模式识别引论

An Introduction to Pattern Recognition

主讲: 李春光

www.pris.net.cn/teacher/lichunguang

模式识别与智能系统实验室

网络搜索教研中心 信息与通信工程学院 北京邮电大学

核方法 内容提要

- 引子: 2个类别的分类问题
- 对偶表示
  - 正则化最小二乘法求解线性回归
- 核函数的构造
  - 生成法则
  - **Fisher** 核
- 高斯过程(Gaussian Processes)
  - 高斯过程 **for** 回归
  - 高斯过程 **for** 分类

引言

- 训练数据: keep or discard？
- **基于参数(parameter-based)的方法**
  - 获得参数向量**w**, 丢弃训练数据
    - 用于回归和分类的线性参数模型
    - 神经网络(非线性模型)
- **基于记忆(memory-based)的方法**(Non-**parametric**)
  - 保存训练数据或训练数据的子集
    - **Parzen Windows**法用于密度估计
    - 最近邻**(nearest neighbors)**方法用于分类

“核方法”的几个特点

- 使用对偶表示(dual representation)
- 基于核函数的线性组合进行预测
- 在训练数据上计算核函数，获得核矩阵
- 获得一个固定的非线性特征空间映射**φ(x)**

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

核函数与核技巧

- 最简单的核函数——线性核
  - $\phi(x) = x$
  - $\boldsymbol{k}(x, x) = x^T x$

- 核技巧(kernel trick)
  - 把“内积”替换为核函数

    举例:
    - **Kernel PCA**
    - **Kernel Fisher 鉴别分析**

核方法 内容提要

- 引子: 2个类别的分类问题
- 对偶表示
  - 正则化最小二乘法求解线性回归
- 核函数的构造
  - 生成法则
  - **Fisher** 核
- 高斯过程(Gaussian Processes)
  - 高斯过程 **for** 回归
  - 高斯过程 **for** 分类

线性回归模型的对偶表示 1

- 线性回归模型

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

- 使用带正则化项的最小二乘法估计参数w

$$J(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{ \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n) - t_n \right\}^2 + \frac{\lambda}{2} \mathbf{w}^{\mathrm{T}} \mathbf{w}$$

  – 对w求梯度，整理后得到：

$$\Rightarrow \quad \mathbf{w} = -\frac{1}{\lambda} \sum_{n=1}^{N} \left\{ \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n) - t_n \right\} \phi(\mathbf{x}_n) = \sum_{n=1}^{N} a_n \phi(\mathbf{x}_n) = \mathbf{\Phi}^{\mathrm{T}} \mathbf{a}$$

    - 其中

$$a_n = -\frac{1}{\lambda} \left\{ \mathbf{w}^{\mathrm{T}} \phi(\mathbf{x}_n) - t_n \right\}$$

并不去直接计算w，而是借助一个中间表示量

线性回归模型的对偶表示 2

- 把w的中间表示带入到目标函数J(w)中, 得到:

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{a} - \mathbf{a}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}\mathbf{a}$$

  – 其中 $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}}$ 称为**Gram**矩阵

$$K_{nm} = \phi(\mathbf{x}_n)^{\mathrm{T}}\phi(\mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m)$$

  – 把**K**带入**J(w):**

$$J(\mathbf{a}) = \frac{1}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{K}\mathbf{a} - \mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{t} + \frac{\lambda}{2}\mathbf{a}^{\mathrm{T}}\mathbf{K}\mathbf{a}$$

  – 求梯度，令梯度为**0**

$$\Rightarrow \quad \mathbf{a} = (\mathbf{K} + \lambda\mathbf{I}_N)^{-1}\mathbf{t}$$

线性回归模型的对偶表示 3

- 给定输入数据x, 预测其输出

$$y(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) = \mathbf{a}^{\mathrm{T}}\boldsymbol{\Phi}\phi(\mathbf{x}) = \mathbf{k}(\mathbf{x})^{\mathrm{T}}\left(\mathbf{K} + \lambda\mathbf{I}_N\right)^{-1}\mathbf{t}$$

– 其中 $k_n(\mathbf{x}) = k(\mathbf{x}_n, \mathbf{x})$

N × N 矩阵求逆

- 对比线性回归模型的原表示(primal representation)

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j\phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\boldsymbol{\phi}(\mathbf{x})$$

$$\Rightarrow \quad y(x, \mathbf{w}) = \boldsymbol{\varphi}(x)^T \mathbf{w} = \boldsymbol{\varphi}(x)^T \left(\Phi^T\Phi + \lambda \cdot \mathrm{I}\right)^{-1} \Phi^T \mathbf{t}$$

M × M 矩阵求逆

核方法 内容提要

- 引子: 2个类别的分类问题
- 对偶表示
  - 正则化最小二乘法求解线性回归
- 核函数的构造
  - 生成法则
  - **Fisher** 核
- 高斯过程(Gaussian Processes)
  - 高斯过程 **for** 回归
  - 高斯过程 **for** 分类

核函数的构造: 显式映射法

- 显式映射法
  - 选定一个特征映射 ***φ(x)***, 然后获得对应的核函数

$$k(x, x') = \boldsymbol{\phi}(x)^{\mathrm{T}}\boldsymbol{\phi}(x') = \sum_{i=1}^{M} \phi_i(x)\phi_i(x')$$

核函数构造

- 举例: 定义核函数为内积的平方

$$k(\mathbf{x},\mathbf{z}) = \left(\mathbf{x}^{\mathrm{T}}\mathbf{z}\right)^2$$

  - 其中

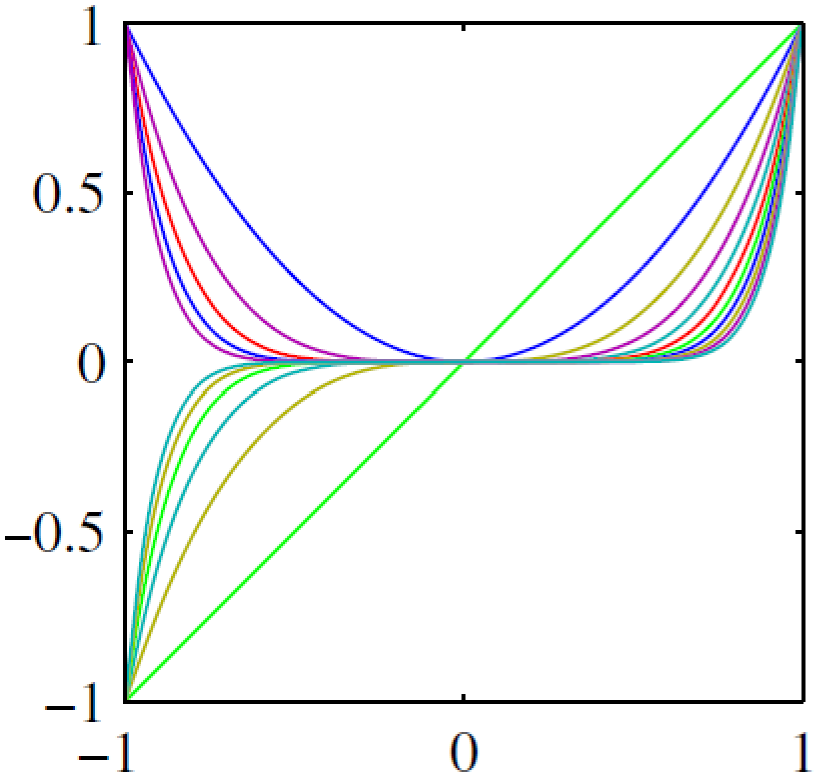
$$\begin{aligned} k(\mathbf{x},\mathbf{z}) &= \left(\mathbf{x}^{\mathrm{T}}\mathbf{z}\right)^2 = (x_1z_1 + x_2z_2)^2 \\ &= x_1^2z_1^2 + 2x_1z_1x_2z_2 + x_2^2z_2^2 \\ &= (x_1^2, \sqrt{2}x_1x_2, x_2^2)(z_1^2, \sqrt{2}z_1z_2, z_2^2)^{\mathrm{T}} \\ &= \phi(\mathbf{x})^{\mathrm{T}}\phi(\mathbf{z}). \end{aligned}$$

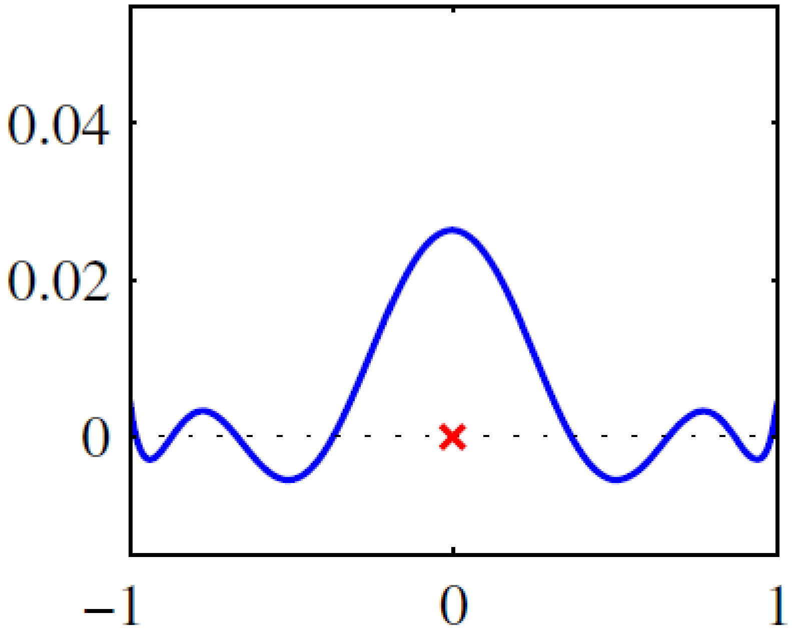
  - 所对应的特征映射为:

$$\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^{\mathrm{T}}$$

图示:特征映射与对应核函数

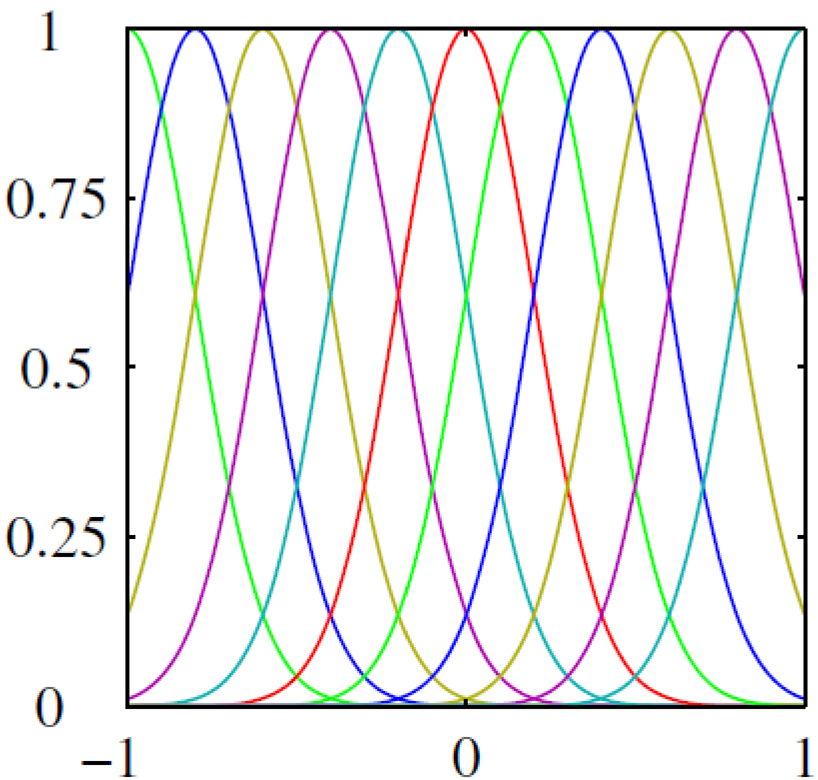
- 多项式基函数与对应的核函数

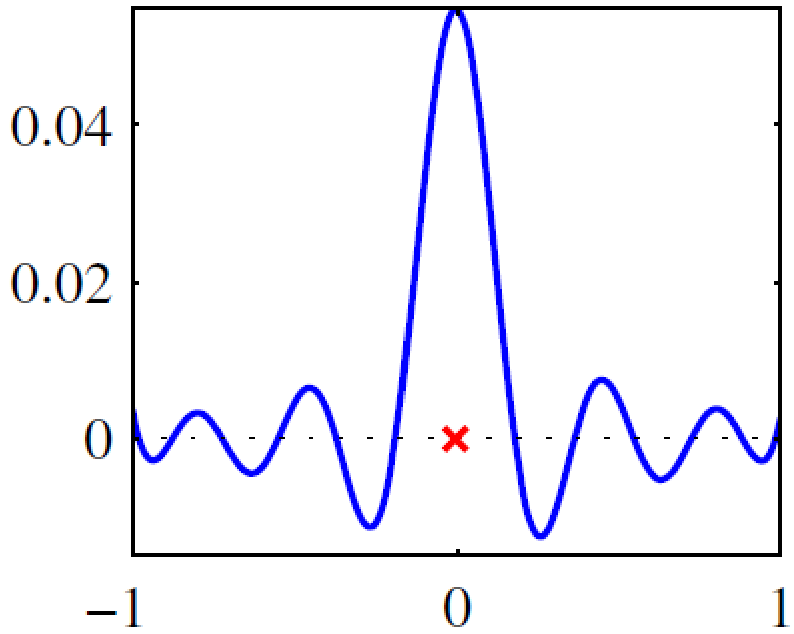




图示:特征映射与对应核函数

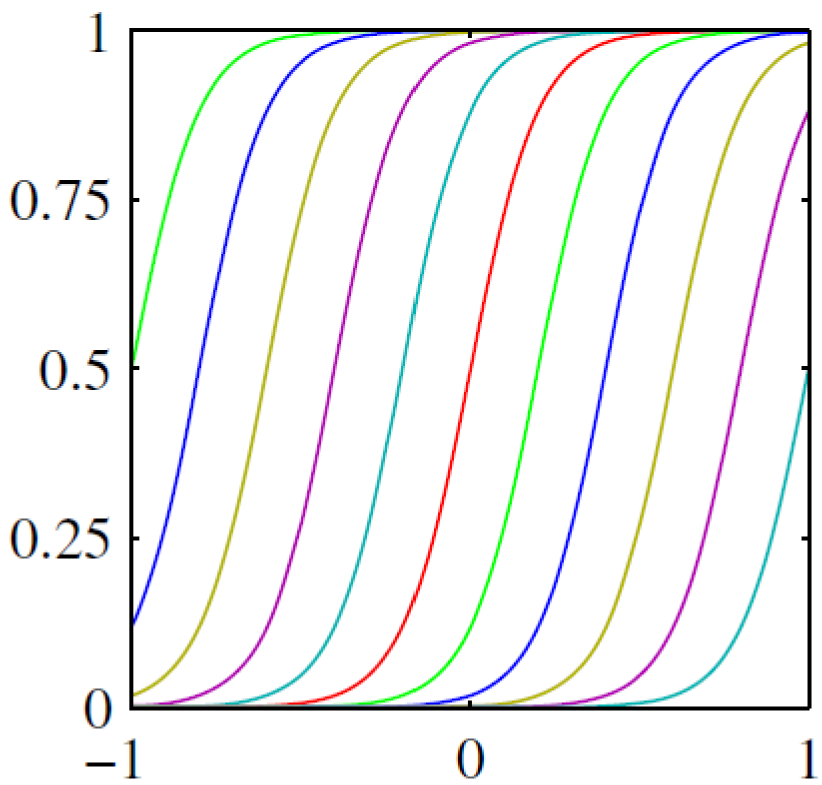
- 高斯基函数与对应的核函数

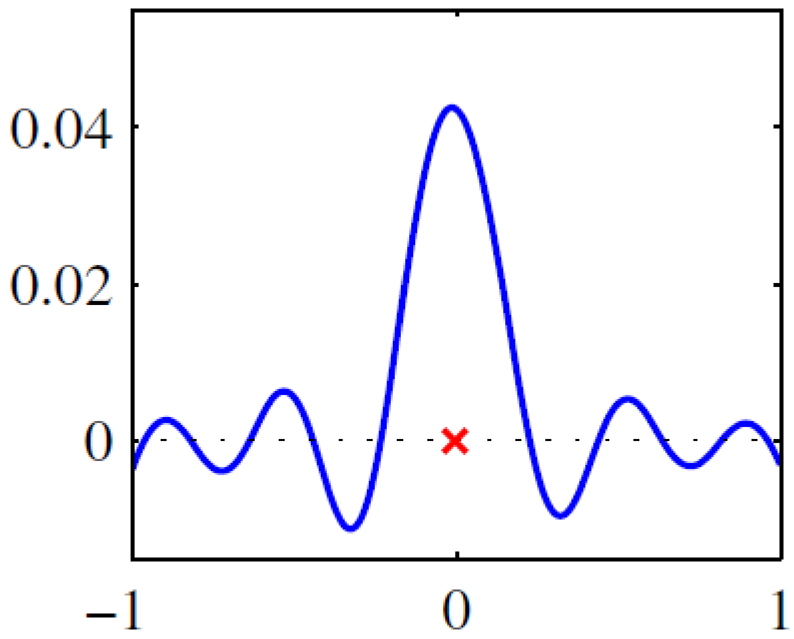




图示：特征映射与对应核函数

- Sigmoid基函数与对应的核函数





核函数的构造: 合成法

- 由给定的核函数合成新的核函数

$$
\begin{aligned}
k(\mathbf{x},\mathbf{x}') &= c k_1(\mathbf{x},\mathbf{x}') \\
k(\mathbf{x},\mathbf{x}') &= f(\mathbf{x}) k_1(\mathbf{x},\mathbf{x}') f(\mathbf{x}') \\
k(\mathbf{x},\mathbf{x}') &= q\left(k_1(\mathbf{x},\mathbf{x}')\right) \\
k(\mathbf{x},\mathbf{x}') &= \exp\left(k_1(\mathbf{x},\mathbf{x}')\right) \\
k(\mathbf{x},\mathbf{x}') &= k_1(\mathbf{x},\mathbf{x}') + k_2(\mathbf{x},\mathbf{x}') \\
k(\mathbf{x},\mathbf{x}') &= k_1(\mathbf{x},\mathbf{x}') k_2(\mathbf{x},\mathbf{x}') \\
k(\mathbf{x},\mathbf{x}') &= k_3\left(\boldsymbol{\phi}(\mathbf{x}),\boldsymbol{\phi}(\mathbf{x}')\right) \\
k(\mathbf{x},\mathbf{x}') &= \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}' \\
k(\mathbf{x},\mathbf{x}') &= k_a(\mathbf{x}_a,\mathbf{x}'_a) + k_b(\mathbf{x}_b,\mathbf{x}'_b) \\
k(\mathbf{x},\mathbf{x}') &= k_a(\mathbf{x}_a,\mathbf{x}'_a) k_b(\mathbf{x}_b,\mathbf{x}'_b)
\end{aligned}
$$

常用的核函数

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

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

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

  - 包含全部**M**阶单项式

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^T\boldsymbol{x}')^M$$

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

常用的核函数

- 高斯核(Gaussian kernel)

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

Note: can substitute $\boldsymbol{x}^T \boldsymbol{x}'$ with a nonlinear kernel $\kappa(\boldsymbol{x}, \boldsymbol{x}')$

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

- Sigmoid kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \tanh(a\boldsymbol{x}^T \boldsymbol{x}' + b)$$

- 定义在非向量类型数据的Kernel:

$$k(A_1, A_2) = 2^{|A_1 \cap A_2|}$$

概率生成模型

- 基于生成模型的核

$$
\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{x}') &= p(\boldsymbol{x})p(\boldsymbol{x}') \\
k(\boldsymbol{x}, \boldsymbol{x}') &= \sum_i p(\boldsymbol{x}|i)p(\boldsymbol{x}'|i)p(i) \\
k(\boldsymbol{x}, \boldsymbol{x}') &= \int p(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{x}'|\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z}
\end{aligned}
$$

