

机器学习与数据科学

Machine Learning and Data Science

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

信息与通信工程学院 网络搜索教研中心

北京邮电大学



专题 三：线性模型的扩展

- 内容提要

- 引言

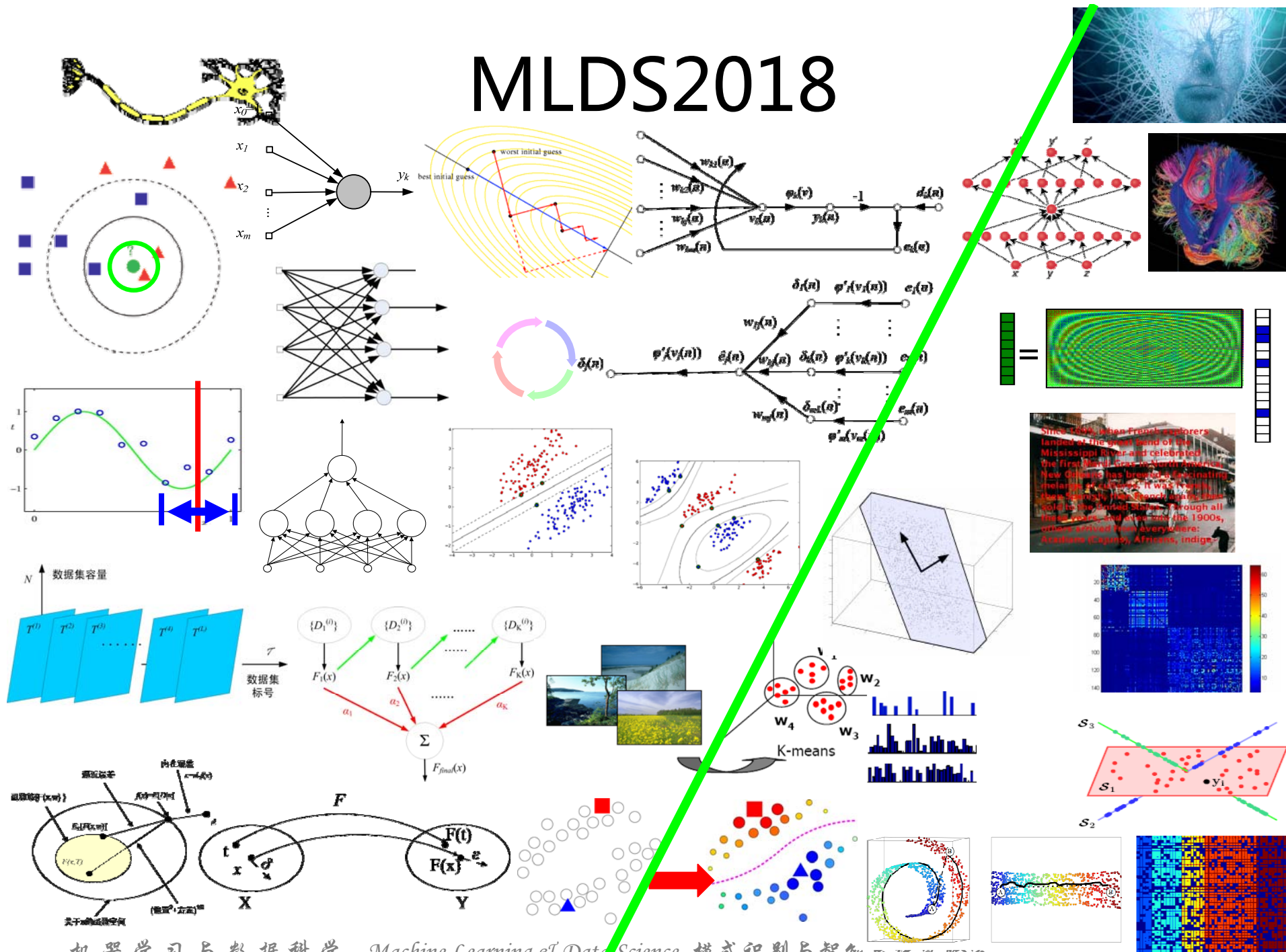
- 广义线性模型

- 核方法

- 多层感知器(MLP)

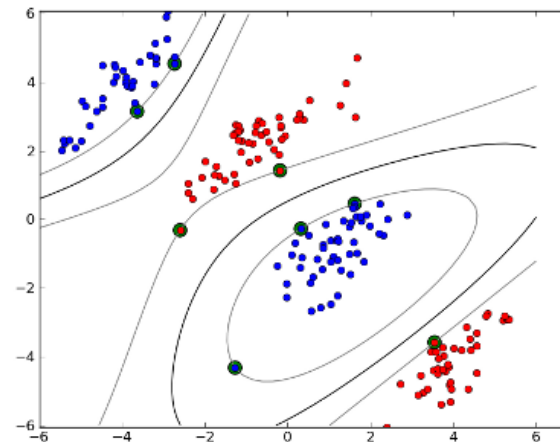
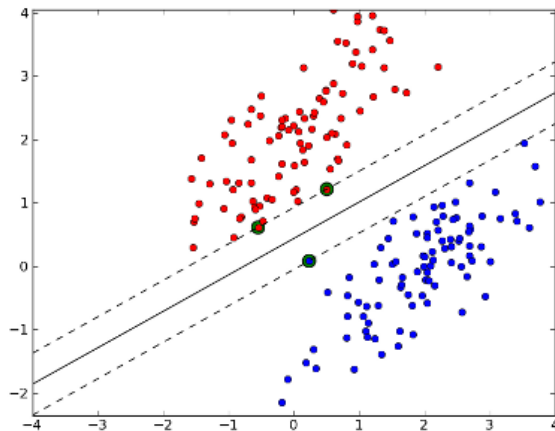
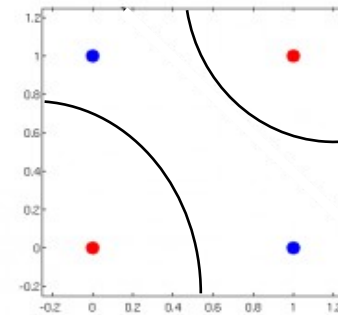
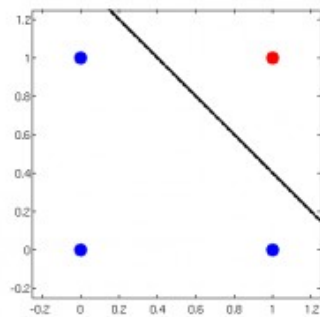
- 误差反向传播算法

