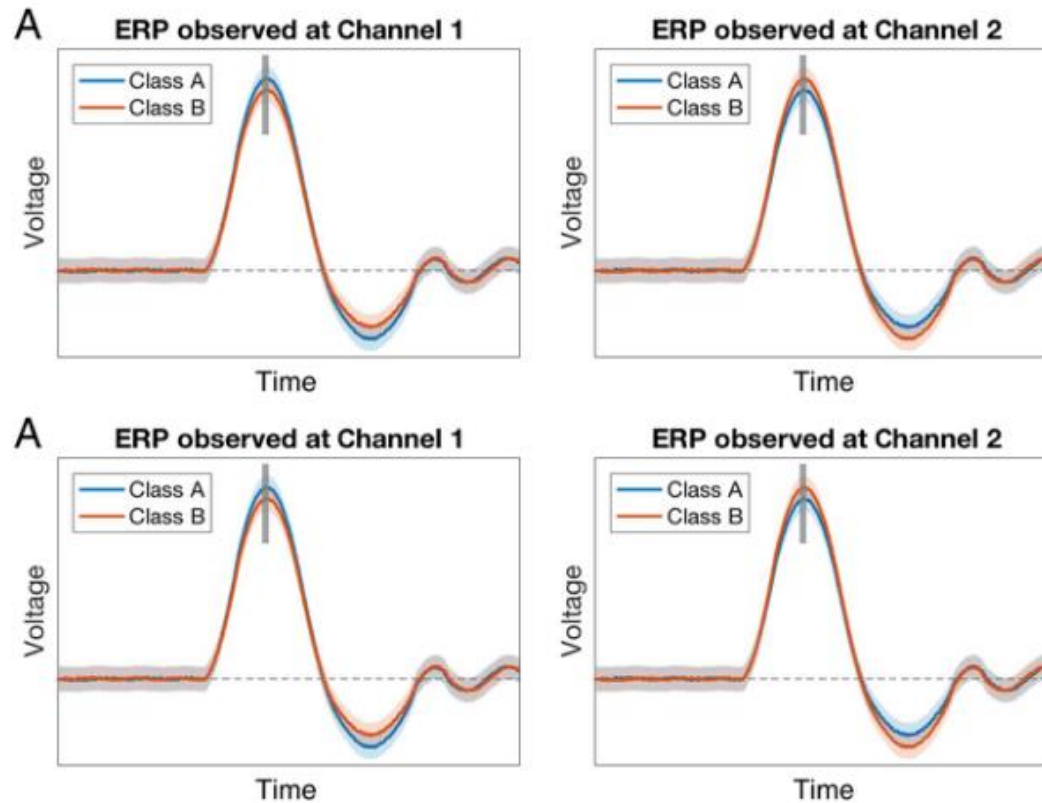
$$\text{SampEn}(m, r, N) = -\log \left(\frac{B^{m+1}(r)}{B^m(r)} \right)$$

序列下所有小于 r 的 D 值总数为 A

如果时间序列是完全规则的，那么无论维数是 m 还是 $m+1$ 匹配的概率都不会发生显著变化

Feature Extraction


- ERP特征：ERP幅值



Feature Extraction

 **erp_all**

80x20400 double

 **all_mean_amplitude**

80x6 double

n样本 x m维度

标签 Label

Feature Extraction

- 特征选择方法

过滤法:

- | | | | |
|----|---|---------|--------------------------|
| 回归 | { | • 皮尔逊相关 | Feature 和 Label |
| 分类 | | • T检验 | Feature 和 Feature 做t检验 |
| | { | • F-分数 | 正类和负类的差异/(正类组内方差+负类组内方差) |

Lasso回归(最小绝对收缩和选择算子\L1正则)