- 基于隐马尔科夫模型(HMM)的核

$$
k(\mathbf{X}, \mathbf{X}') = \sum_{\mathbf{Z}} p(\mathbf{X}|\mathbf{Z})p(\mathbf{X}'|\mathbf{Z})p(\mathbf{Z})
$$

$\mathbf{X} = \{x_1, ..., x_L\}$ - observations
$\mathbf{Z} = \{z_1, ..., z_L\}$ - hidden states

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

$$\mathbf{F} \simeq \frac{1}{N}\sum_{n=1}^{N} g(x_n,\theta)g(x_n,\theta)^T$$

径向基函数(RBF)网络

- 径向基函数(Radial Basis Function)
  - 基函数定义为距离的函数 $\phi_j(\mathbf{x}) = h(\|\mathbf{x} - \boldsymbol{\mu}_j\|)$
- 插值函数 $f(\mathbf{x}) = \sum_{n=1}^{N} w_n h(\|\mathbf{x} - \mathbf{x}_n\|)$
- 插值问题 $E = \frac{1}{2}\sum_{n=1}^{N}\int \{y(\mathbf{x}_n + \boldsymbol{\xi}) - t_n\}^2 \nu(\boldsymbol{\xi})\,\mathrm{d}\boldsymbol{\xi}.$

$$y(\mathbf{x}_n) = \sum_{n=1}^{N} t_n h(\mathbf{x} - \mathbf{x}_n)$$

- 其中基函数为 $h(\mathbf{x} - \mathbf{x}_n) = \dfrac{\nu(\mathbf{x} - \mathbf{x}_n)}{\sum_{n=1}^{N}\nu(\mathbf{x} - \mathbf{x}_n)}$

Nadaraya-Watson核回归模型推导

- 核密度估计(KDE: Kernel Density Estimation)

$$p(\mathbf{x}, t) = \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x} - \mathbf{x}_n, t - t_n)$$

- 回归函数 y(**x**)

$$\begin{aligned} y(\mathbf{x}) &= \mathbb{E}[t|\mathbf{x}] = \int_{-\infty}^{\infty} t p(t|\mathbf{x}) \, \mathrm{d}t = \frac{\int t p(\mathbf{x}, t) \, \mathrm{d}t}{\int p(\mathbf{x}, t) \, \mathrm{d}t} \\ &= \frac{\sum_n \int t f(\mathbf{x} - \mathbf{x}_n, t - t_n) \, \mathrm{d}t}{\sum_m \int f(\mathbf{x} - \mathbf{x}_m, t - t_m) \, \mathrm{d}t} \end{aligned}$$

其中使用到 $$\int_{-\infty}^{\infty} f(\mathbf{x}, t) t \, \mathrm{d}t = 0$$

Nadaraya-Watson核回归模型

- 通过使用变量代换技术：

$$y(\mathbf{x}) = \frac{\sum_n g(\mathbf{x}-\mathbf{x}_n)t_n}{\sum_m g(\mathbf{x}-\mathbf{x}_m)} = \frac{\sum_n \int t f(\mathbf{x}-\mathbf{x}_n, t-t_n)\,\mathrm{d}t}{\sum_m \int f(\mathbf{x}-\mathbf{x}_m, t-t_m)\,\mathrm{d}t}$$

$$= \boxed{\sum_n k(\mathbf{x},\mathbf{x}_n)t_n}$$

➢ Nadaraya-Watson 核回归

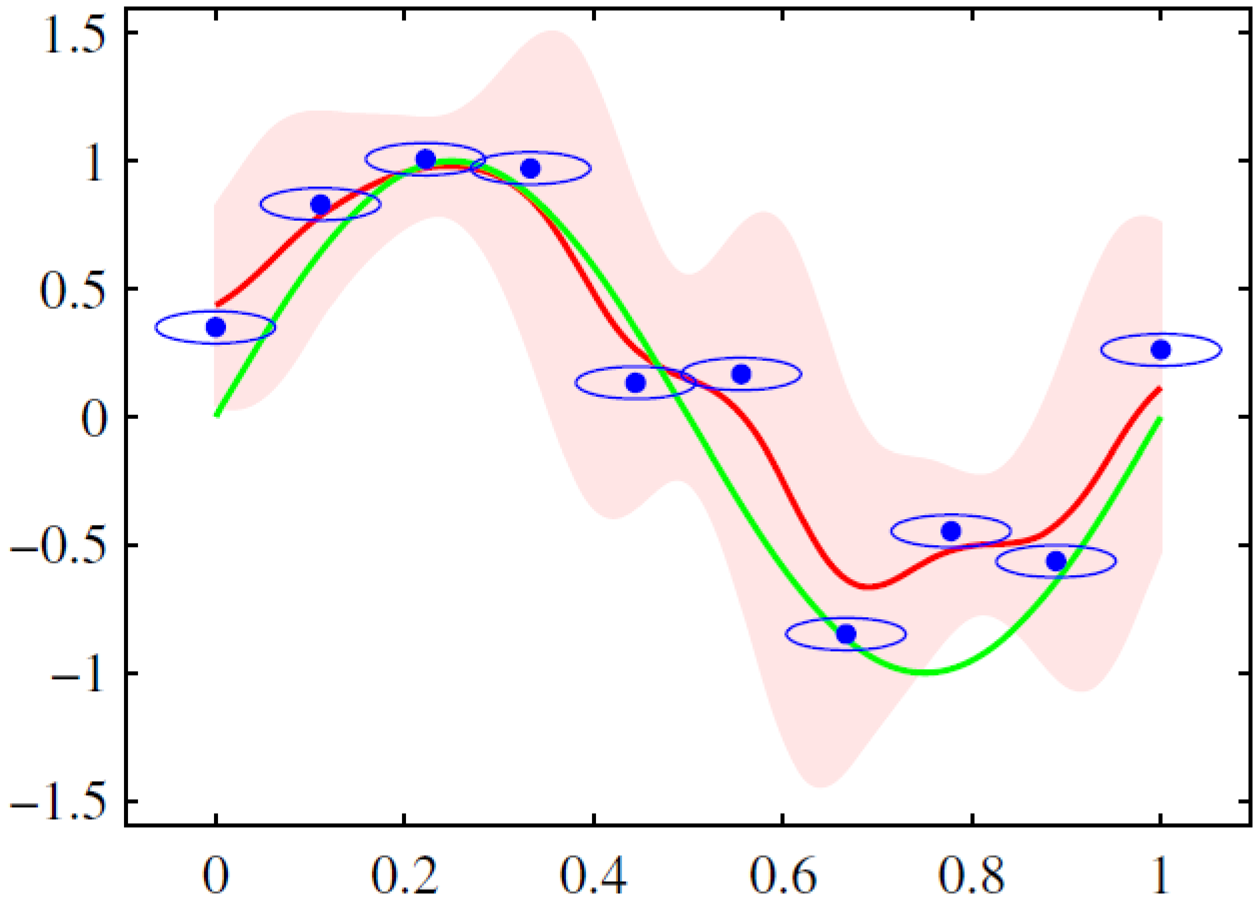
– 其中核函数定义为：

$$k(\mathbf{x},\mathbf{x}_n) = \frac{g(\mathbf{x}-\mathbf{x}_n)}{\sum_m g(\mathbf{x}-\mathbf{x}_m)}$$

$$g(\mathbf{x}) = \int_{-\infty}^{\infty} f(\mathbf{x},t)\,\mathrm{d}t$$

示例: Nadaraya-Watson核回归模型

- 红色曲线为回归模型的结果(绿色曲线为理论曲线)



核方法 内容提要

- 引子: 2个类别的分类问题
- 对偶表示
  - 正则化最小二乘法求解线性回归
- 核函数的构造
  - 生成法则
  - **Fisher** 核
- 高斯过程(Gaussian Processes)
  - 高斯过程 **for** 回归
  - 高斯过程 **for** 分类

高斯过程的基本思路

- 类似于使用一组固定基函数的线性回归

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

- 把核方法应用于概率鉴别模型中——高斯过程，与使用固定基函数的线性回归相比，其区别在于:
  - 我们通过引入定义在函数上的概率分布而使用无限多基函数
  - 在实际问题上,我们只考虑在训练和测试数据上的函数值

$$K_{mn} = k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha}\phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m)$$

回顾线性回归

- 考虑由M个固定基函数定义的回归模型

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$$

- 对权值向量w使用高斯分布先验

$$p(\mathbf{w}) = N\left(\mathbf{w} \mid \mathbf{0}, \alpha^{-1}\mathbf{I}\right)$$

  – 其中参数 α 为超参数

回顾线性回归

- 在N个训练数据点上计算回归函数y(x), 我们获得一个联合分布: $\mathbf{y} = \mathbf{\Phi w}$

  其中 $y_n = \mathbf{w}^T \phi(\mathbf{x}_n) \propto \mathcal{N}$ $\qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix},$

  – 矩阵 $\mathbf{\Phi}$ 的元素定义为: $\mathbf{\Phi}_{nk} = \phi_k(\mathbf{x}_n)$

  – 则向量 **y** 也是高斯分布，其参数为 $\mathbb{E}[\mathbf{y}] = \mathbf{\Phi}\mathbb{E}[\mathbf{w}] = 0$

  $$\operatorname{cov}[\mathbf{y}] = \mathbb{E}\left[\mathbf{y}\mathbf{y}^{\mathrm{T}}\right] = \mathbf{\Phi}\mathbb{E}\left[\mathbf{w}\mathbf{w}^{\mathrm{T}}\right]\mathbf{\Phi}^{\mathrm{T}} = \frac{1}{\alpha}\mathbf{\Phi}\mathbf{\Phi}^{\mathrm{T}} = \mathbf{K}$$

  - 其中K为Gram矩阵

  $$K_{mn} = k(\mathbf{x}_n, \mathbf{x}_m) = \frac{1}{\alpha}\phi(\mathbf{x}_n)^T \phi(\mathbf{x}_m)$$