MLDS2018



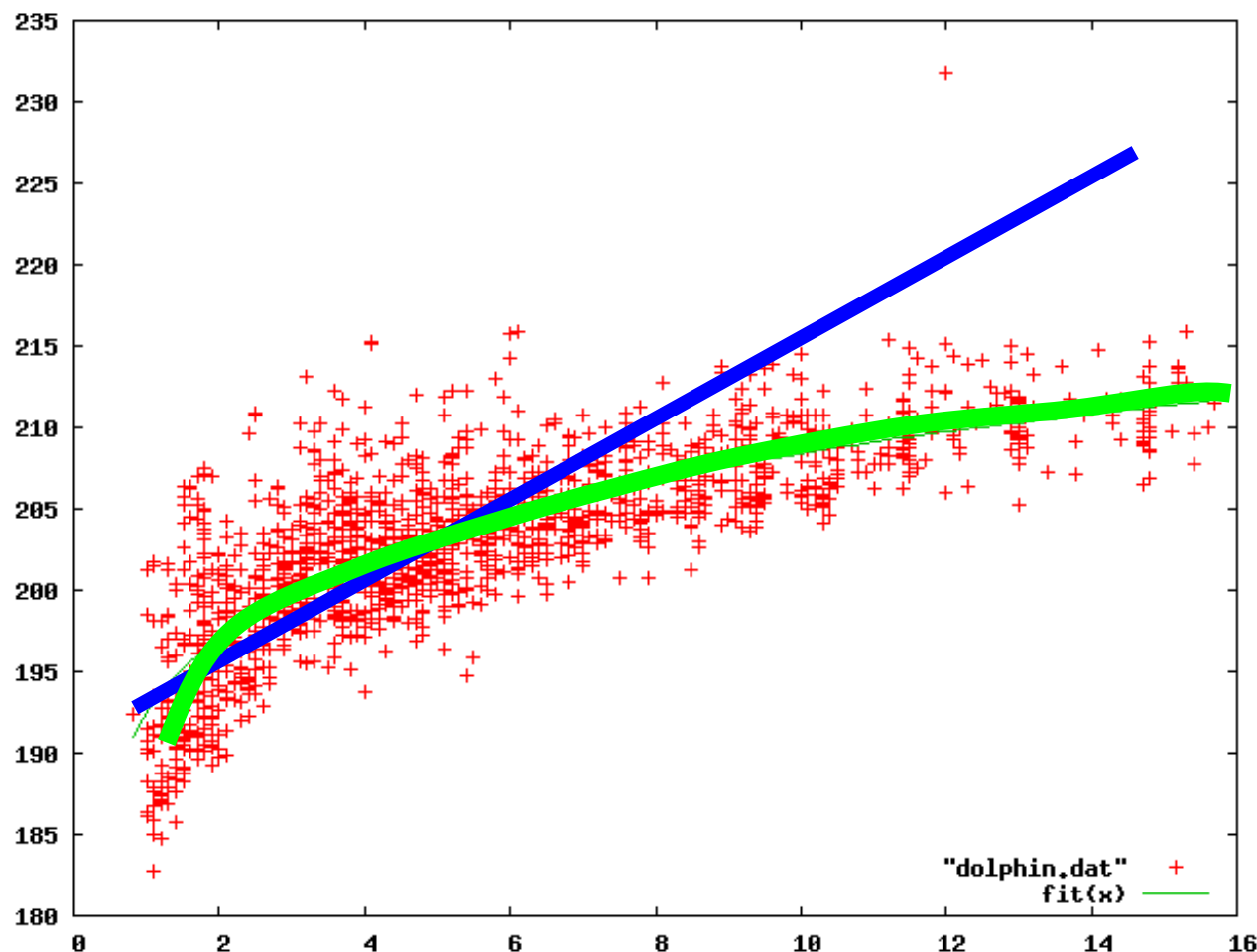
数据并不总是线性可分的...

- 线性可分(linearly separable) \rightarrow 非线性可分(not linearly separable)



变化规律也并不总是线性的...

