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

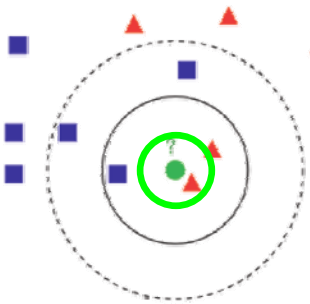
模式识别与智能系统实验室

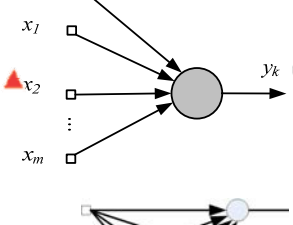
信息与通信工程学院 网络搜索教研中心

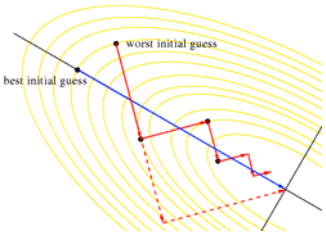
北京邮电大学

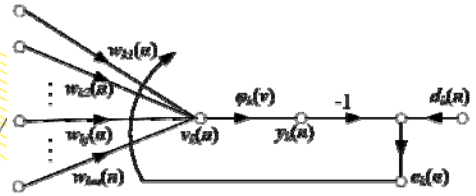
专题 七：正则化理论

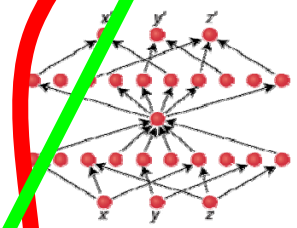
- **内容提要**
  - 引言
  - 正则化理论与一般形式
  - 典型算法
    - 正则化最小二乘
    - SVM
    - 正则化网络
    - 核密度估计

