今天(jīntiān)内容

- 核回归
- 核方法
  - Kernel trick
  - 正则化理论

非参数(cānshù)回归

- 参数回归（线性回归）时，假设$r(x)$ 为线性的 。当$r(x)$ 不是x的线性函数时，基于最小二乘的回归效果不佳

- 非参数回归：不对$r(x)$的形式做任何假定
  - 参考核密度估计
  - 局部加权方法：用点x附近的Y_i的加权平均表示$r(x)$
    - 权重为核函数的值，邻域由核函数的宽度控制

核回归(huíguī)：Nadaraya-Watson

- 回忆一下回归方程的定义：

$$r(x)=\mathbb{E}(Y\mid X=x)=\int yf(y\mid x)\,dy$$

$$=\frac{\int yf(x,y)\,dy}{\int f(x,y)\,dy}=\frac{\int yf(x,y)\,dy}{f(x)}$$

- 分别对$f(x), f(x,y)$用核密度估计，得到

$$\hat{r}(x)=\frac{\sum_{i=1}^{n}K_h(x,x_i)\,y_i}{\sum_{j=1}^{n}K_h(x,x_j)}$$

核回归(huíguī)：Nadaraya-Watson

- 证明：$\hat{f}(x,y)=\frac{1}{n}\sum_{i=1}^{n}K_{h_1}(x,x_i)K_{h_2}(y,y_i)$

$$\int y\hat{f}(x,y)dy=\frac{1}{n}\sum_{i=1}^{n}\int K_{h_1}(x,x_i)\,yK_{h_2}(y,y_i)\,dy$$

$$=\frac{1}{n}\sum_{i=1}^{n}K_{h_1}(x,x_i)\int\frac{y}{h_2}K\left(\frac{y-y_i}{h_2}\right)dy$$

$$=\frac{1}{n}\sum_{i=1}^{n}K_{h_1}(x,x_i)\int(sh_2+y_i)K(s)ds$$

$$=\frac{1}{n}\sum_{i=1}^{n}K_{h_1}(x,x_i)\,y_i$$

$$\boxed{\begin{aligned}&\int K(x)dx=1,\\&\int xK(x)dx=0\end{aligned}}$$

核回归(huíguī)：Nadaraya-Watson

- 证明（续）

$$r(x)=\frac{\int f(x,y)ydy}{f(x)}$$

$$\hat{r}(x)=\frac{\frac{1}{n}\sum_{i=1}^{n}K_{h_1}(x,x_i)y_i}{\frac{1}{n}\sum_{j=1}^{n}K_{h_1}(x,x_j)}=\frac{\sum_{i=1}^{n}K_{h_1}(x,x_i)y_i}{\sum_{j=1}^{n}K_{h_1}(x,x_j)}=\frac{\sum_{i=1}^{n}K_{h}(x,x_i)y_i}{\sum_{j=1}^{n}K_{h}(x,x_j)}$$

核回归(huíguī)：Nadaraya-Watson

- 这可以被看作是对y取一个加权平均，对x附近的值给予更高的权重：

$$\hat{r}(x)=\sum_{i=1}^{n} w_i(x)\,y_i$$

- 其中

$$w_i(x)=\frac{K_h(x,x_i)}{\sum_{j=1}^{n} K_h(x,x_j)}$$

核回归(huíguī)：Nadaraya-Watson

- 将核回归估计写成如下形式： $\hat{r}(x)=\dfrac{\hat{g}_h(x)}{\hat{f}_h(x)}$
- 其中 $\hat{g}_h(x)=\dfrac{1}{n}\sum_{i=1}^{n}K_h(x,x_i)y_i$, $\hat{f}_h(x)=\dfrac{1}{n}\sum_{i=1}^{n}K_h(x,x_i)$

$$
\begin{aligned}
\mathbb{E}\left(\hat{g}_h(x)\right) &= \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}K_h(x,x_i)y_i\right)\\
&= \mathbb{E}\left(K_h(x,x_i)y_i\right)\\
&= \int\int K_h(x,u)\,y\underline{f(y\mid u)f(u)}\,dy\,du\\
&= \int K_h(x,u)f(u)\underbrace{\left(\int yf(y\mid u)\,dy\right)}_{r(u)}du\\
&= \int K_h(x,u)\,g(u)\,du
\end{aligned}
$$

核回归(huíguī)： Nadaraya-Watson

- 类似核密度估计中求期望的展开，得到

$$\mathbb{E}\left(\hat{g}_h\left(x\right)\right)\approx g\left(x\right)+\frac{h^2}{2}g''\left(x\right)\int x^2K^2\left(x\right)dx$$

- 同理，

$$\mathbb{V}\left(\hat{g}_h\left(x\right)\right)\approx\frac{1}{nh}\sigma^2\left(x\right)\int K^2\left(x\right)dx$$