- 线性回归 vs. 非线性回归



专题 三：线性模型的扩展

- 内容提要

- 引言

- 广义线性模型

- 核方法

- 多层感知器(MLP)

- 误差反向传播算法

线性可分 vs. Φ 可分

- 两个类别是线性可分的, 即存在权值向量 \mathbf{w} , 使得
 - 对类别 \mathbf{C}_1 的样本 \mathbf{x} , 有 $\mathbf{w}^T \mathbf{x} > 0$
 - 对类别 \mathbf{C}_2 的样本 \mathbf{x} , 有 $\mathbf{w}^T \mathbf{x} < 0$
- 两个类别 C_1 和 C_2 是 Φ 可分的
 - 对每个样本 \mathbf{x} , 定义一个由一组实值函数组成的向量 $\Phi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \dots, \varphi_q(\mathbf{x}))^T$, 其中 $\varphi_i : R^m \rightarrow R, i = 1, \dots, q$
 - 如果存在一个权值向量 \mathbf{w} 满足:
 - 对于 $\mathbf{x} \in C_1$, $\mathbf{w}^T \Phi(\mathbf{x}) > 0$
 - 对于 $\mathbf{x} \in C_2$, $\mathbf{w}^T \Phi(\mathbf{x}) < 0$
 - 则两个类别是 Φ 可分的

Φ : 把 \mathbf{x} 从 R^m 映射到
 R^q 空间中

Cover定理

- 非线性映射 Φ 的引入动机
 - 为何引入 Φ ?
 - 如何设计 Φ ?
- Cover定理(定性表述):
 - 把复杂的模式分类问题非线性地投射到高维空间将比投射到低维空间更可能是线性可分的
 - 两个条件
 - (1) 变换 Φ 为非线性的
 - (2) 特征空间的维数足够高, i.e., $q \geq m$

[1] Cover, T.M. (1965). "Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition". IEEE Transactions on Electronic Computers. EC-14: 326–334.8

机器学习与数据科学 - *Machine Learning & Data Science* 模式识别与智能系统实验室

Cover定理

- Theorem :

- A complex pattern-classification problem, cast in a high-dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated.

- Proof:

- 考虑一个2分类问题, 设样本数为 N .
- 把 N 个样本映射到 $N-1$ 维空间中的单纯形(simplex)顶点上. 那么, 每个到两类别的划分(partition)都是线性可分的.

[1] Cover, T.M. (1965). "Geometrical and Statistical properties of systems of linear inequalities with applications in pattern recognition". IEEE Transactions on Electronic Computers. EC-14: 326–334.9

机器学习与数据科学 - Machine Learning & Data Science 模式识别与智能系统实验室

广义线性模型

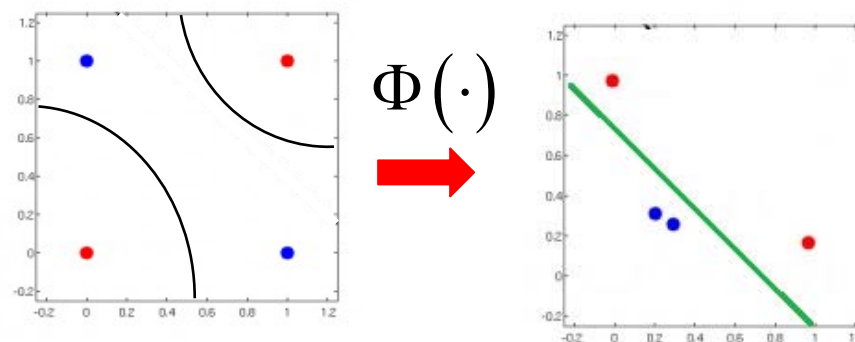
- 对输入数据首先进行非线性变换，然后对变换后的数据进行线性分类或线性回归
 - 首先定义非线性变换： $\Phi: R^m \rightarrow R^q$
 - 对每个输入数据 \mathbf{x} ，计算特征空间中的新表达
$$\Phi(\mathbf{x}) = (\varphi_1(\mathbf{x}), \varphi_2(\mathbf{x}), \dots, \varphi_q(\mathbf{x}))^T$$
 - 然后，基于特征空间中的新表达进行线性分类
$$F(\mathbf{x}) = \text{sgn}(\mathbf{w}^T \Phi(\mathbf{x}))$$

或线性回归
$$F(\mathbf{x}) = \mathbf{w}^T \Phi(\mathbf{x})$$

非线性变换举例 1: XOR

- **XOR问题：**

- 类别 1: (0,0), (1,1)
- 类别 2: (0,1), (1,0)
- 非线性可分！

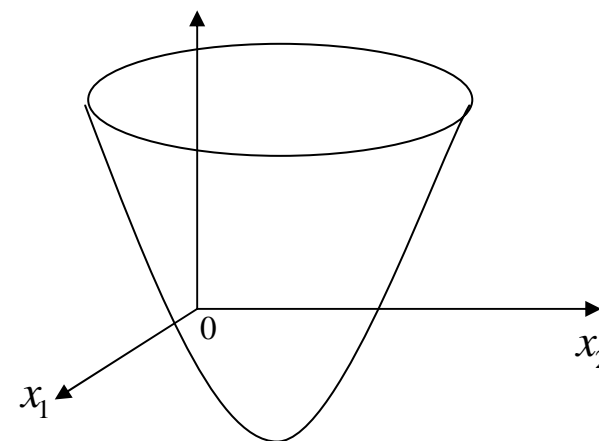


- 采用RBF函数进行非线性变换 $\Phi(\mathbf{x}) = \begin{pmatrix} \varphi_1(\mathbf{x}) \\ \varphi_2(\mathbf{x}) \end{pmatrix}$,
 $\varphi_i(\mathbf{x}) = \exp\left(-\|\mathbf{x} - \mathbf{c}_i\|_2^2\right), i = 1, 2$
 - 其中，设选择中心参数 \mathbf{c}_1 和 \mathbf{c}_2 为(0,0)和(1,1)
- 非线性变换后
 - 类别 1: (0.1353,1.0000), (1.0000,0.1353)
 - 类别 2: (0.3687,0.3687), (0.3687,0.3687)
 - 变成线性可分！

非线性变换举例 2 $f(x_1, x_2)$

- 非线性变换用于回归：

- 设 $f(\mathbf{x}) = x_1^2 + x_2^2 - 2x_1 - 4x_2$
 $= (x_1 - 1)^2 + (x_2 - 2)^2 - 5$