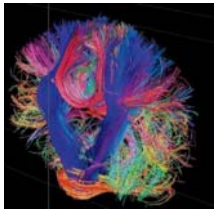


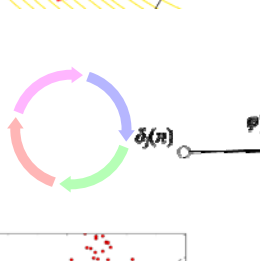


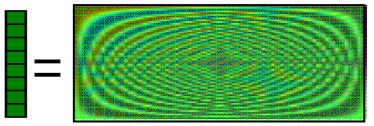


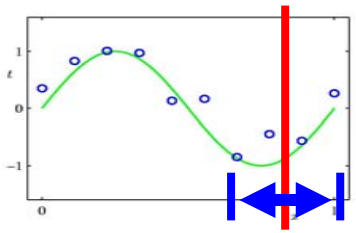


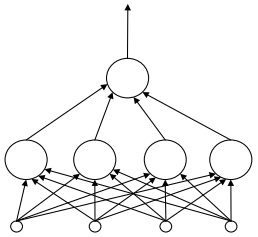


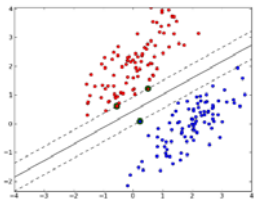


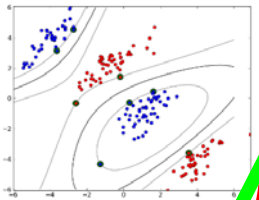


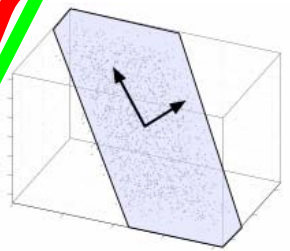


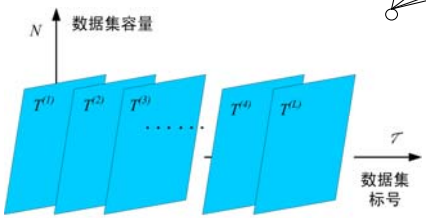


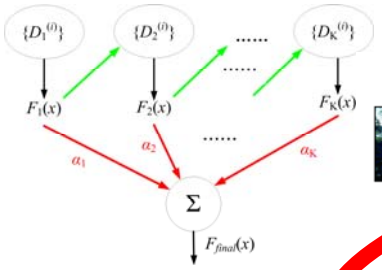


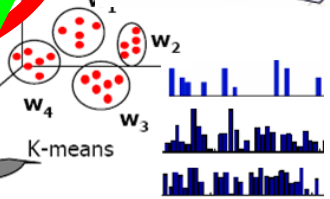


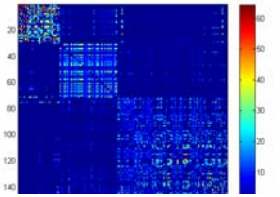


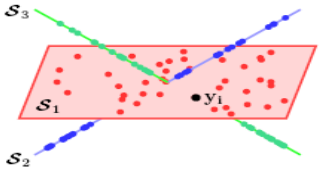


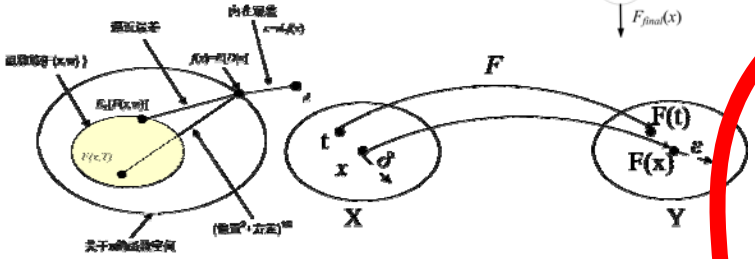


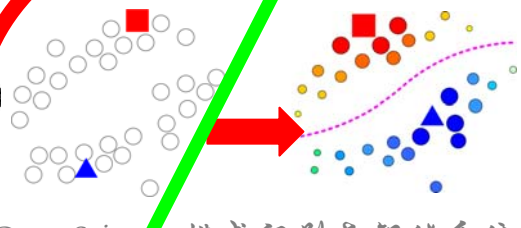


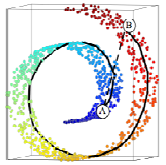


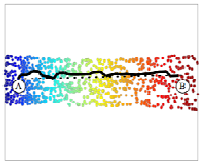


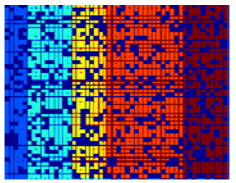












MLDS 2018

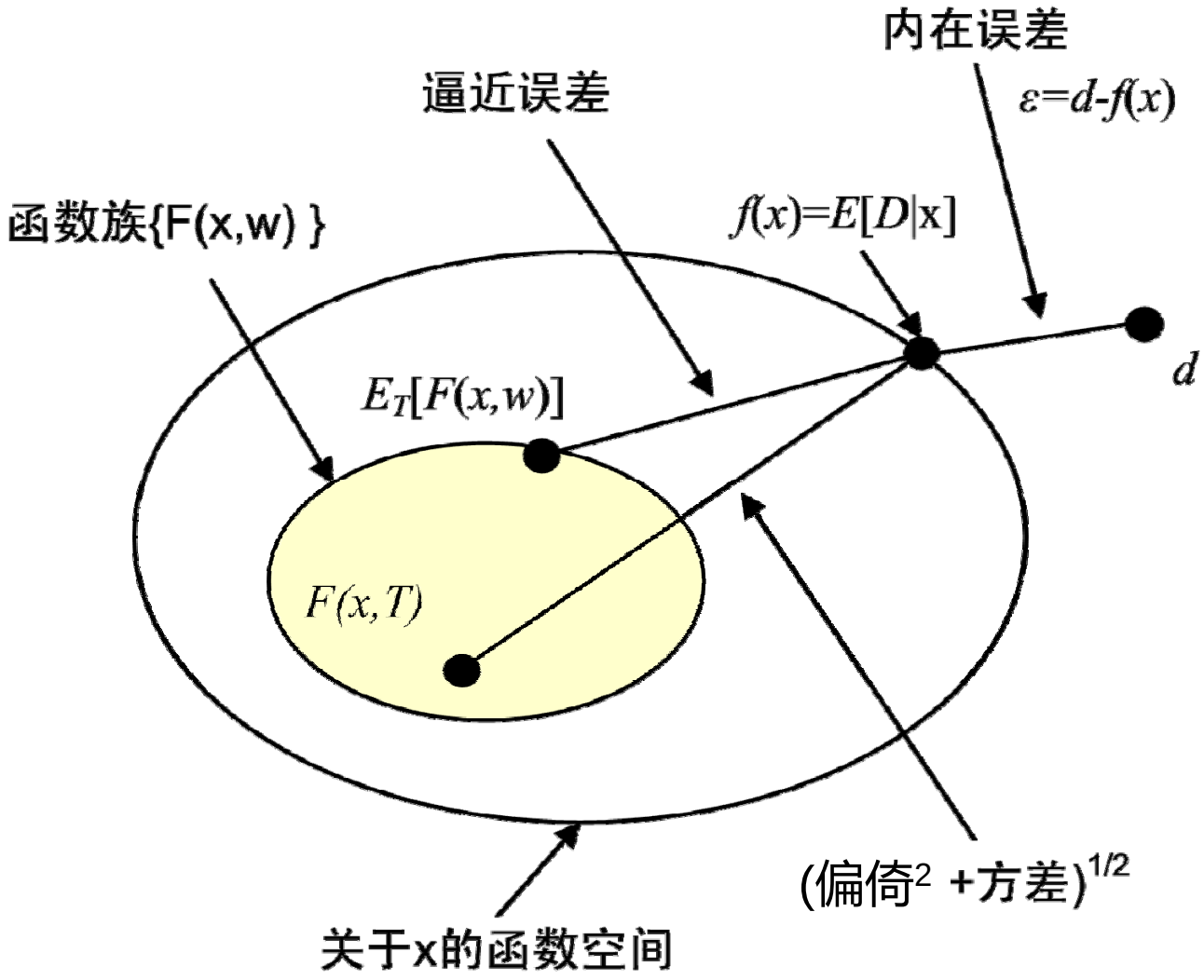
x_0
x_1
x_2
x_m
y_k
worst initial guess
best initial guess
$w_{k1}(n)$
$w_{k2}(n)$
$w_{kj}(n)$
$w_{km}(n)$
$\varphi_k(v)$
$v_k(n)$
$y_k(n)$
-1
$d_k(n)$
$e_k(n)$
$\delta_j(n)$
$\varphi'_j(v_j(n))$
$e_j(n)$
$w_{1j}(n)$
$w_{kj}(n)$
$w_{mj}(n)$
$\delta_1(n)$
$\delta_k(n)$
$\delta_{m}(n)$
$\varphi'_1(v_1(n))$