主成分分析(PCA)

$$\min_{w, b, \xi} \quad \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i$$

$$\text{s.t.} \quad y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1, 2, \dots, N$$

$$\xi_i \geq 0, \quad i = 1, 2, \dots, N$$

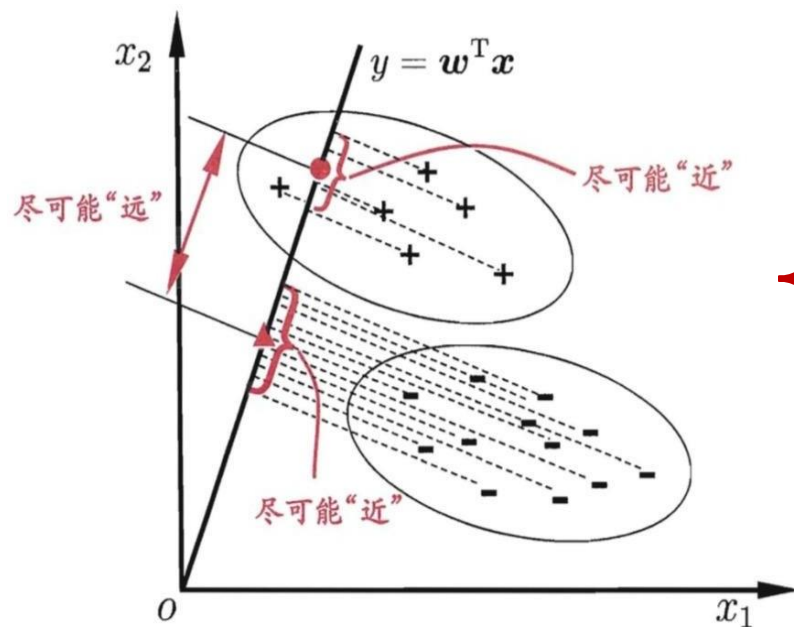
Classifier

- 模型的选择。机器学习的模型很多，包括：感知机、k近邻、贝叶斯、决策树、逻辑斯蒂回归与最大熵模型、支持向量机、k-means聚类、隐马尔科夫、深度神经网络，以及其他如AdaBoost之类的提升方法。
- 同一类模型也可包含各种不同“复杂度”（或不同“参数”）的模型。如，多项式模型可分为1次、2次、3次、... n次多项式模型。k近邻算法有3-近邻、4-近邻、5-近邻算法等。
- “没有免费的午餐”定理：没有任何一种方法能在各种数据集里完胜其他所有的方法。
- 一般来说，简单的模型解释性强一些，复杂的模型解释性弱一些。

Classifier

线性判别分析 (LDA, Linear Discriminant Analysis)

- 模型原理
- 线性判别分析：将样本投射到“超平面”上，使得同类样本的投影点尽可能接近，异类样本的投影点尽可能远离。对新样本进行分类时，将新样本投射到“超平面”上，根据投影点的位置确定新样本的类别。

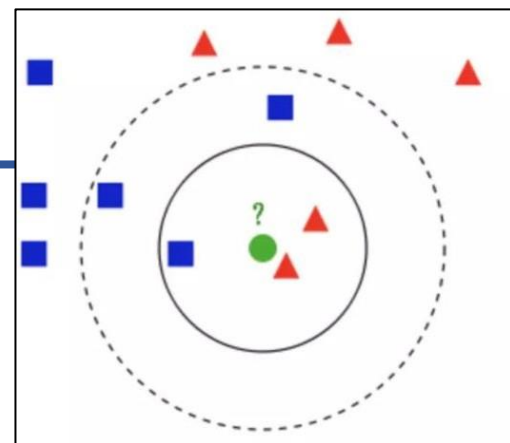


类内方差最小化

类间均值差异最大化

Classifier

K近邻 (kNN, k-Nearest Neighbor) :



模型原理:

- K近邻: 如果一个样本在特征空间中的k个最相邻的样本中的大多数属于某一个类别, 则该样本也属于这个类别, 并具有这个类别上样本的特性。该方法在确定分类决策上只依据最邻近的一个或者几个样本的类别来决定待分样本所属的类别。kNN方法在类别决策时, 只与极少量的相邻样本有关。由于kNN方法主要靠周围有限的邻近的样本, 而不是靠判别类域的方法来确定所属类别的, 因此对于类域的交叉或重叠较多的待分样本集来说, kNN方法较其他方法更为适合。
- kNN模型的三个基本要素: (1) 距离度量 (2) k值的选择 (3) 分类决策规则。