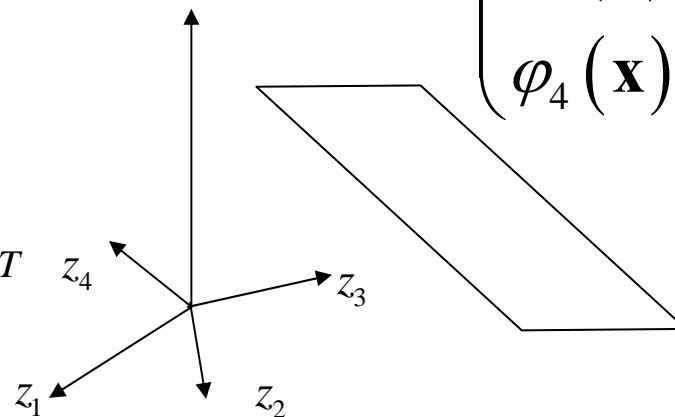


- 若引入非线性变换：

$$\mathbf{z} = \Phi(\mathbf{x}) = \begin{pmatrix} \varphi_1(\mathbf{x}) \\ \varphi_2(\mathbf{x}) \\ \varphi_3(\mathbf{x}) \\ \varphi_4(\mathbf{x}) \end{pmatrix} = \begin{pmatrix} x_1 \\ x_2 \\ x_1^2 \\ x_2^2 \end{pmatrix}$$

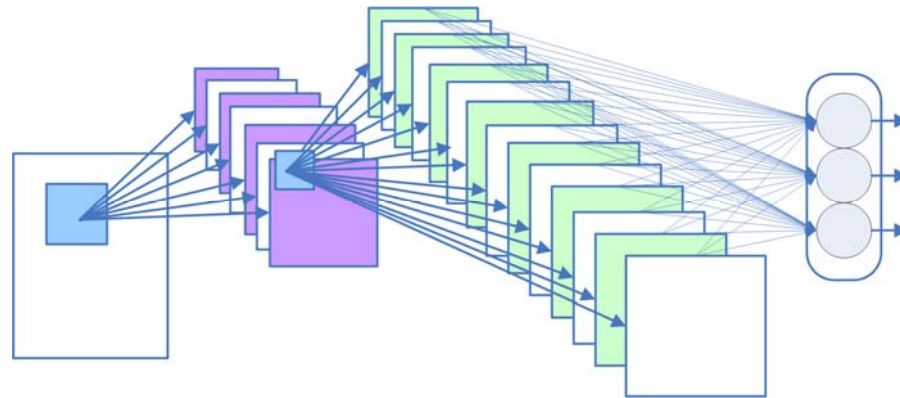
则 $f(\mathbf{z}) = \mathbf{w}^T \mathbf{z}$

其中 $\mathbf{w} = (-2, -4, 1, 1)^T$

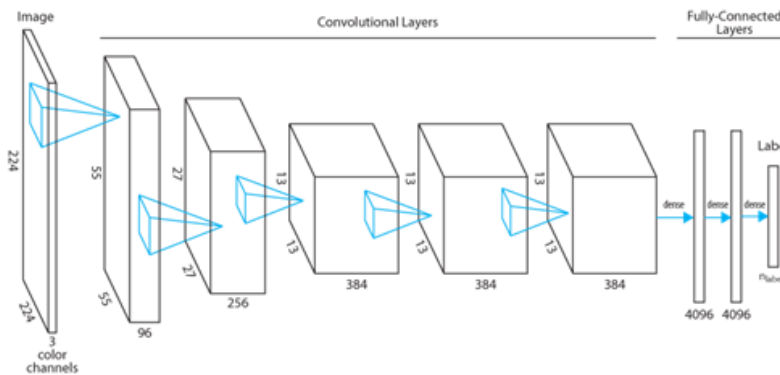


非线性变换举例 3

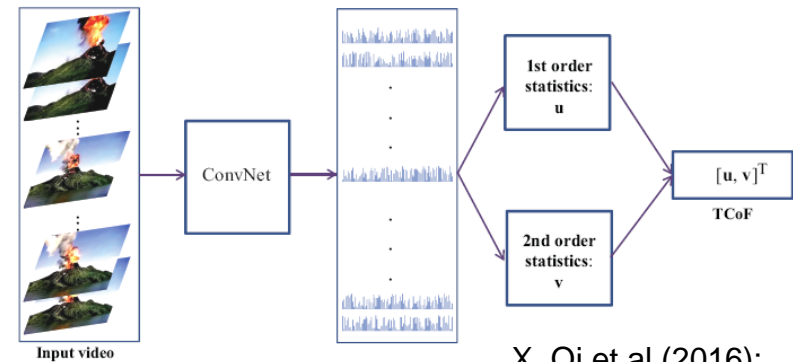
- 卷积神经网络 (Convolutional Neural Network)