$\varphi'_k(v_k(n))$
$e_1(n)$
$e_k(n)$
$e_m(n)$
x
y
z
x'
y'
z'
=
Since 1699, when French explorers landed at the great bend of the Mississippi River and celebrated the first Mardi Gras in North America, New Orleans has brewed a fascinating melange of cultures. It was Spanish, then French, then Spanish again, then sold to the United States. Through all these years, and even into the 1900s, others arrived from everywhere: Acadians (Cajuns), Africans, indige-
1
0
−1
t
N
数据集容量
$T^{(1)}$
$T^{(2)}$
$T^{(3)}$
$T^{(4)}$
$T^{(L)}$
$\mathcal{T}$
数据集
标号
$\{D_1^{(i)}\}$
$\{D_2^{(i)}\}$
$\{D_K^{(i)}\}$
$F_1(x)$
$F_2(x)$
$F_K(x)$
α_1
α_2
α_K
Σ
$F_{final}(x)$
w_1
w_2
w_3
w_4
K-means
s_1
s_2
s_3
y_1
F
t
x
X
F(t)
F(x)
Y
ε
δ
A
B

经验风险最小化(Empirical Risk Minimization)

- ERM原则
  - 通过构建经验风险泛函来代替风险泛函，通过在训练数据集$D=\{(\mathbf{x}_i,y_i)\}_{i=1}^N$上最小化经验风险泛函$R_{\text{emp}}(\mathbf{w})$来寻找逼近函数$F(\mathbf{x},\mathbf{w})$，即寻找$F(\mathbf{x},\mathbf{w}_{emp})$，其中

$$\mathbf{w}_{emp}=\arg\min_{\mathbf{w}} R_{emp}(\mathbf{w})=\arg\min_{\mathbf{w}}\frac{1}{N}\sum_{i=1}^{N}\ell\left(y_i,F(\mathbf{x}_i,\mathbf{w})\right)$$

偏倚-方差分解(Bias-Variance Decomposition)



经验风险最小化原则(ERM)