Classifier

K近邻 (kNN, k-Nearest Neighbor) :

- kNN模型的三个基本要素

(1) 距离度量

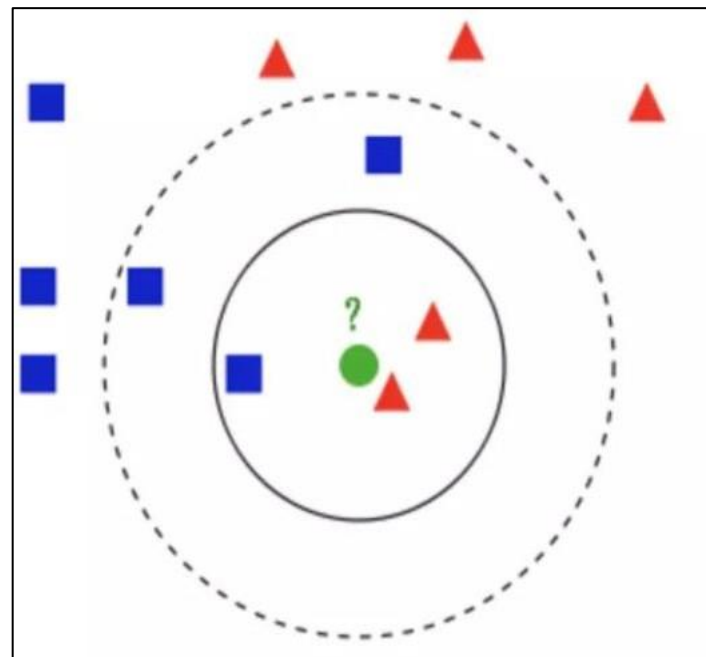
欧式距离:

$$D(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

曼哈顿距离:

$$D(x, y) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n| = \sum_{i=1}^n |x_i - y_i|$$