– “迁移”训练好的CNN网络用于特征抽取



(a) 图像



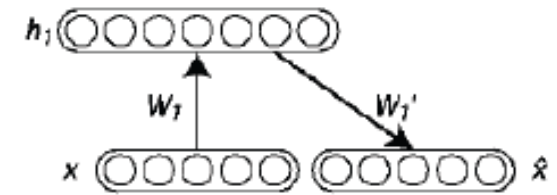
(b) 视频的每帧

X. Qi et al.(2016):
Neurocomputing.

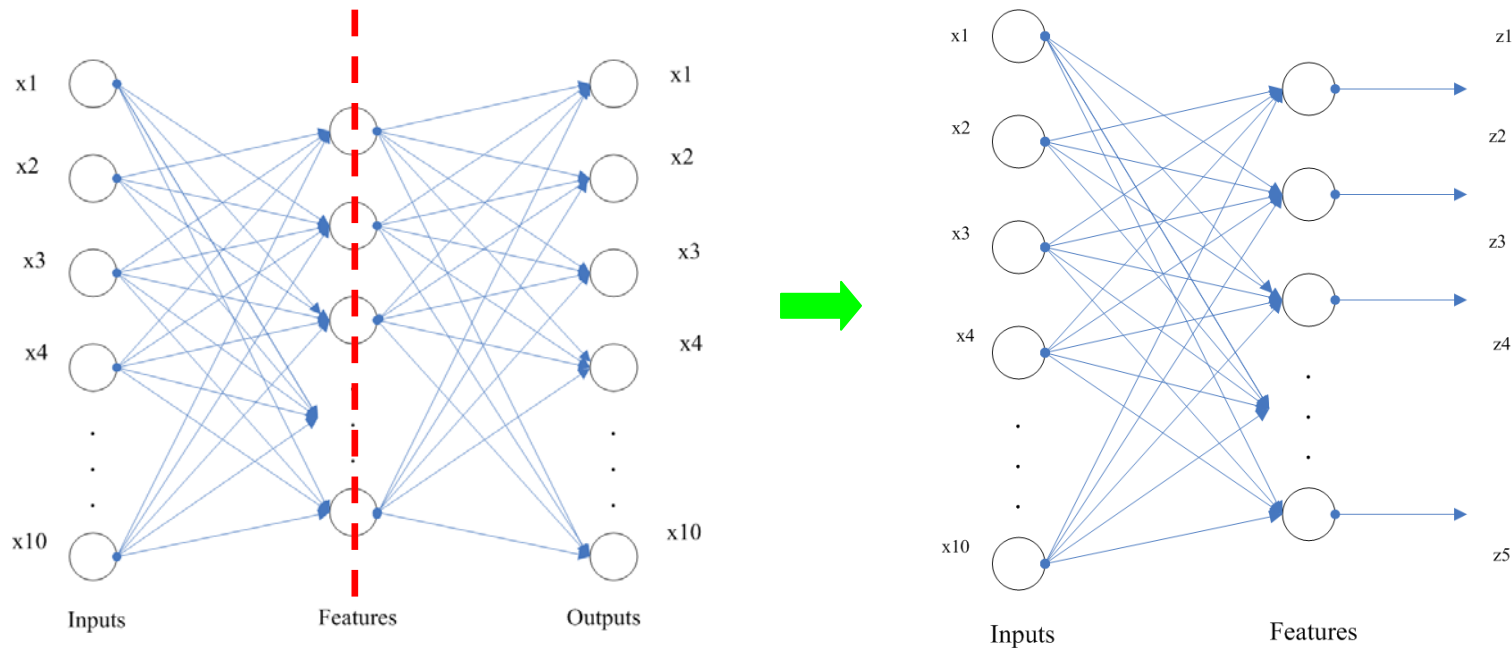
非线性变换举例 4:

- 剪开自编解码(Auto-Encoder)网络
 - 把 $h(\mathbf{x}, W_o)$ 用于特征抽取, 其中

$$W_o = \arg \min_W \sum_{\mathbf{x} \in D} \ell(\mathbf{x}, g(h(\mathbf{x}, W), W))$$



Reconst. $\hat{\mathbf{x}}$



隐含地定义非线性变换

- 定义一个核函数 $K(\cdot, \cdot)$ ，可以隐含地(implicitly)定义从输入空间到特征空间的非线性映射 $\Phi(\mathbf{x})$ ，其中 $\Phi(\mathbf{x})$ 满足：

$$\Phi(\mathbf{x})^T \Phi(\mathbf{x}_i) = K(\mathbf{x}, \mathbf{x}_i)$$

- 算法的核化(kernelization):
 - 若所有涉及到 $\Phi(\mathbf{x})$ 之处，均以其内积形式出现，那么即可把内积替换为核函数的对应形式，即：

$$\Phi(\mathbf{x})^T \Phi(\mathbf{x}_i) \rightarrow K(\mathbf{x}, \mathbf{x}_i)$$

- 这个替换方法也叫做核技巧(kernel trick)或核方法

(可)核化的方法举例

- 在**隐含特征空间**中寻找线性决策超平面
 - 核感知器
 - Kernel Perceptron
 - 支持向量机(**SVM**)
 - 在高维特征空间中寻找最优决策超平面
- 在**隐含特征空间**中寻找聚类(i.e. 向量量化)
 - 核**k**-均值聚类
 - Kernel k-means
- 在**隐含特征空间**中寻找稀疏编码
 - 核稀疏编码
 - Kernel Sparse Coding
- 在**隐含特征空间**中寻找紧致(Compact)编码(i.e. PCA)
 - 核主成分分析
 - Kernel PCA

思考问题: 聚类、稀疏编码、PCA之间有何内在联系?