- ERM原则
  - 1. 可以构建经验风险泛函 $R_{emp}(\mathbf{w})$ 代替风险泛函 $R(\mathbf{w})$, 即基于i.i.d.训练样本 $D=\{(\mathbf{x}_i,y_i)\}_{i=1}^{N}$, 定义

$$\mathbf{w}_{emp}=\arg\min_{\mathbf{w}} R_{emp}(\mathbf{w})=\arg\min_{\mathbf{w}}\frac{1}{N}\sum_{i=1}^{N}\ell\left(y_i,F(\mathbf{x}_i,\mathbf{w})\right)$$

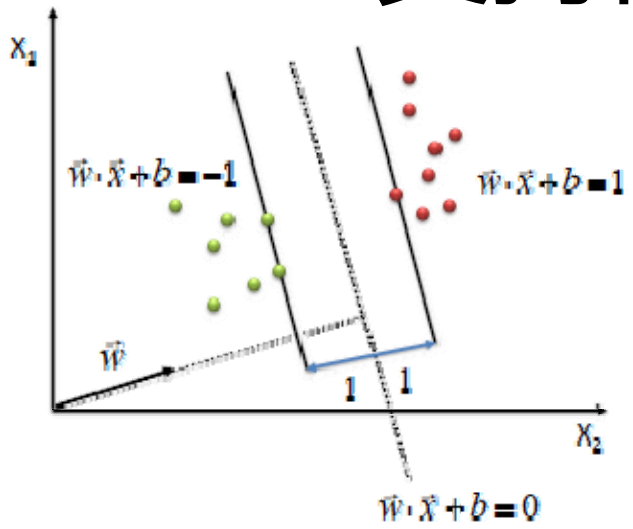
  - 2. 只要当训练样本的数量**N**趋于无穷大时，经验风险泛函**一致收敛**于实际风险泛函，那么，最小化经验风险 $R_{emp}(\mathbf{w})$ 的 $\mathbf{w}_{\text{emp}}$ 所对应的实际风险 $R(\mathbf{w}_{\text{emp}})$ **依概率收敛**到实际风险 $R(\mathbf{w}_{\text{emp}})$ 的最小可能值，即 $\mathrm{P}\{R(\mathbf{w}_{emp})-R(\mathbf{w}_o)<2\varepsilon\}>1-\alpha$

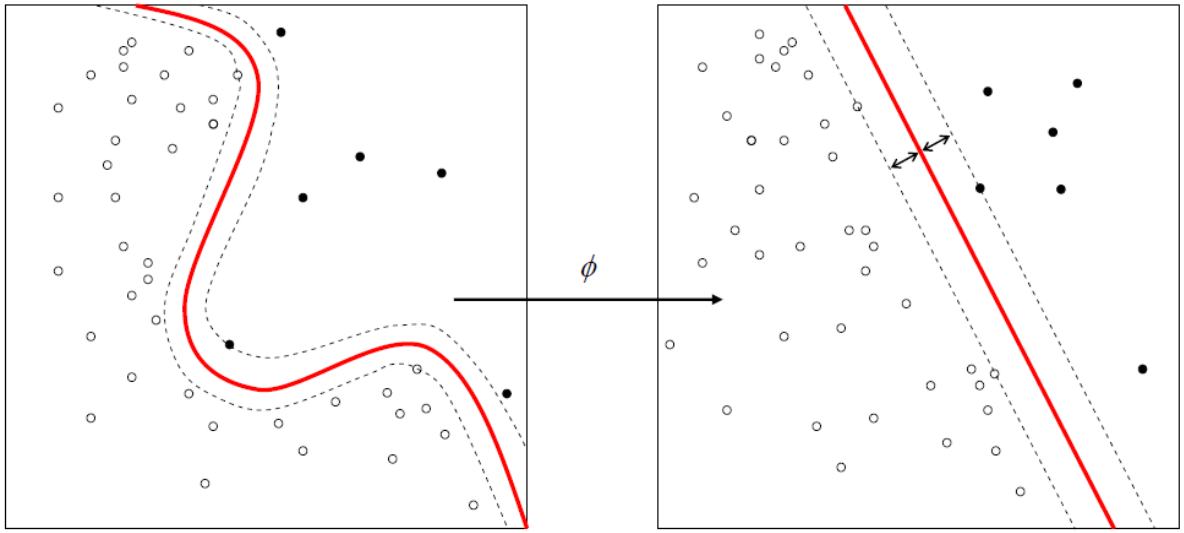
结构风险最小化(Structural Risk Minimization)