高斯过程

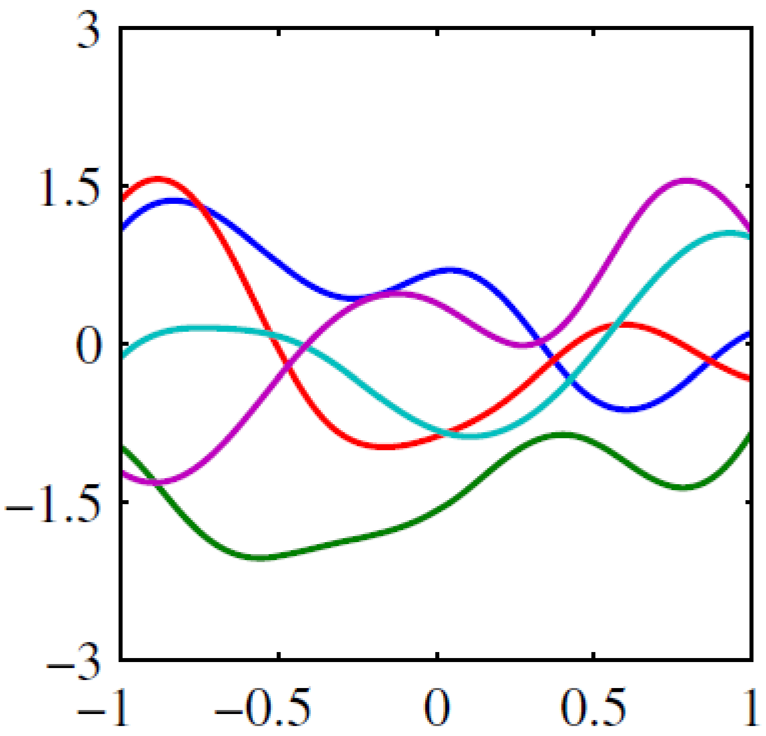
- 高斯过程定义为:
  - 一个定义在函数**y(x)**上的概率分布，其中函数**y(x)**在数据点$\mathbf{x}_1,\ldots\mathbf{x}_N$的任意子集上的取值的联合分布也为高斯分布
- 关键点:
  - 联合分布使用二阶统计量**(**均值和协方差**)**定义
  - 通常，均值为**0**，我们只需考虑协方差，即核函数 $\mathbb{E}\left[y(\mathbf{x}_n)y(\mathbf{x}_m)\right]=k(\mathbf{x}_n,\mathbf{x}_m)$

$$\mathrm{cov}[\mathbf{y}]=\alpha^{-1}\mathbf{\Phi}\mathbf{\Phi}^T=\mathbf{K}$$

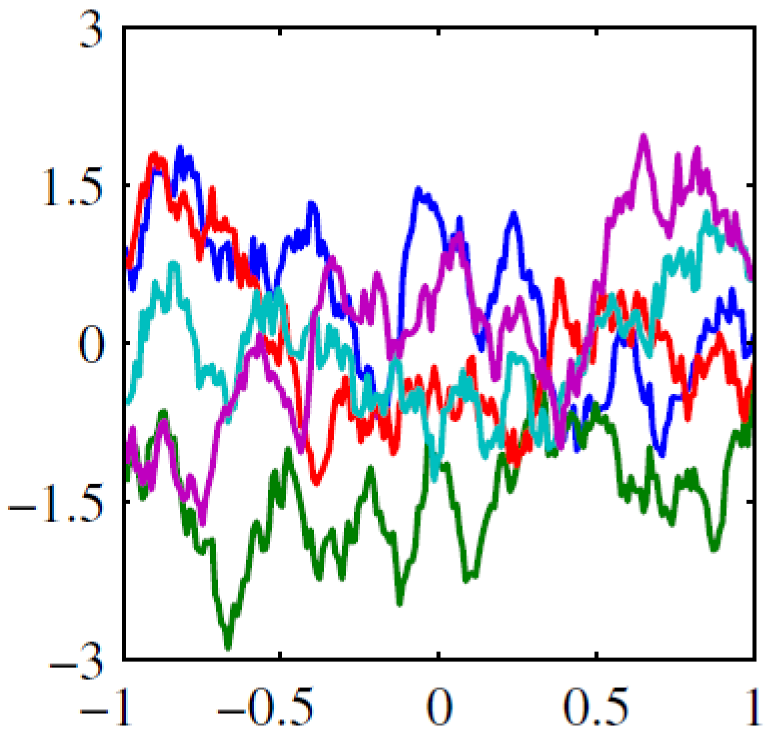
    - 因此，与其去选择一组基函数，我们直接选择一个核函数

示例：不同的核函数

- 直接定义核函数



'Gaussian' kernel



exponential kernel

核方法 内容提要

- 引子: 2个类别的分类问题
- 对偶表示
  - 正则化最小二乘法求解线性回归
- 核函数的构造
  - 生成法则
  - **Fisher** 核
- 高斯过程(Gaussian Processes)
  - 高斯过程 **for** 回归
  - 高斯过程 **for** 分类

高斯过程回归(GPR)

- 使用GP解决回归任务，需要考虑噪声

$$t_n = y_n + \epsilon_n \quad \text{with} \quad y_n = y(\mathbf{x_n})$$

- 假设噪声服从高斯分布 $p(t_n|y_n) = \mathcal{N}(t_n|y_n, \beta^{-1})$

- 令 $\mathbf{t} = (t_1, \ldots, t_n)^T$， $\mathbf{y} = (y_1, \ldots, y_n)^T$ 则联合分布为:

$$p(\mathbf{t}|\mathbf{y}) = \mathcal{N}(\mathbf{t}|\mathbf{y}, \beta^{-1}\mathbf{I}_N)$$

高斯过程回归(GPR)

- 由高斯过程的定义，我们得到

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K})$$

  – 如果数据点是相似的，则具有强相关性

- 对于边缘分布p(t), 我们有对y进行积分

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})\mathrm{d}\mathbf{y} = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C})$$

  – 其中 $\mathbf{C} = \mathbf{K} + \beta^{-1}\mathbf{I}$

高斯过程回归(GPR)

- 广泛用于GPR的核函数

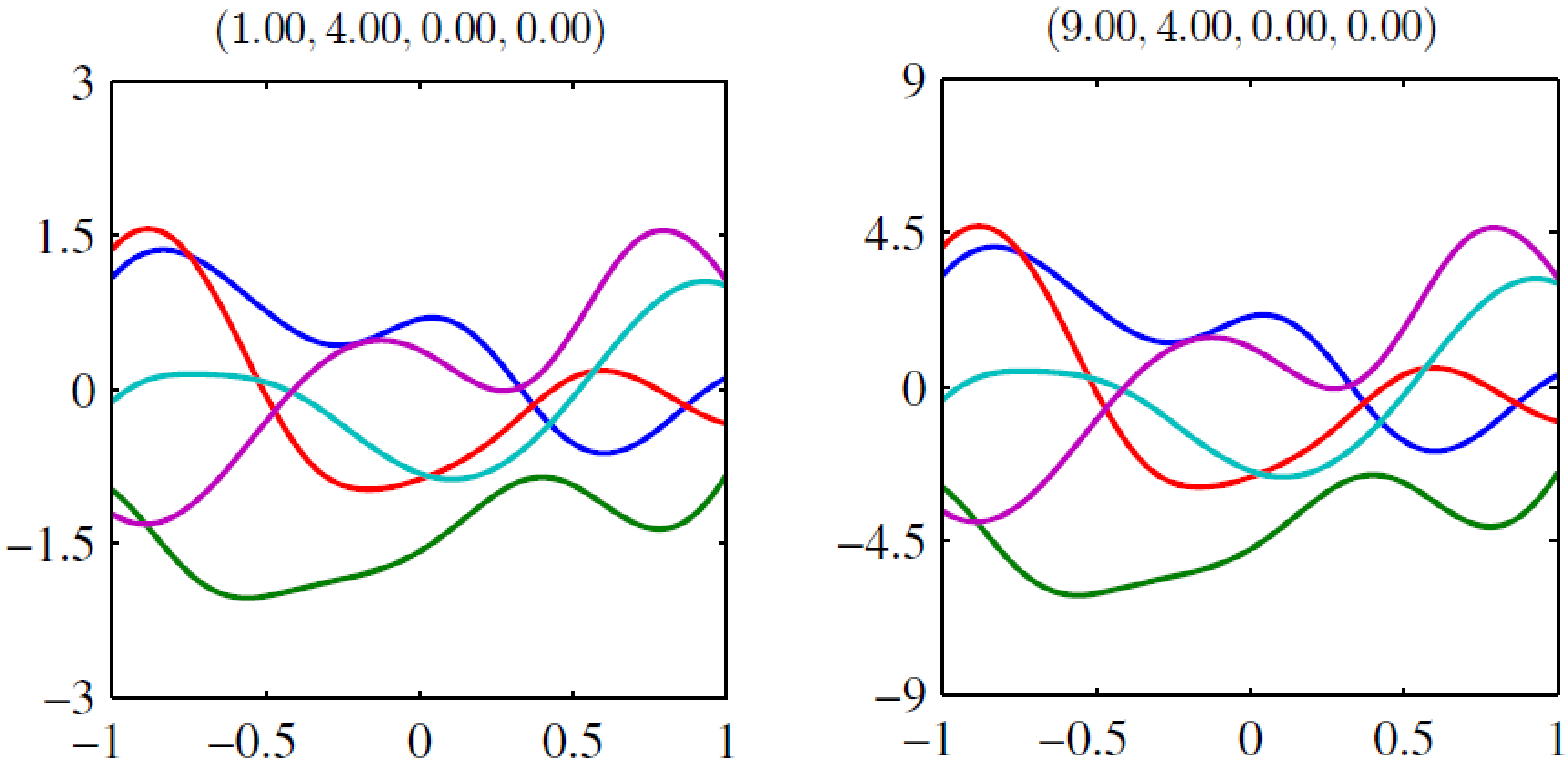
$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right\} + \theta_2 + \theta_3 \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m$$

  - 其中 $(\theta_0, \theta_1, \theta_2, \theta_3)$ 为4个超参数

示例: 高斯过程回归(GPR)