回顾：感知器算法

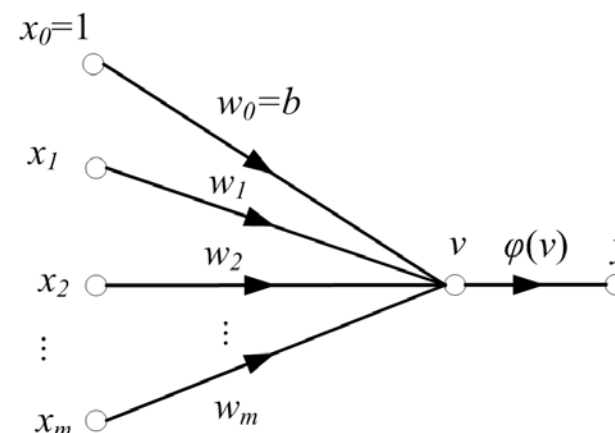
- 感知器模型

$$y = \varphi(\mathbf{w}^T \mathbf{x}) = \text{sgn}(\mathbf{w}^T \mathbf{x})$$

- 感知器的权值更新

$$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \cdot \mathbf{x}_i \quad \text{if } y_i \mathbf{w}^T \mathbf{x}_i < 0$$

$$y_i \mathbf{w}^T \mathbf{x}_i < 0 \quad \text{等价于发生分类错误 } \hat{y}_i \neq y_i, \quad \hat{y}_i = \text{sgn}(\mathbf{w}^T \mathbf{x}_i)$$



— 若 \mathbf{w} 初始化为 $\mathbf{0}$ ，则：

$$\mathbf{w} = \bar{\eta} \sum_{t=1}^T (y_t - \hat{y}_t) \cdot \mathbf{x}_t, \quad \bar{\eta} = \frac{1}{2}$$

$$y = \text{sgn}(\mathbf{w}^T \mathbf{x}) \rightarrow y = \text{sgn} \left(\left(\bar{\eta} \sum_{t=1}^T (y_t - \hat{y}_t) \cdot \mathbf{x}_t \right)^T \mathbf{x} \right)$$

核技巧的使用：例1

- 核感知器：

- 引入非线性变换 $\mathbf{x} \rightarrow \Phi(\mathbf{x})$

- 发生分类错分时，更新权值：

$$\mathbf{w} \leftarrow \mathbf{w} + \eta y_i \cdot \Phi(\mathbf{x}_i) \quad \rightarrow \quad \mathbf{w}^T \Phi(\mathbf{x}) \leftarrow (\mathbf{w} + \eta y_i \cdot \Phi(\mathbf{x}_i))^T \Phi(\mathbf{x})$$

$$y_i \mathbf{w}^T \Phi(\mathbf{x}_i) < 0 \quad \text{等价于分类错误：} \quad \hat{y}_i \neq y_i, \quad \hat{y}_i = \text{sgn}(\mathbf{w}^T \mathbf{x}_i)$$

- 注意到：

$$y = \text{sgn}(\mathbf{w}^T \Phi(\mathbf{x})) = \text{sgn} \left(\left(\eta \sum_{t=1}^T (y_t - \hat{y}_t) \cdot \Phi(\mathbf{x}_t) \right)^T \Phi(\mathbf{x}) \right)$$

定义内积核： $K(\mathbf{x}, \mathbf{x}_i) = \Phi(\mathbf{x})^T \Phi(\mathbf{x}_i)$

$$\rightarrow y = \text{sgn} \left(\eta \sum_{t=1}^T (y_t - \hat{y}_t) K(\mathbf{x}_t, \mathbf{x}) \right)$$

核感知器的训练

•

Algorithm 6 Kernel Perceptron Algorithm

Input: train data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$,

Output: α_t where $t = 1, \dots, T$

Initialize $\alpha_t \leftarrow 0$ for $t = 1, \dots, T$

for $t = 1$ **to** T **do**

$\hat{y}_t \leftarrow \text{sgn} \sum_{s=1}^T \alpha_s y_s K(\mathbf{x}_s, \mathbf{x}_t)$

If $\hat{y}_t \neq y_t$ **then**

$\alpha_t \leftarrow \alpha_t + 1$

end-for

Until no sample is misclassified

核技巧的使用：例2

- 在特征空间中寻找线性决策超平面

$$\mathbf{w}^T \Phi(\mathbf{x}) = 0$$

- 定义决策超平面：

发现内积了没有?



$$\sum_{j=1}^q w_j \varphi_j(\mathbf{x}) + b = 0 \quad \Rightarrow \quad \sum_{j=0}^q w_j \varphi_j(\mathbf{x}) = 0 \quad \Rightarrow \quad \mathbf{w}^T \Phi(\mathbf{x}) = 0$$

假设最优权值为 $\mathbf{w}_o = \sum_{i=1}^N \alpha_i y_i \Phi(\mathbf{x}_i)$

为什么可以这样假设？

– 决策超平面方程可以写为

$$\sum_{i=1}^N \alpha_i y_i \Phi^T(\mathbf{x}_i) \Phi(\mathbf{x}) = 0$$

– 若定义内积核 $K(\mathbf{x}_i, \mathbf{x}) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x})$ ，则决策超平面方程变为

$$\sum_{i=1}^N \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) = 0$$

不含有 $\Phi(\mathbf{x})$

核技巧的使用：例3

- 欧氏距离中的核技巧：


- 欧氏距离： $\|\mathbf{x}_j - \mathbf{x}_i\|_2^2 = (\mathbf{x}_j - \mathbf{x}_i)^T (\mathbf{x}_j - \mathbf{x}_i)$

- 引入非线性变换： $\mathbf{x} \rightarrow \Phi(\mathbf{x})$

- 隐含特征空间中的欧氏距离：

$$\|\Phi(\mathbf{x}_j) - \Phi(\mathbf{x}_i)\|_2^2 = \Phi(\mathbf{x}_j)^T \Phi(\mathbf{x}_j) + \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_i) - 2\Phi(\mathbf{x}_j)^T \Phi(\mathbf{x}_i)$$