(2) k值的选择

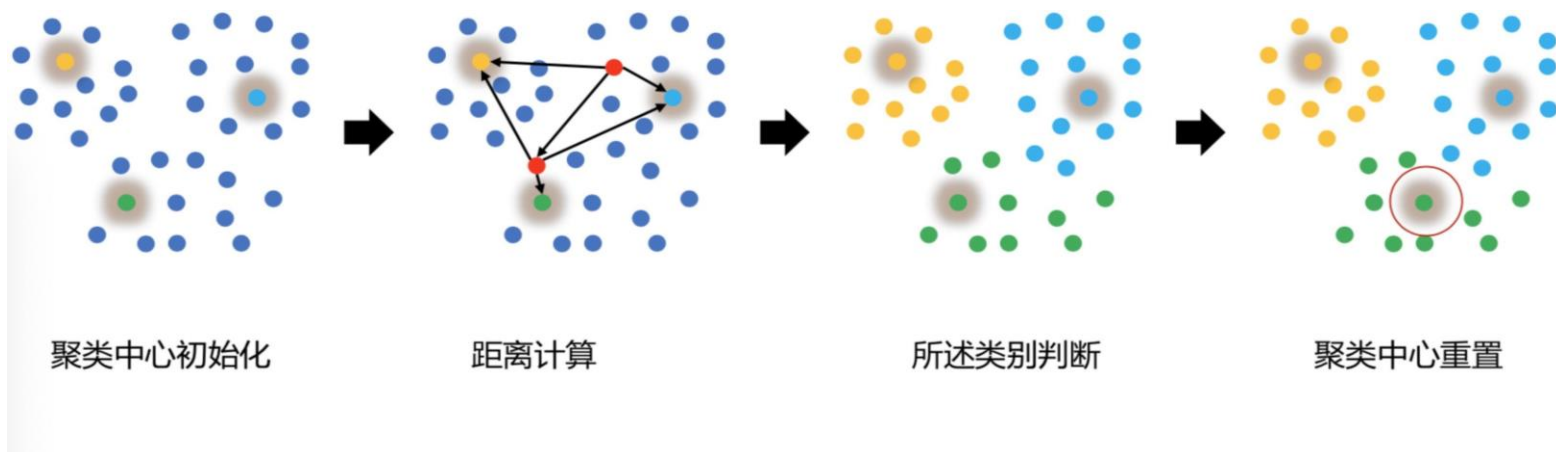


通常采用交叉验证法来选取最优的k值。

Classifier

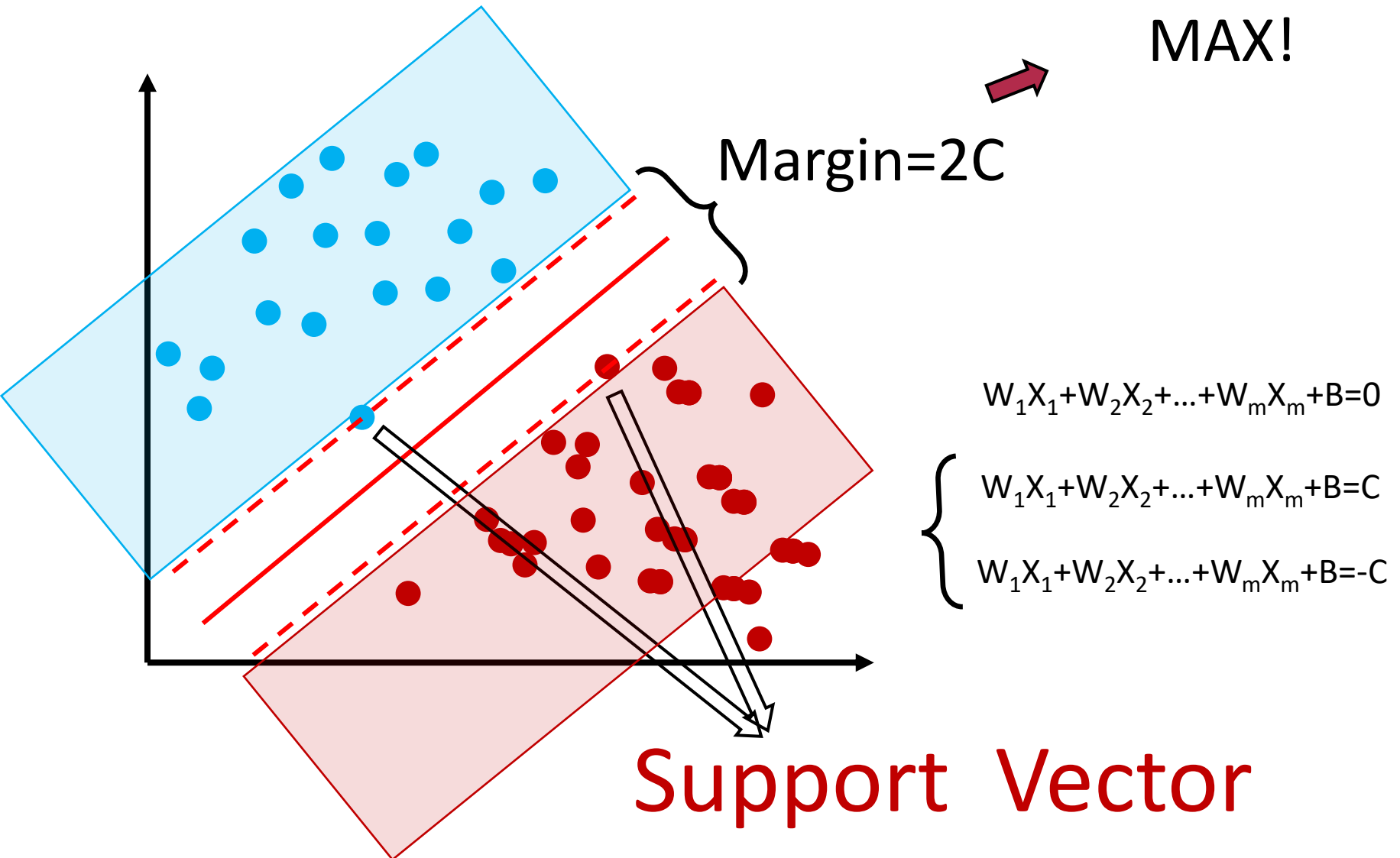
K-Means Cluster原理

- 首先, 选择 n 个数值型变量参与聚类分析, 最后要求的聚类数为 k 个;
- 其次, 由系统选择 k 个(聚类的类数) 观测量 (也可由用户指定) 作为聚类的种子;