- 分别使用两组超参数 $(\theta_0, \theta_1, \theta_2, \theta_3)$

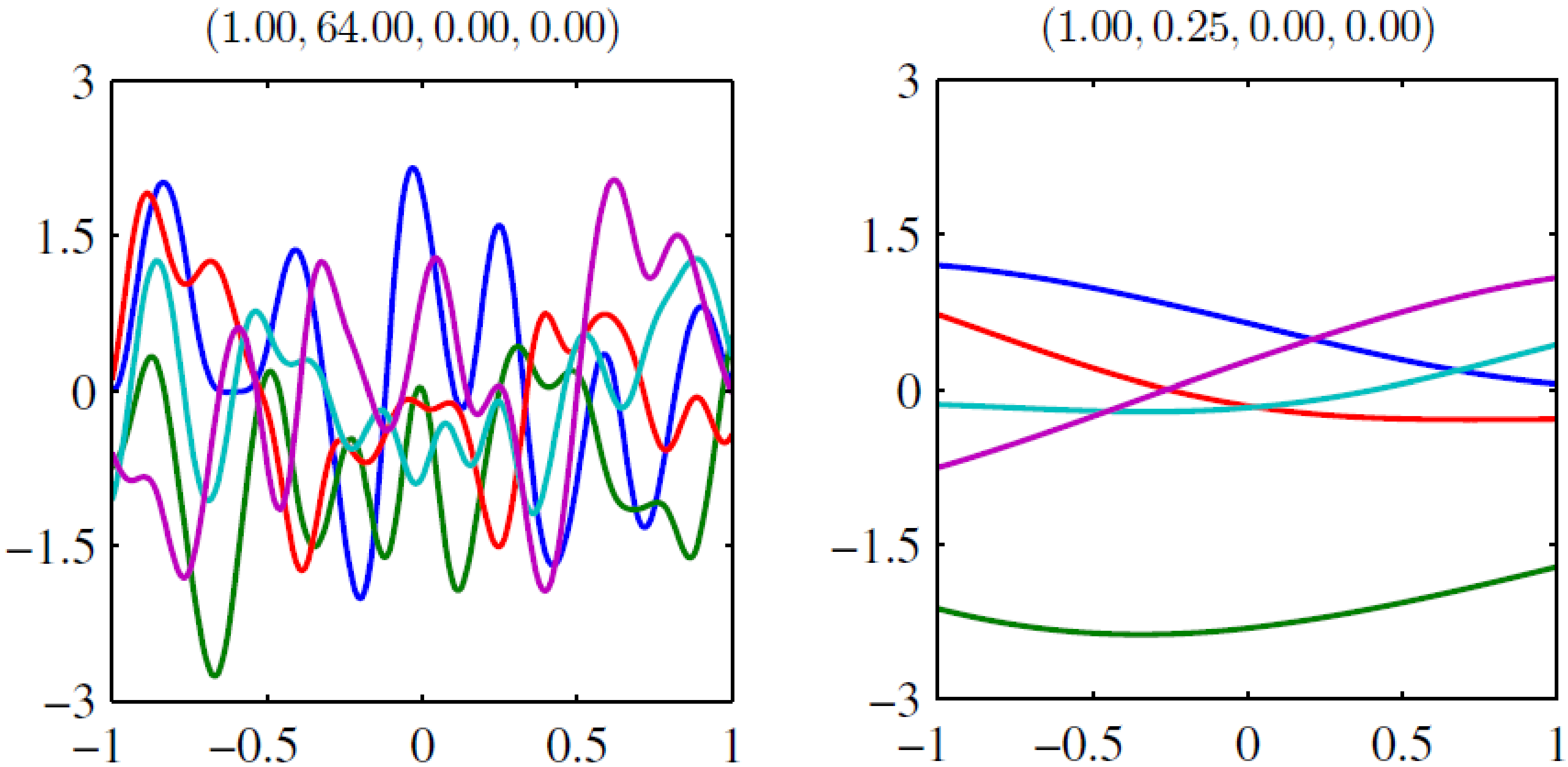
$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right\} + \theta_2 + \theta_3 \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m$$



示例: 高斯过程回归(GPR)

- 分别使用两组超参数 $(\theta_0,\theta_1,\theta_2,\theta_3)$

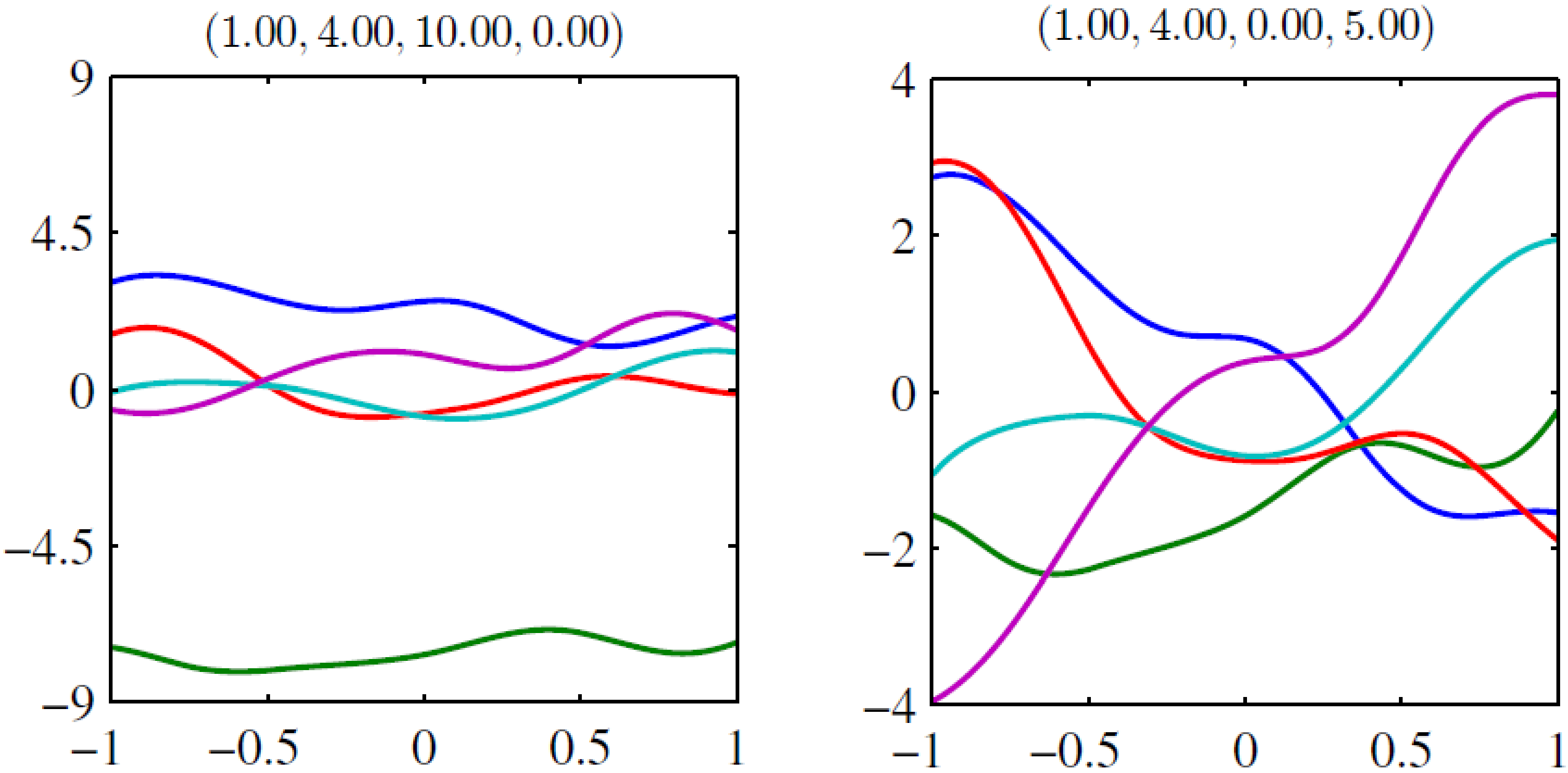
$$k(\mathbf{x}_n,\mathbf{x}_m)=\theta_0\exp\left\{-\frac{\theta_1}{2}\|\mathbf{x}_n-\mathbf{x}_m\|^2\right\}+\theta_2+\theta_3\mathbf{x}_n^{\mathrm{T}}\mathbf{x}_m$$



示例: 高斯过程回归(GPR)

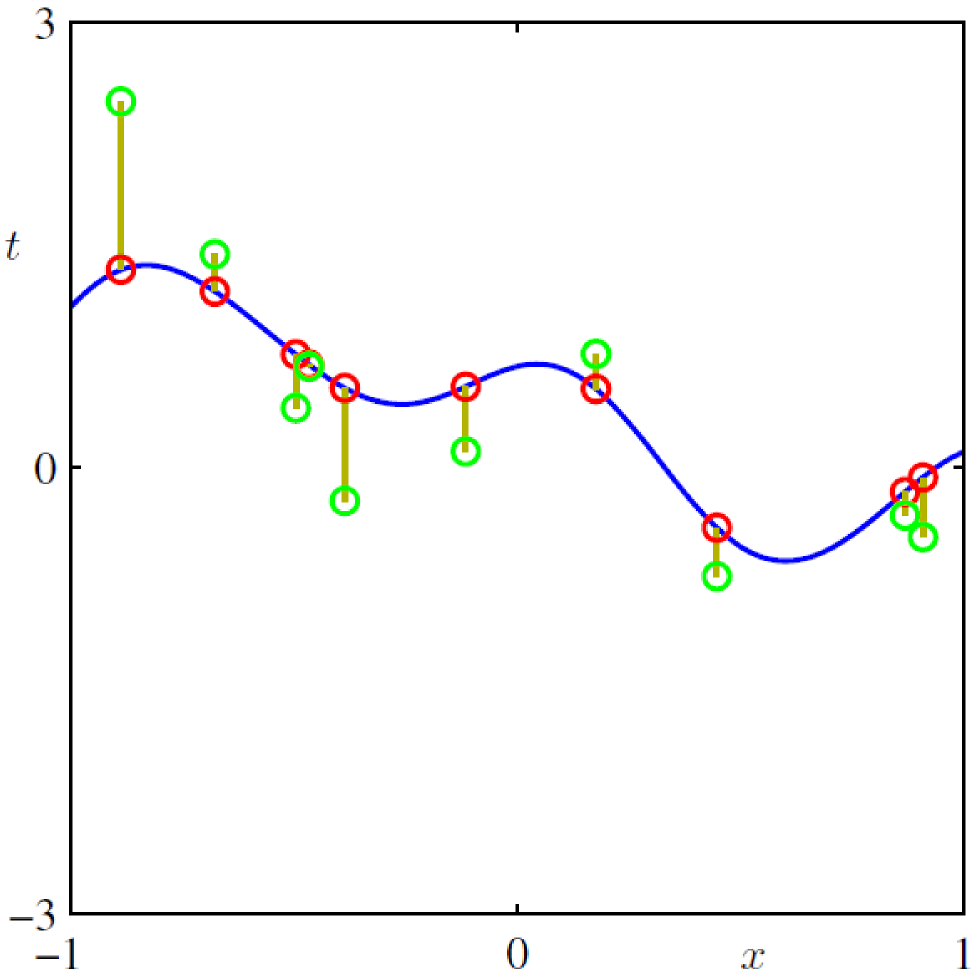
- 分别使用两组超参数 $(\theta_0, \theta_1, \theta_2, \theta_3)$