定义内积核： $K(\mathbf{x}_j, \mathbf{x}_i) = \Phi(\mathbf{x}_j)^T \Phi(\mathbf{x}_i)$

 $\|\Phi(\mathbf{x}_j) - \Phi(\mathbf{x}_i)\|_2^2 = K(\mathbf{x}_j, \mathbf{x}_j) + K(\mathbf{x}_i, \mathbf{x}_i) - 2K(\mathbf{x}_j, \mathbf{x}_i)$

常用的核函数

- 多项式核(polynomial kernel)
 - 包含全部**2**阶单项式

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}')^2$$

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + c)^2$$

- 包含全部**M**阶单项式

$$k(x, x') = (x^T x')^M$$

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}' + c)^M$$

常用的核函数

- 高斯核(Gaussian kernel)

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\|\mathbf{x} - \mathbf{x}'\|^2 / 2\sigma^2)$$

$$k(\mathbf{x}, \mathbf{x}') = \exp(-\mathbf{x}^T \mathbf{x} / 2\sigma^2) \exp(\mathbf{x}^T \mathbf{x}' / \sigma^2) \exp(-(\mathbf{x}')^T \mathbf{x}' / 2\sigma^2)$$

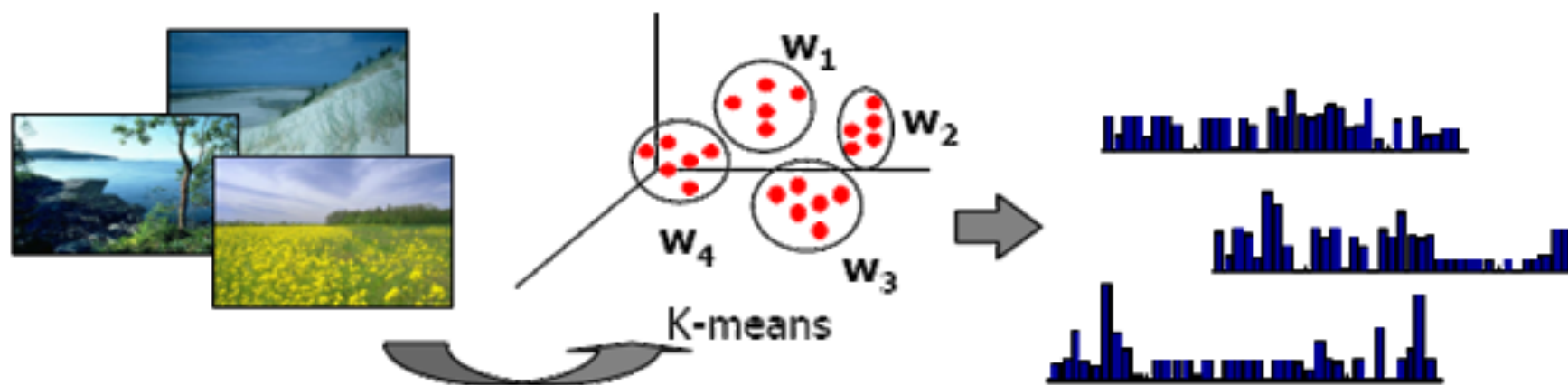
- Sigmoid kernel:

$$k(x, x') = \tanh(ax^T x' + b)$$

直方图相交核

- 直方图相交核(Histogram Intersection Kernel)
 - 两个直方图相交部分的和

$$K(h, g) = \sum_{i=1}^m \min(h_i, g_i)$$



[1] M. Swain and D. Ballard. Color indexing. IJCV, 7(1):11–32, 1991.

Fisher核

- 设参数化生成模型为 $p(x|\theta)$
 - 则**Fisher**得分定义为: $g(x, \theta) = \nabla_{\theta} \ln p(x|\theta)$
- Fisher 核和信息矩阵为:

$$k(x, x') = g(x, \theta)^T \mathbf{F}^{-1} g(x', \theta)$$

- 其中**F**为信息矩阵

$$\mathbf{F} = \mathbb{E}_x [g(x, \theta) g(x, \theta)^T | \theta]$$

- **Fisher** 核对应重参数化具有不变性

- **Fisher**核的样本估计 $\mathbf{F} \simeq \frac{1}{N} \sum_{n=1}^N g(x_n, \theta) g(x_n, \theta)^T$

[1] Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. NIPS 1999.

[2] Jorge Sánchez et al. : Image Classification with the Fisher Vector: Theory and Practice, 2013. 25

Fisher Vector (FV)

- 图像表达方式：

- 抽取局部特征
- 建立高斯混合模型 $G(\pi, \mu_1, \dots, \mu_m, \Lambda_1, \dots, \Lambda_m)$
- 计算对数似然函数对各个参数的梯度

$$\nabla_{\theta} \ln L(\theta, X)$$

- 构造整个图片的表示向量**FV**:

$$\left(\nabla_{\mu_1}, \dots, \nabla_{\mu_m}, \nabla_{\Lambda_1}, \dots, \nabla_{\Lambda_m} \right)^T$$

[1] Jorge Sánchez et al. : Image Classification with the Fisher Vector: Theory and Practice, 2013.

[2] X. Qi, et al.: “HEp-2 Cell Classification via Combining Multi-resolution Co-occurrence Texture and Large Regional Shape Information”, IEEE Journal of Biomedical and Health Informatics, 2017.