- 第三, 按照距离这些类中心的距离最小的原则把所有观测量 (样品) 分派到各类重心所在的类中去;
- 第四, 这样每类中可能由若干个样品, 计算每个类中各个变量的均值, 以此作为第二次迭代的中心;
- 第五, 然后根据这个中心重复第三、第四步, 直到中心的迭代标准达到要求时, 聚类过程结束。



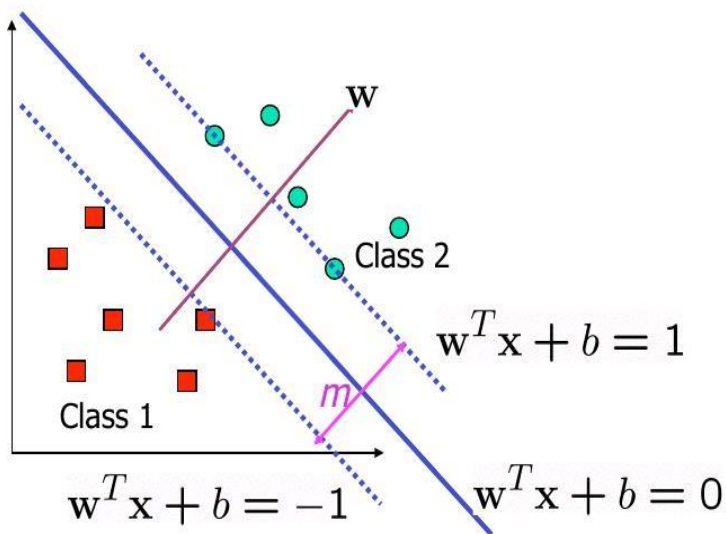
Classifier

Support Vector Machine (SVM)



Classifier

Support Vector Machine (SVM)