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|\mathbf{x}_n - \mathbf{x}_m\|^2\right\} + \theta_2 + \theta_3 \mathbf{x}_n^{\mathrm{T}} \mathbf{x}_m$$



示例: 高斯过程回归(GPR)

•



基于高斯过程回归(GPR)的预测

- 假设已经得到数据集上的联合概率分布模型

$$p(\mathbf{t}) = \int p(\mathbf{t}|\mathbf{y})p(\mathbf{y})\mathrm{d}\mathbf{y} = \mathcal{N}(\mathbf{t}|\mathbf{0}, \mathbf{C})$$

- 对于新数据点x_{N+1}，我们要计算其预测性分布

$$p(t_{N+1}|\mathbf{t})$$

- 基于数据点 $\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{x}_{N+1}$ 上的联合分布，我们可以借助均值和方差获得 $p(t_{N+1}|\mathbf{t})$

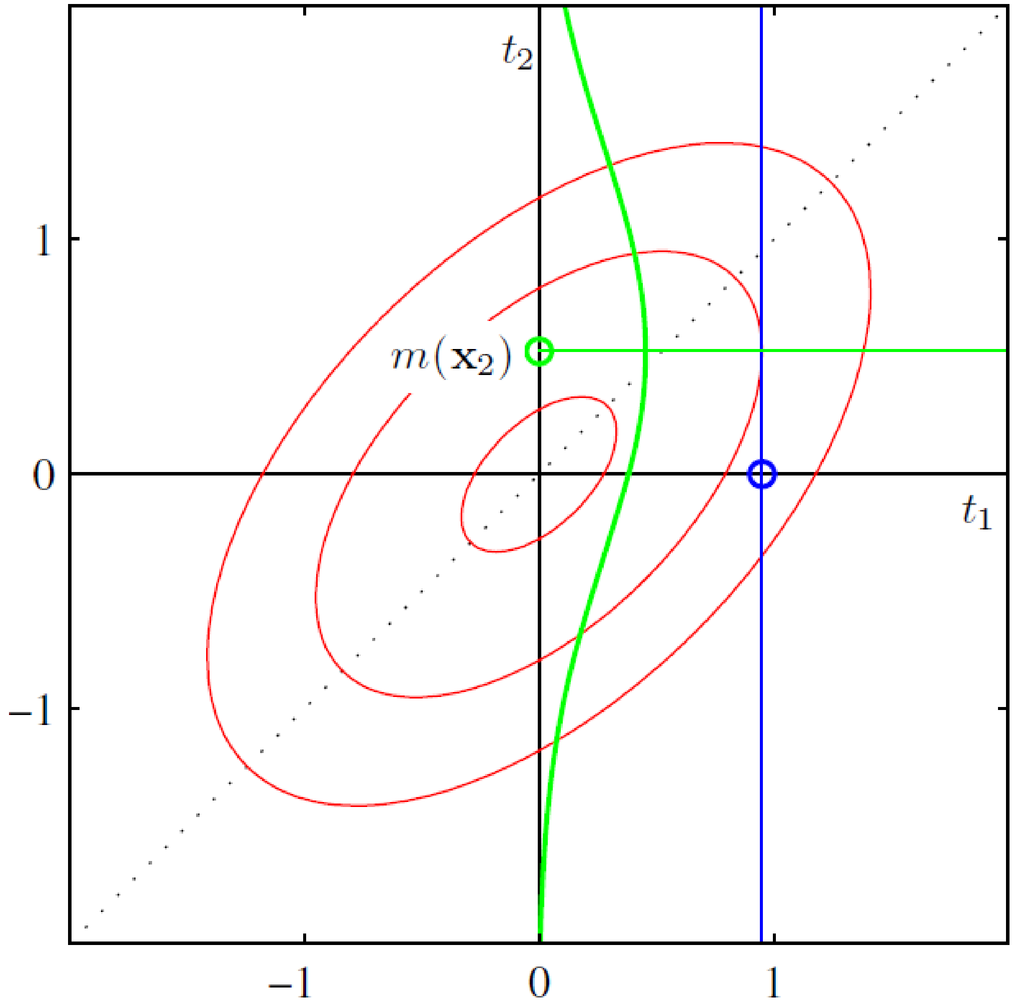
$$m(\mathbf{x}_{N+1}) = \mathbf{k}^T\mathbf{C}^{-1}\mathbf{t} \qquad \sigma^2(\mathbf{x}_{N+1}) = c - \mathbf{k}^T\mathbf{C}^{-1}\mathbf{k}$$

  - 其中 $\mathbf{k} = (k(\mathbf{x}_1, \mathbf{x}_{N+1}), \ldots, k(\mathbf{x}_N, \mathbf{x}_{N+1}))^T$

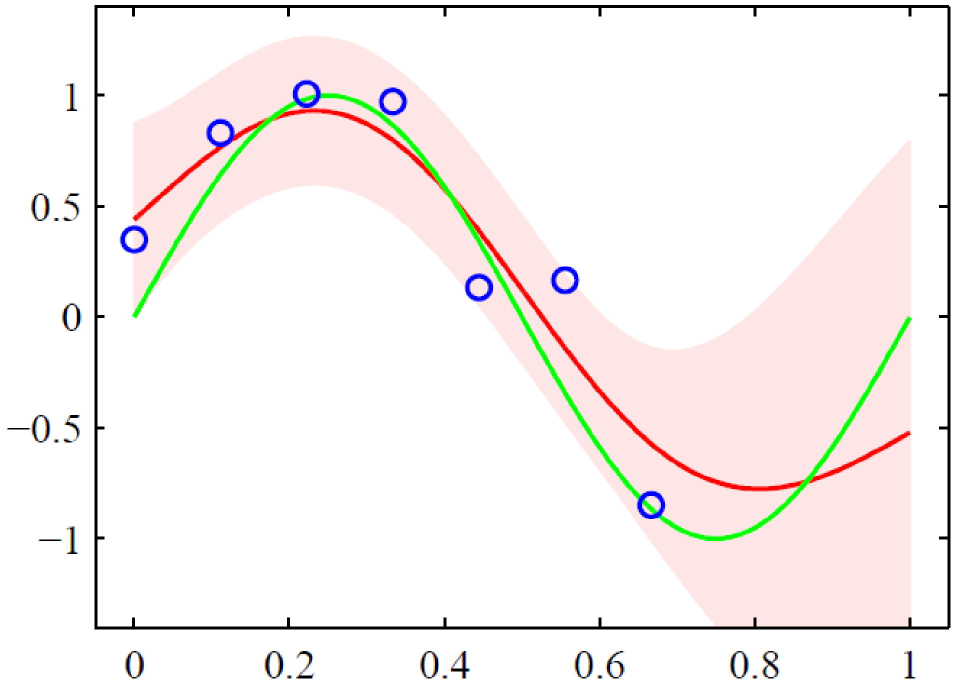
$$c = k(\mathbf{x}_{N+1}, \mathbf{x}_{N+1}) + \beta^{-1}$$

示例: 基于GPR的预测

•



示例: 基于GPR的预测

- 

Green curve: original sinusoidal function; blue points: sampled training data points with additional noise; read line: mean estimate; shaded regions: $+/-\ 2\sigma$

确定GPR模型中的超参数

- 高斯过程模型部分地依赖于协方差函数的选择
  - 在实际中，一般使用一个函数的参数族，然后基于数据去推理合理的参数

- 最简单的方法:
  - 最大似然准则
    - 寻找最大化对数似然函数**lnp(t|Θ)** 的参数Θ
  - 困难
    - **lnp(t|Θ)**一般是**non-convex**且存在多个最大值

确定GPR模型中的超参数

- 解决策略:
  - 引入一个先验分布p(Θ)，然后最大化对数后验概率 ln p(t| Θ) + ln p(Θ)

$$\ln p(\mathbf{t}|\boldsymbol{\theta}) = -\frac{1}{2}\ln|\mathbf{C}_N| - \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{t} - \frac{N}{2}\ln(2\pi)$$

$$\frac{\partial}{\partial\theta_i}\ln p(\mathbf{t}|\boldsymbol{\theta}) = -\frac{1}{2}\mathrm{Tr}\left(\mathbf{C}_N^{-1}\frac{\partial\mathbf{C}_N}{\partial\theta_i}\right) + \frac{1}{2}\mathbf{t}^{\mathrm{T}}\mathbf{C}_N^{-1}\frac{\partial\mathbf{C}_N}{\partial\theta_i}\mathbf{C}_N^{-1}\mathbf{t}$$