非向量类型数据的核函数

- 定义在非向量类型数据的核:
 - 字符串(包括**DNA**序列)之间
 - 可借助编辑距离(Edit Distance)来定义
 - 编辑距离：定义在两个字符串之间，是指由一个字符串转换成另一个字符串所需要进行的最少编辑操作次数, 其中编辑操作包括把字符替换、插入和删除
 - 集合之间
$$k(A_1, A_2) = 2^{|A_1 \cap A_2|}$$
 - 其它类型数据(比如图或树)

核函数的构造: 合成法

- 由给定的核函数合成新的核函数

$$k(\mathbf{x}, \mathbf{x}') = ck_1(\mathbf{x}, \mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = f(\mathbf{x})k_1(\mathbf{x}, \mathbf{x}')f(\mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = q(k_1(\mathbf{x}, \mathbf{x}'))$$

$$k(\mathbf{x}, \mathbf{x}') = \exp(k_1(\mathbf{x}, \mathbf{x}'))$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$$

$$k(\mathbf{x}, \mathbf{x}') = k_3(\phi(\mathbf{x}), \phi(\mathbf{x}'))$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}'$$

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a) + k_b(\mathbf{x}_b, \mathbf{x}'_b)$$

$$k(\mathbf{x}, \mathbf{x}') = k_a(\mathbf{x}_a, \mathbf{x}'_a)k_b(\mathbf{x}_b, \mathbf{x}'_b)$$

核方法的意义

- 提供一种在隐含特征空间中构造学习模型的通用框架
 - 1. 作为线性模型的扩展
 - 对输入数据进行非线性变换，在隐含特征空间中构造线性模型
 - 2. 处理非向量类型数据
 - 对于非向量类型数据，比如字符串、图、树、序列等，通过定义核矩阵或相似度矩阵，即可在适当的特征空间中构造学习模型

Q / A

- Any Question? ...

思考问题

- 1. 对于一个给定核函数 $K(\dots)$ ，它一定对应着从输入空间到特征空间的非线性映射 Φ 的内积吗？
 - 由Mercer定理可知，不一定...
- 2. 给定包含 N 个样本点的数据集和一个核函数 $K(\dots)$ ，核方法的几何意义在哪里？
 - 提供了再生核希尔伯特空间(**RKHS**)的基函数
 - 核方法如何应用在大数据集上？

Mercer定理

- $K(x,y)$ 表示一个连续对称核，其中 x 和 y 定义在闭区间 $[a,b]$ 上，核 $K(x,y)$ 可以被展开成级数

$$K(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(\mathbf{x}) \varphi_i(\mathbf{y}), \quad \lambda_i > 0$$

特征函数,
特征值

为保证该展开式是合理的且绝对一致收敛，其充要条件是：

$$\int_a^b \int_a^b K(\mathbf{x}, \mathbf{y}) \psi(\mathbf{x}) \psi(\mathbf{y}) d\mathbf{x} d\mathbf{y} \geq 0$$

对于所有满足 $\int_a^b \psi^2(\mathbf{x}) d\mathbf{x} < +\infty$ 的 $\psi(\cdot)$ 成立

- 解释：
 - 满足**Mercer**定理的核**K**是正定的，具有**再生性质**
 - 满足**Mercer**定理的核，均存在对应的非线性变换(**Mercer**核映射)
 - 核的特征函数，即为其所隐含的非线性变换
 - 理论上，特征空间的维数可以是无穷大

再生核Hilbert空间

- 由核函数 $K(\cdot, \cdot)$ 所诱导的Hilbert空间，满足如下两条：

- H_K 由 K 展开而成，即
$$H_K = \overline{\text{span}\{K(\mathbf{x}_i, \cdot), \mathbf{x}_i \in X\}}$$
- K 具有再生性质，即
$$\langle f(\cdot), K(\cdot, \mathbf{x}_i) \rangle_{H_K} = f(\mathbf{x}_i)$$


- 几点解释

- 集合 \rightarrow 空间
 - 集合 + 运算法则 (e.g. 8 条)
- 线性空间 \rightarrow 内积空间
 - 定义内积
- 内积空间 \rightarrow Hilbert 空间
 - 完备性
- 从Kernel函数张成的集合 \rightarrow 内积空间
 - 定义双线性形式，验证对称、双线性特性和平方范数
- Kernel的Reproducing 性质
 - 从双线性形式出发，验证再生性质
- RKHS表现定理
$$f(\mathbf{x}) = \sum_{i=1}^N \alpha_i K(\mathbf{x}_i, \mathbf{x})$$

正定核 K 具有 “可再生” 性质

- K的可再生性质，即 $\langle f(\cdot), K(\cdot, \mathbf{x}_i) \rangle_{H_K} = f(\mathbf{x}_i)$,

给定 $f(\mathbf{x}) = \sum_{i=1}^N \alpha_i K(\mathbf{x}_i, \mathbf{x})$, $\langle K(\cdot, \mathbf{x}_i), K(\cdot, \mathbf{x}_j) \rangle_{H_K} = K(\mathbf{x}_i, \mathbf{x}_j)$


$$\begin{aligned} \langle f(\cdot), K(\cdot, \mathbf{x}_i) \rangle_{H_K} &= \left\langle \sum_{j=1}^N \alpha_j K(\mathbf{x}_j, \cdot), K(\cdot, \mathbf{x}_i) \right\rangle_{H_K} \\ &= \sum_{j=1}^N \alpha_j \langle K(\mathbf{x}_j, \cdot), K(\cdot, \mathbf{x}_i) \rangle_{H_K} \\ &= \sum_{j=1}^N \alpha_j K(\mathbf{x}_i, \mathbf{x}_j) \\ &= f(\mathbf{x}_i) \end{aligned}$$

表现(Representer)定理

- 表现定理的意义——给出“表达形式”

$$\min_{f \in H_K} \sum_{i=1}^N \ell(y_i, f(\mathbf{x}_i)) + \lambda \cdot \Omega(f)$$

– 若正则化项形式为： $\Omega(f) = \|f\|_{H_K}^2$

则其解的形式：
$$f(\mathbf{x}) = \sum_{i=1}^N \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

Q / A

- Any Question? ...