最大化间隔 $m = 2C = 2/\|w\|$

最小化： $\frac{1}{2} \|w\|^2$

约束： $y_i (w^T x_i + b) \geq 1 \quad \forall i$

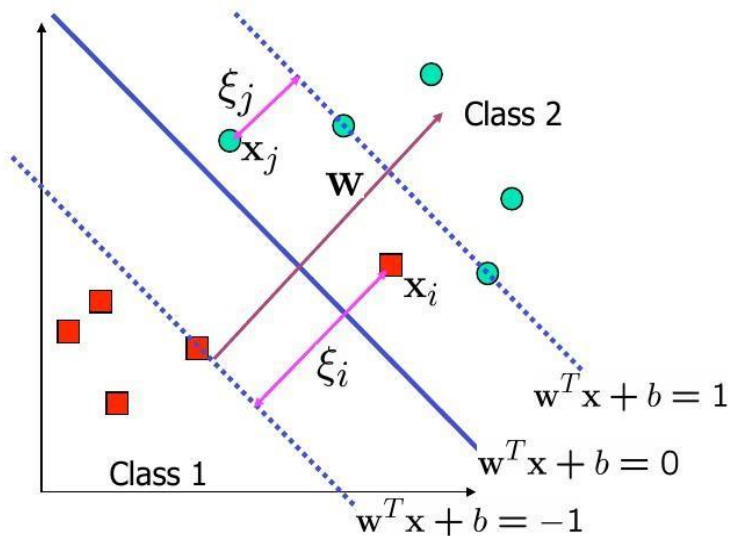
间隔带距离

间隔带中点距离和

$\frac{1}{2} \|w\|^2 + C \sum_{i=1}^m \xi_i$

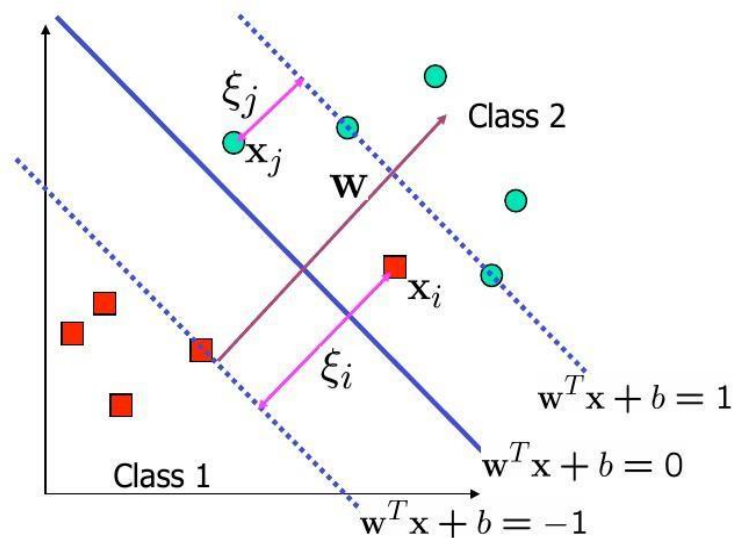
惩罚系数

$y_i (w^T x_i + b) \geq 1 - \xi_i$ 。注意： $\xi_i \geq 0$ 。



Classifier

Support Vector Machine (SVM)



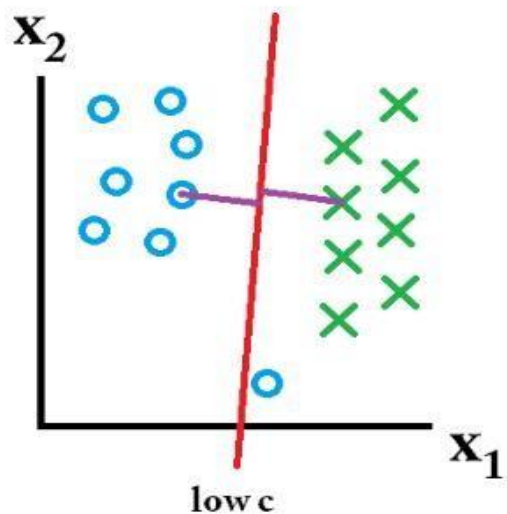
间隔带距离

间隔带中点距离和

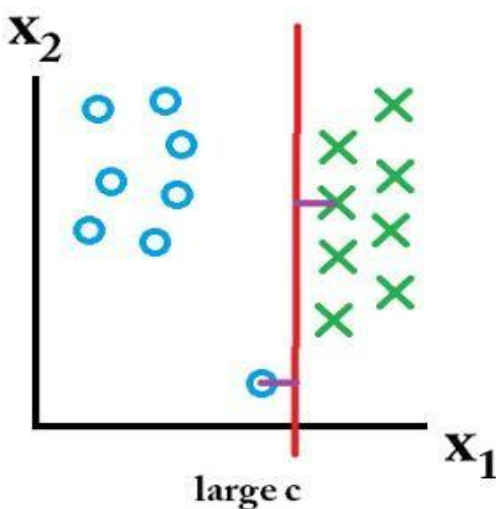
$$\frac{1}{2} ||w||^2 + C \sum_{i=1}^m \xi_i$$