- **结构风险最小化方法**
  - 通过控制VC维，以使学习机器的容量与训练数据有效数量相匹配

  - 考虑学习机器 $\Psi = \{F(\mathbf{x},\mathbf{w});\mathbf{w}\in W\}$，定义n个这样学习机器的嵌套结构 $\Psi_t = \{F(\mathbf{x},\mathbf{w});\mathbf{w}\in W_t\}, t=1,...,n$，使得 $\Psi_1 \subset \Psi_2 \subset \cdots \subset \Psi_n$ 且各个学习机器的VC维满足递增条件 $h_1 \le h_2 \le \cdots \le h_n$，那么，结构风险最小化方法可以如下进行：
    - 对每个学习机器，**最小化其经验风险**
      - 即最小化训练误差
    - 寻找具有**最小保证风险**的学习机器
      - 即在训练误差与逼近函数复杂性之间寻求折衷

支持向量机

- 
X_1
$\vec{w}\cdot\vec{x}+b=-1$
$\vec{w}\cdot\vec{x}+b=1$
$\vec{w}$
1
1
X_2
$\vec{w}\cdot\vec{x}+b=0$


ϕ

引子：学习问题的另一种观点

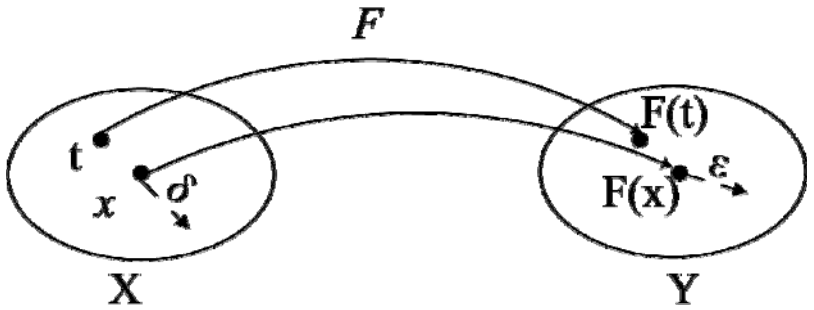
- 学习问题
  - 统计意义下的最优化问题
  - 看作——基于给定一组数据点的超曲面重建问题
    - **训练一个学习机器使其根据输入模式找到相应的输出模式，相当于学习一个超曲面(即多维映射)使其能够根据输入确定输出**
- 互逆问题
  - 对于两个相关问题，如果系统地解决其中任意一个问题都必须部分地或者全部地知道关于另一个问题的知识
    - **正问题(direct problem)：**
      - **研究得比较早、比较透彻的问题**
    - **逆问题(inverse problem)**
      - **区别：从数学角度看，是否“适定”**

适定 vs. 非适定

集合X上规定了一个度量

- 适定的(well-posed)
  - 在度量空间中一个定义域 X 和一个值域Y，由一个固定的未知映射F所联系，如果下面关于映射F的3个条件都满足的话，称映射 F的重建问题是适定的
    - **1. 存在性**
    - **2. 唯一性**
    - **3. 连续性(也称为稳定性)**

      $\|F(x+\delta)-F(x)\| \le \varepsilon$



  - 任何一项条件不满足，则称为不适定的(ill-posed)

    如果问题不适定，则说明大量数据中只包含很少一部分的有用信息

  - *举例*：生成训练数据的物理过程是适定的正问题，而从这些数据进行的对物理过程的学习问题(视为超曲面重建问题)是不适定问题
    - 唯一准则可能不满足：
    - 存在准则可能不满足：
    - 噪声和不精确性的影响，**连续性**可能不满足：

如果一个学习问题不具有连续性，则所得输入-输出映射与其真解将毫无关系！

- **内容提要**

  - 引言
  - 正则化理论与一般形式
  - 典型算法
    - 正则化最小二乘
      - 正则化最小二乘的核扩展
    - SVM
    - 正则化网络
    - 核密度估计

正则化(Regularization)

- 目的:
  - 把不适定问题转化为适定问题
- 策略:
  - 通过引入某些含有解的先验知识的非负的辅助泛函来使问题的解稳定
    - 这个非负辅助泛函称为**正则化泛函**
      - 从统计学习理论角度，正则化泛函施加对学习机器容量的控制
      - 在贝叶斯观点中，正则化泛函相当于一个先验分布
- 正则化方法:
  - 把适当的正则化泛函引入到待求解问题的目标泛函中，从而求解一个增加了正则化项的问题的方法