自动的相关特征检测

- 通过在每个维度上引入加权超参数 η_i ，可以实现自动检测相关的特征

$$k(\mathbf{x}_n, \mathbf{x}_m) = \theta_0 \exp\left(-\frac{1}{2}\sum_{i=1}^{D} \eta_i (x_{ni} - x_{mi})^2\right) + \theta_2 + \theta_3 \mathbf{x}_n^T \mathbf{x}_m$$

  – 权值可以基于最大似然准则学习得到

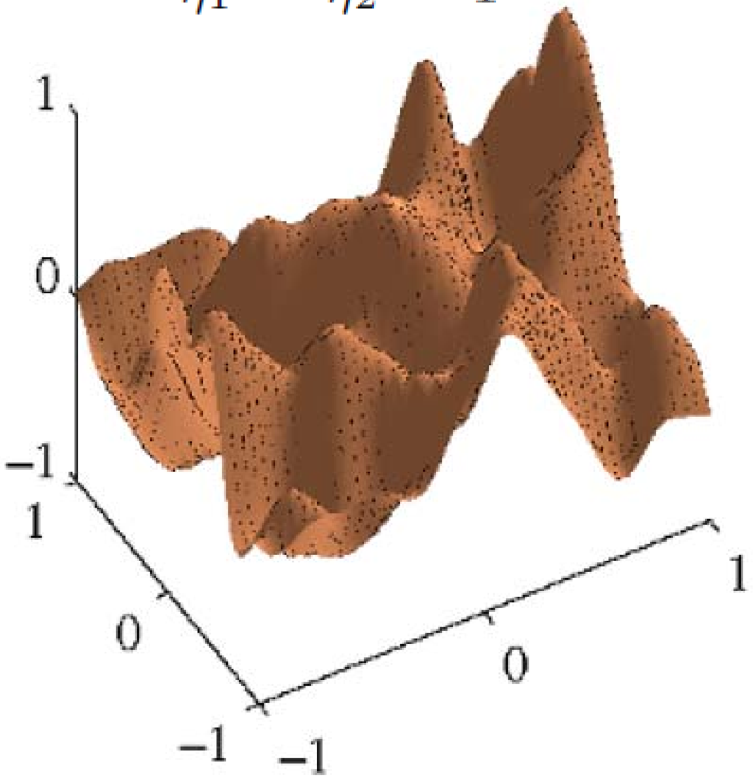
  – 不相关的特征对应于小的权值
    - 可以把这些维度丢弃

示例:自动的相关特征检测

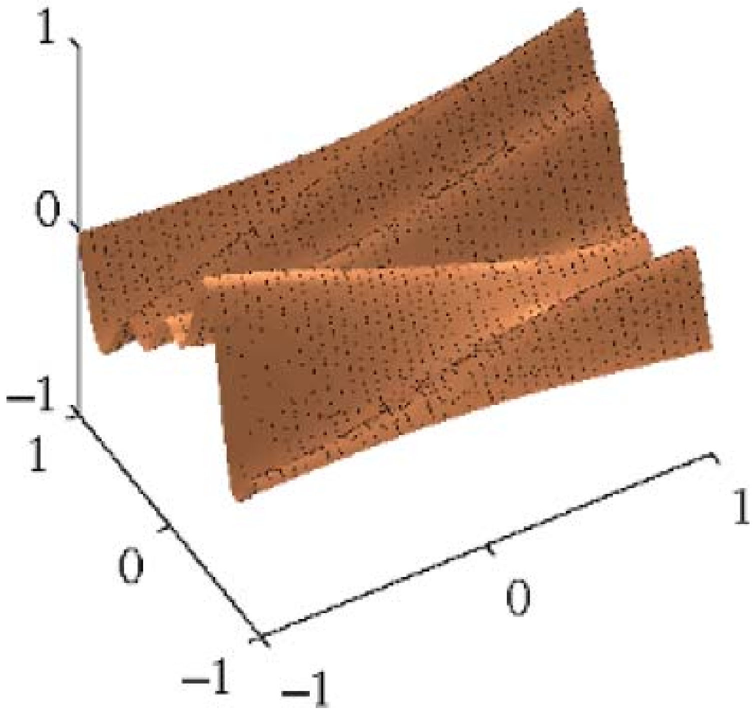
- 不同的权值

$$k(\mathbf{x}, \mathbf{x}') = \theta_0 \exp\left\{ -\frac{1}{2} \sum_{i=1}^{2} \eta_i (x_i - x_i')^2 \right\}$$

$\eta_1 = \eta_2 = 1$



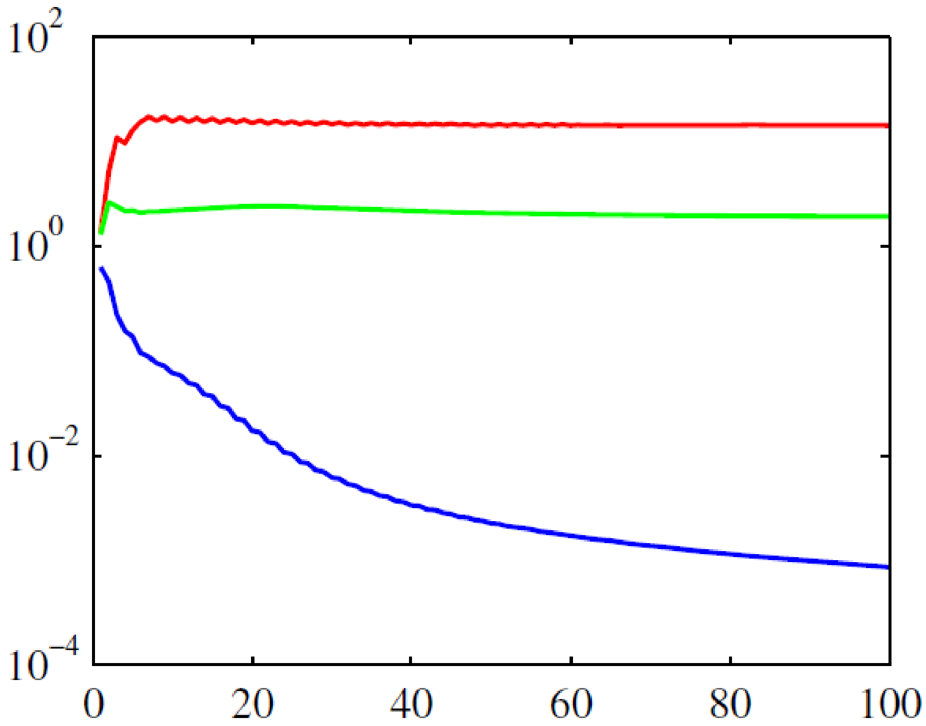
$\eta_1 = 1, \eta_2 = 0.01$



示例:自动的相关特征检测

- 小的权值对应于不相关的维度，可以被确定并丢掉

$$k(\mathbf{x}, \mathbf{x}') = \theta_0 \exp\left\{-\frac{1}{2}\sum_{i=1}^{2} \eta_i (x_i - x_i')^2\right\}$$

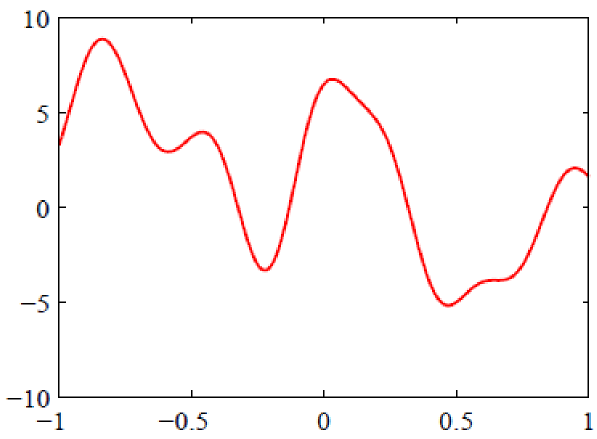


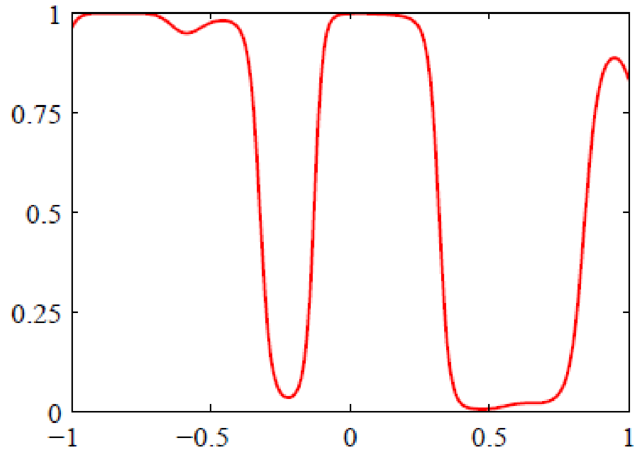
核方法 内容提要

- 引子: 2个类别的分类问题
- 对偶表示
  - 正则化最小二乘法求解线性回归
- 核函数的构造
  - 生成法则
  - **Fisher** 核
- 高斯过程(Gaussian Processes)
  - 高斯过程 **for** 回归
  - 高斯过程 **for** 分类

高斯过程 for 分类

- 目标:
  – 建模新输入数据的对应输出的后验概率
- 问题:
  – 把输入数据映射到一个区间 [0，1]
- 解决方案：
  – 使用高斯过程和非线性激活函数





高斯过程 for 分类

- 考虑一个2类分类问题
  - 两个类别对应于目标输出为 **0**和**1**
- 用于分类的高斯过程模型:
  - 在函数**a(x)**上定义一个高斯过程，并使用**Logistic**函数把函数**a(x)**的输出转换为**[0,1]**区间内的值

$$y = \sigma(a(\mathbf{x}))$$

  - 我们需要计算条件分布

$$p(t_{N+1} = 1|\mathbf{t}) = \int p(t_{N+1} = 1|a_{N+1})p(a_{N+1}|\mathbf{t})\mathrm{d}a_{N+1}$$

$$= \int \sigma(a_{N+1})p(a_{N+1}|\mathbf{t})\mathrm{d}a_{N+1}.$$

  - 困难:
    - 积分不能解析地处理，需要使用近似或数值计算技术

近似计算技术

- 目标:
  - 计算 $p(t_{N+1}|\mathbf{t})$
  - 其中

$$\mathbf{t} = (t_1, \dots, t_N)^{\mathrm{T}}$$

$$p(t|a) = \sigma(a)^t (1 - \sigma(a))^{1-t}$$

$$y = \sigma(a)$$

$$p(\mathbf{a}_{N+1}) = \mathcal{N}(\mathbf{a}_{N+1}|\mathbf{0}, \mathbf{C}_{N+1})$$

$$C(\mathbf{x}_n, \mathbf{x}_m) = k(\mathbf{x}_n, \mathbf{x}_m) + \nu\delta_{nm}$$

$$p(t_{N+1} = 1|\mathbf{t}_N) = \int p(t_{N+1} = 1|a_{N+1}) p(a_{N+1}|\mathbf{t}_N)\, \mathrm{d}a_{N+1}$$