惩罚系数

$$y_i(w^T x_i + b) \geq 1 - \xi_i. \text{ 注意: } \xi_i \geq 0.$$



low c



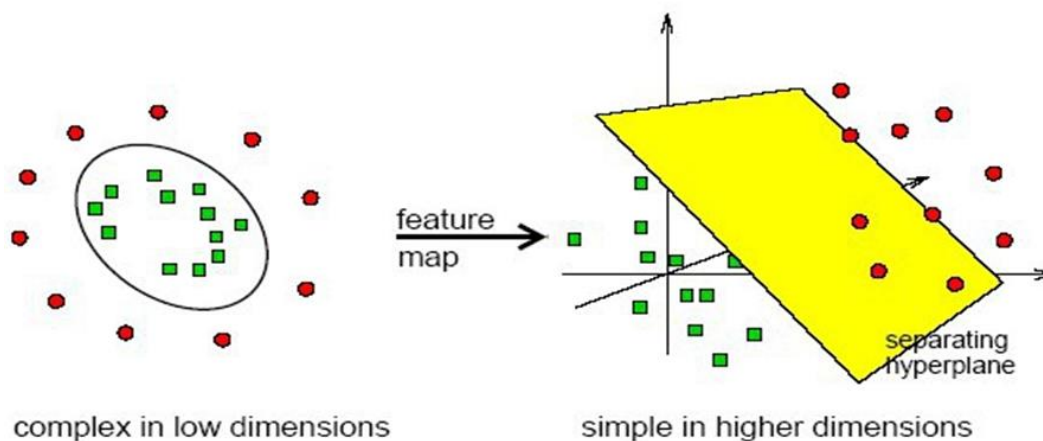
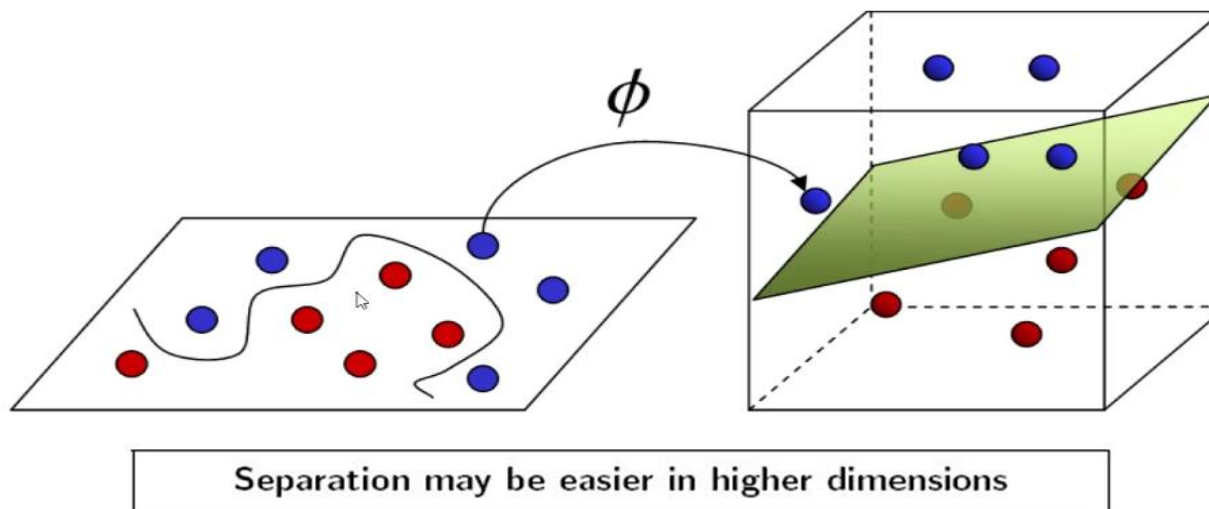
large c

不同的 C 对超平面有不同的影响

超参数寻优

- SVM中的核技巧 (Kernel Trick)

计算两个向量在隐式映射过后的空间中的内积的函数叫做核函数



核函数	公式	调参
linear kernel	$K(x_i, x_j) = x_i^T x_j$	
Polynomial kernel	$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, d > 1$	-d : 多项式核函数的最高次项次数, -g : gamma参数, -r : 核函数中的coef0
Gaussian radial basis function (RBF)	$k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma \ \mathbf{x}_i - \mathbf{x}_j\ ^2)$	-g : gamma参数, 默认值是1/k
Sigmoid kernel	$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r), \gamma > 0, r < 0$	-g : gamma参数, -r : 核函数中的coef0

RacLab

Feature Extraction & Classifier

Yang Ziyang

2024.08.23