正则化的3种类型

- Tikhonov型(变分法)

$$\arg\min_f \varepsilon(f) + \lambda \cdot \Omega(f)$$

  - 其中 $\varepsilon(f)$：标准误差项

    $\Omega(f)$：正则化泛函　　$\lambda \geq 0$：正则化参数

- Ivanov型(残差法)

$$\arg\min_f \ \Omega(f) \ \text{s.t.} \ \varepsilon(f) \leq \delta$$

  - 其中 $\delta \geq 0$：预定义的常数

- Philips型(拟解法)

$$\arg\min_f \ \varepsilon(f) \ \text{s.t.} \ \Omega(f) \leq \rho$$

  - 其中 $\rho \geq 0$：预定义的常数

Tikhonov正则化框架

- 求解最小化Tikhonov型的目标泛函

$$f_\lambda^* = \arg\min_f \varepsilon(f) + \lambda \cdot \Omega(f)$$

其中

$\varepsilon(f)$：标准误差项，依赖于训练样本

$\Omega(f)$：正则化项，包含关于问题的解的先验知识

λ：正则化参数

– 正则化参数看作一个指示器，指示所给的数据用于确定问题的解的样本的充分性

  - 当λ→0时，表明该问题不受约束，问题的解完全取决于所提供的训练样本
  - 当λ→∞时，表明仅由正则化项就足以得到问题的解，说明所给训练数据完全不可信

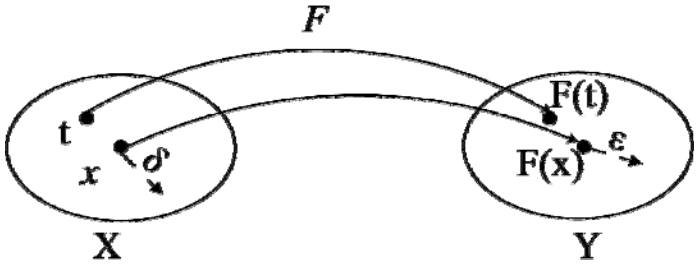
从统计学习理论角度看，正则化参数控制着学习机器(或函数族)的容量

Tikhonov正则化的理论保证

- 考虑不适定方程： $A \cdot f = \mathbf{y}$
  - 若给定的观测不是准确的**y**，而是其近似 $\mathbf{y}_\delta$ ：

$$\left\| \mathbf{y} - \mathbf{y}_\delta \right\|_2 \le \delta$$

  - 假定正则化参数 $\lambda(\delta)$ 满足如下条件
    - 1. $\delta \to 0, \lambda(\delta) \to 0$
    - 2. $\lim_{\delta \to 0} \delta^2 / \lambda(\delta) \le t < +\infty$



  - 那么，当 $\delta \to 0$ 时，最小化**Tikhonov**泛函的解 $f^*_{\lambda(\delta)}$ 收敛于问题的准确解f^*，其中

$$f^*_{\lambda(\delta)} = \arg\min_f \left\| A \cdot f - \mathbf{y}_\delta \right\|_2^2 + \lambda \cdot \Omega(f)$$

Tikhonov & Arsenin, 1977

- **内容提要**

  - 引言
  - 正则化理论
  - 正则化问题的一般形式
  - 典型算法
    - 正则化最小二乘
      - 正则化最小二乘的核扩展
    - SVM
    - 正则化网络
    - 核密度估计

正则化问题的一般形式

- 正则化问题的一般形式

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

1. 正则化项具有不同形式
2. 损失函数具有不同形式
3. 函数空间具有不同形式

正则化项的不同形式

- **正则化项的常用形式**
  - 线性函数的权向量L_p范数定义的正则化项

$$\Omega(f)=\|\mathbf{w}\|_p^p$$

  - 再生核Hilbert空间中L_2范数定义的正则化项

$$\Omega(f)=\|f\|_{H_K}^2$$

  - 采用线性微分算子的正则化项

$$\Omega(f)=\|\mathbf{D}f\|^2$$

  - 采用傅立叶变换的正则化项 $\Omega(f)=\int_{R^d}\frac{|\tilde{f}(s)|}{\tilde{G}(s)}ds$
  - 流形(manifold)正则化项 $\Omega(f)=\mathbf{f}^T L\mathbf{f}=\frac{1}{2}\sum_{i,j=1}^{N}w_{ij}\left(f_i-f_j\right)^2$