- 使用**Laplace**近似

Laplace近似

- 后验概率的计算

$$p(t_{N+1}=1|\mathbf{t}_N)=\int p(t_{N+1}=1|a_{N+1})p(a_{N+1}|\mathbf{t}_N)\,\mathrm{d}a_{N+1}$$

$$\begin{aligned}
p(a_{N+1}|\mathbf{t}_N) &= \int p(a_{N+1},\mathbf{a}_N|\mathbf{t}_N)\,\mathrm{d}\mathbf{a}_N \\
&= \frac{1}{p(\mathbf{t}_N)}\int p(a_{N+1},\mathbf{a}_N)p(\mathbf{t}_N|a_{N+1},\mathbf{a}_N)\,\mathrm{d}\mathbf{a}_N \\
&= \frac{1}{p(\mathbf{t}_N)}\int p(a_{N+1}|\mathbf{a}_N)p(\mathbf{a}_N)p(\mathbf{t}_N|\mathbf{a}_N)\,\mathrm{d}\mathbf{a}_N \\
&= \int p(a_{N+1}|\mathbf{a}_N)p(\mathbf{a}_N|\mathbf{t}_N)\,\mathrm{d}\mathbf{a}_N
\end{aligned}$$

Laplace近似

- 后验概率的计算

$$p(a_{N+1}|\mathbf{a}_N) = \mathcal{N}(a_{N+1}|\mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{a}_N, c - \mathbf{k}^{\mathrm{T}}\mathbf{C}_N^{-1}\mathbf{k})$$

  - 其中先验概率$p(\mathbf{a}_N)$是一个零均值高斯过程

$$p(\mathbf{t}_N|\mathbf{a}_N) = \prod_{n=1}^{N} \sigma(a_n)^{t_n}(1-\sigma(a_n))^{1-t_n} = \prod_{n=1}^{N} e^{a_n t_n}\sigma(-a_n)$$

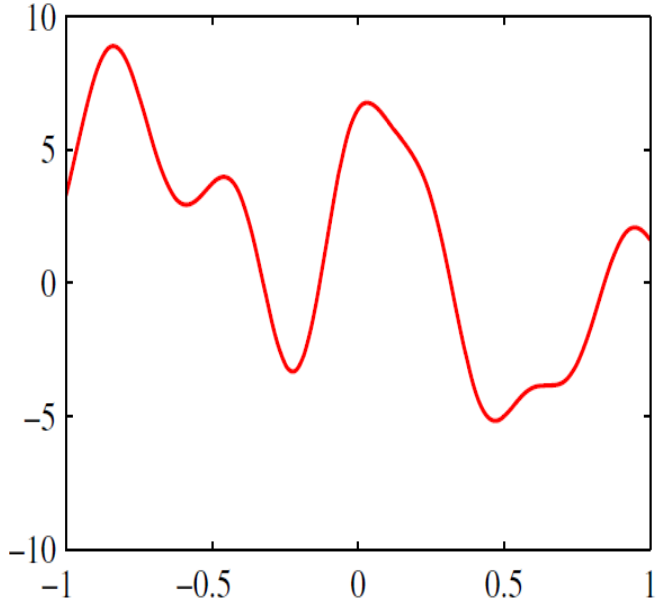
$$\Psi(\mathbf{a}_N) = \ln p(\mathbf{a}_N) + \ln p(\mathbf{t}_N|\mathbf{a}_N)$$

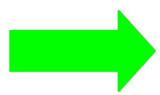
$$q(\mathbf{a}_N) = \mathcal{N}(\mathbf{a}_N|\mathbf{a}_N^\star, \mathbf{H}^{-1})$$

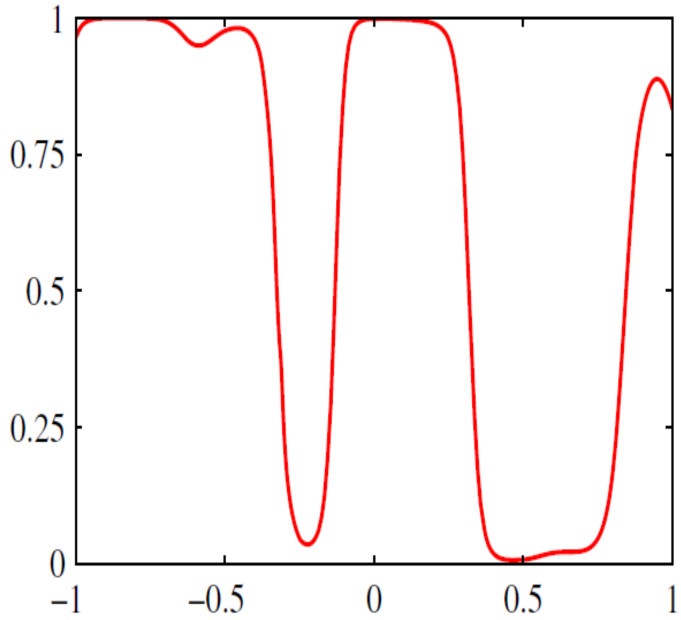
  - 还需要确定协方差函数中的参数 θ

示例: GP for 分类

•

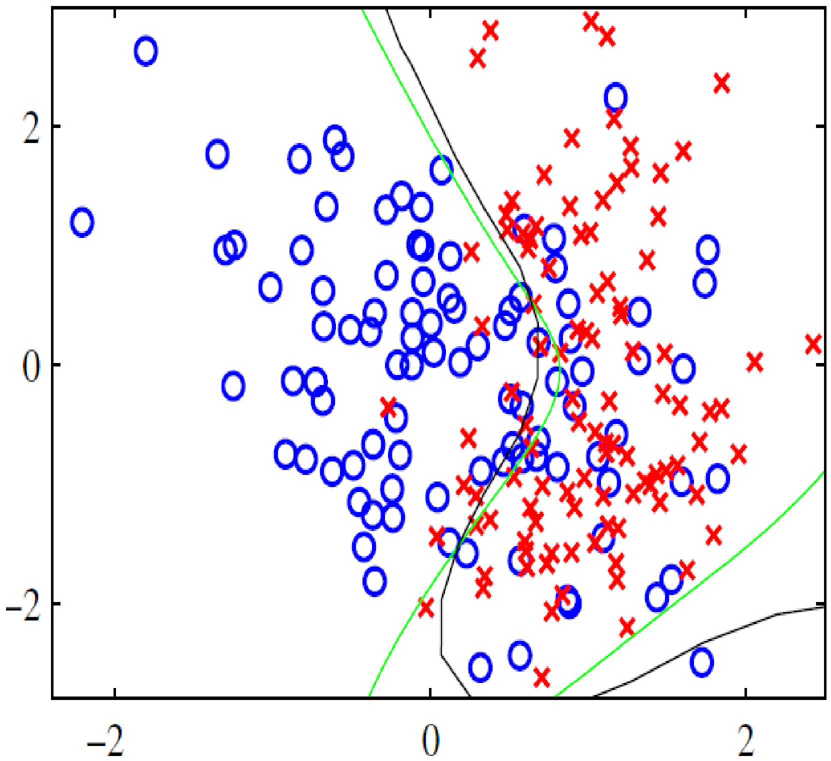


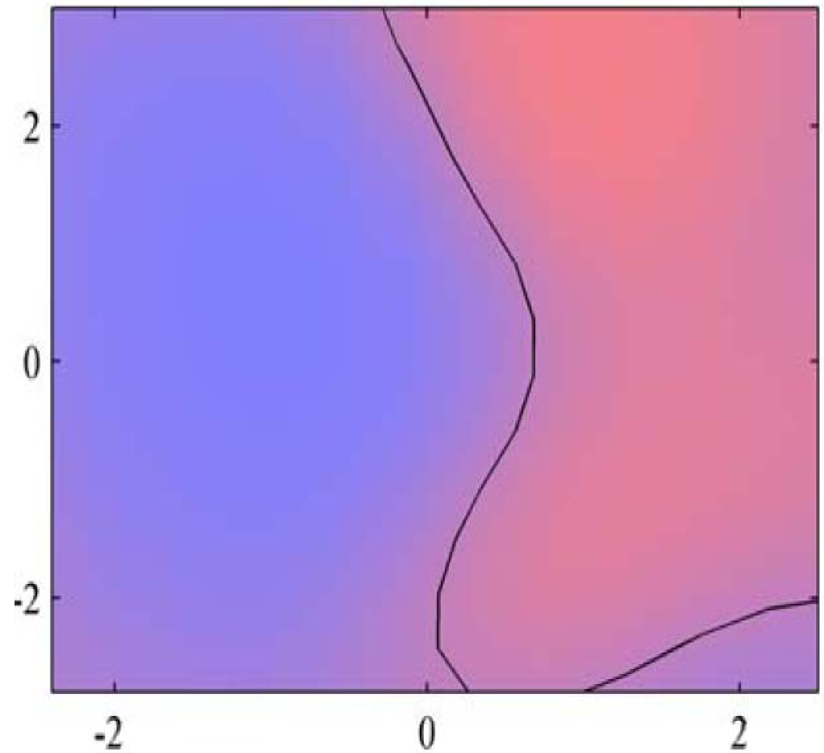




示例: GP for 分类

- 决策边界与置信度(热度图)





Q / A

- Any Questions...