损失函数的不同形式

- 损失函数的几种常用形式
  - 误差的平方

$$\ell\left(y_i - f\left(\mathbf{x}_i\right)\right) = \left(y_i - f\left(\mathbf{x}_i\right)\right)^2$$

  - **Epsilon-insensitive norm**
    - SVM for regression

$$\ell\left(y_i, f\left(\mathbf{x}_i\right)\right) = \left|y_i - f\left(\mathbf{x}_i\right)\right|_\varepsilon = \begin{cases} \left|y_i - f\left(\mathbf{x}_i\right)\right| - \varepsilon, & \text{if } \left|y_i - f\left(\mathbf{x}_i\right)\right| \geq \varepsilon \\ 0, & \text{otherwise} \end{cases}$$

  - **Soft Margin loss function**
    - SVM for classification

$$\ell\left(y_i, f\left(\mathbf{x}_i\right)\right) = \left|1 - y_i f\left(\mathbf{x}_i\right)\right|_+ = \begin{cases} 1 - y_i f\left(\mathbf{x}_i\right), & \text{if } 1 - y_i f\left(\mathbf{x}_i\right) > 0 \\ 0, & \text{otherwise} \end{cases}$$

    - Hard margin loss function

$$\ell\left(y_i, f\left(\mathbf{x}_i\right)\right) = \theta\left(1 - y_i f\left(\mathbf{x}_i\right)\right)$$

Hinge-Loss

    - Misclassification loss function

$$\ell\left(y_i, f\left(\mathbf{x}_i\right)\right) = \theta\left(-y_i f\left(\mathbf{x}_i\right)\right)$$

$\theta\left(\cdot\right)$：**阶跃函数**

函数空间的不同形式

- 线性函数
  - 线性泛函
    - Rietz表现定理

$$\forall \mathbf{x} \in H,\ \exists \mathbf{w} \in H,\quad f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{w} \rangle$$

- 非线性函数
  - 如何表示？如何参数化？
    - 简单函数的复合，比如多层感知器(MLP)
  - 再生核希尔伯特空间
    - Reproducing Kernel Hilbert Space
      - RKHS表现定理

$$f \in H_K \Rightarrow f(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x})$$

再生核Hilbert空间

- 由核函数K(.,.)所诱导的Hilbert空间，满足如下两条：
  - **H_k**由**K**展开而成，即 $$H_k = \overline{span\left\{k\left(\mathbf{x}_i, \cdot\right), \mathbf{x}_i \in X\right\}}$$
  - **K**具有再生性质，即 $$\left\langle f\left(\cdot\right), k\left(\cdot, \mathbf{x}_i\right)\right\rangle_{H_K} = f\left(\mathbf{x}_i\right)$$
- 几点解释
  - 集合 → 空间
    - 集合 + 运算法则 (e.g. 8 条)
  - 线性空间 → 内积空间
    - 定义内积
  - 内积空间 → **Hilbert**空间
    - 完备性
  - 从**Kernel**函数张成的集合 → 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 $$f\left(\mathbf{x}\right) = \sum_{i=1}^{N} \alpha_i k\left(\mathbf{x}_i, \mathbf{x}\right)$$

表现定理(Representer Theorem)

- **表现定理的意义——给出“表达形式”**
  - 在这里，表现定理给出正则化问题解的基本形式

- 对于正则化问题 $\min_{f\in H_K}\sum_{i=1}^{N}\ell\left(y_i, f\left(\mathbf{x}_i\right)\right)+\lambda\cdot\Omega\left(f\right)$

  - 若正则化项形式为：$\Omega\left(f\right)=\left\|f\right\|_{H_K}^2$，则其解的形式：

$$f\left(x\right)=\sum_{i=1}^{N}\alpha_i k\left(x, x_i\right)$$

  - 若采用傅立叶变换的正则化项 $\Omega\left(f\right)=\int_{R^d}\frac{\left|\tilde{f}\left(s\right)\right|}{\tilde{G}\left(s\right)}ds$

则其解的形式为

$$f\left(x\right)=\sum_{i=1}^{N}\beta_i G\left(x-x_i\right)+\sum_{k=1}^{K}\alpha_k\phi_k\left(x\right)$$

G(s)的傅里叶反变换

正则化项对应的零空间

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

Q / A

- Any Question? …

- **内容提要**

  - 引言
  - 正则化理论
  - 正则化问题的一般形式
  - 典型算法
    - 正则化最小二乘
      - 正则化最小二乘的核扩展
    - 正则化网络
    - 核密度估计