- 其中 $\mathbb{V}\left(\varepsilon_i\right)=\sigma^2\left(x\right)$

核回归(huíguī)：Nadaraya-Watson

- 最后，得到估计的风险为

$$R(r,\hat{r}_n) \approx \frac{1}{4}h^4\left(\int\left(x^2K^2(x)\right)dx\right)^4\int\left(r''(x)+2r'(x)\frac{f'(x)}{f''(x)}\right)dx$$

$$+\int\frac{\int\sigma^2K^2(x)dx}{nhf(x)}dx$$

  - 最佳带宽以 $n^{-1/5}$ 的速率减少，在这种选择下风险以 $n^{-4/5}$ 的速率减少，这是最佳收敛速率（同核密度估计）

核回归(huíguī)：Nadaraya-Watson

- 实际应用中，利用交叉验证对求最佳带宽h。交叉验证对风险的估计为

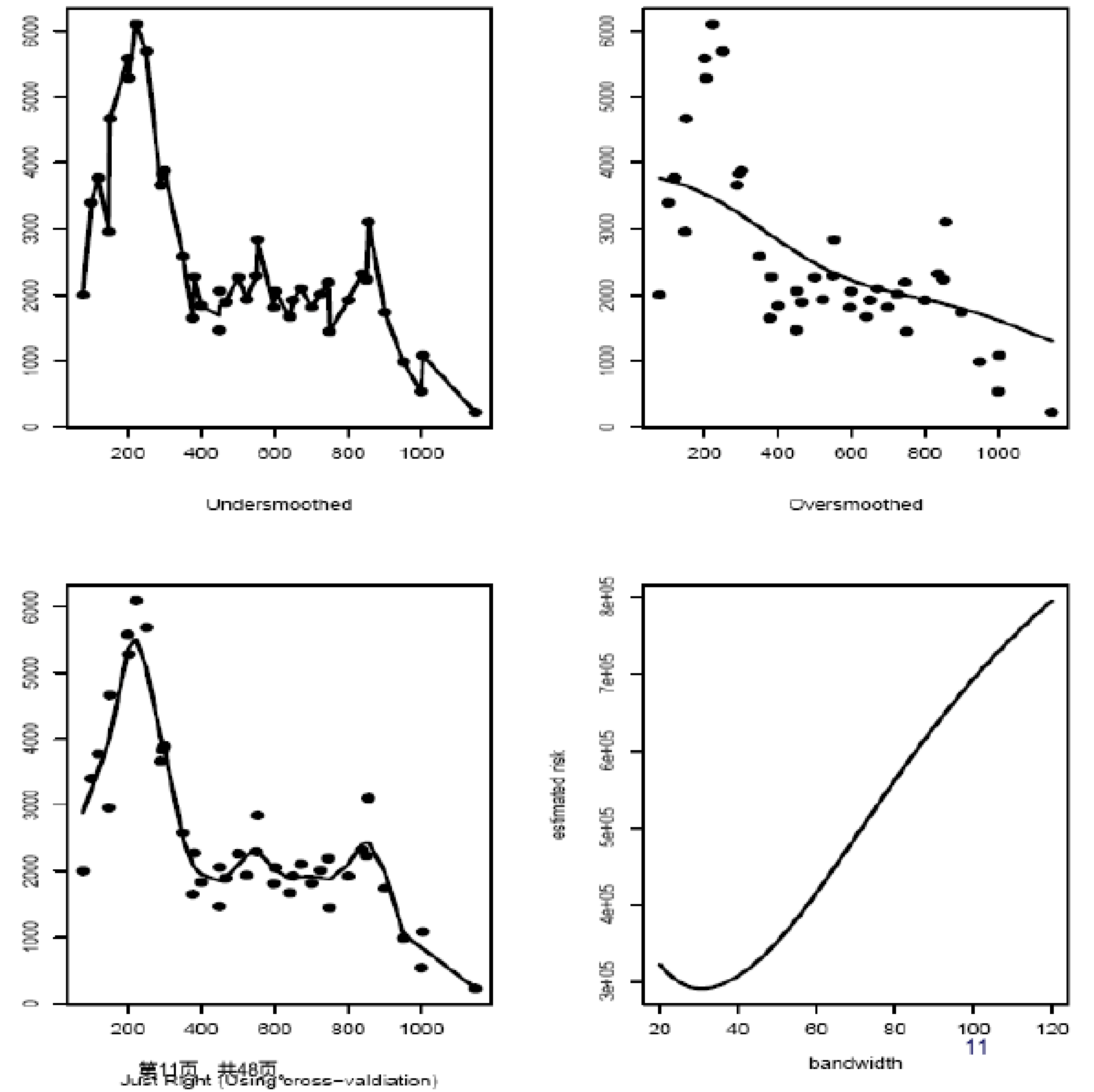
$$\hat{J}(h)=\sum_{i=1}^{n}\left(Y_i-\hat{r}_{-i}(x_i)\right)^2$$

- 实际上不必每次留下一个计算单独估计，可以写成以下形式

$$\hat{J}(h)=\sum_{i=1}^{n}\left(Y_i-\hat{r}(x_i)\right)^2\frac{1}{\left(1-\frac{K(0)}{\sum_{j=1}^{n}K\left(\frac{x_i-x_j}{h}\right)}\right)^2}$$

例：Example 20.23

不同带宽(dài kuān)下Nadaraya-Watson回归的结果



核回归(huíguī)：Nadaraya-Watson

- 模型类型：非参数
- 损失：平方误差
- 参数选择：留一交叉验证

局部(júbù)线性回归

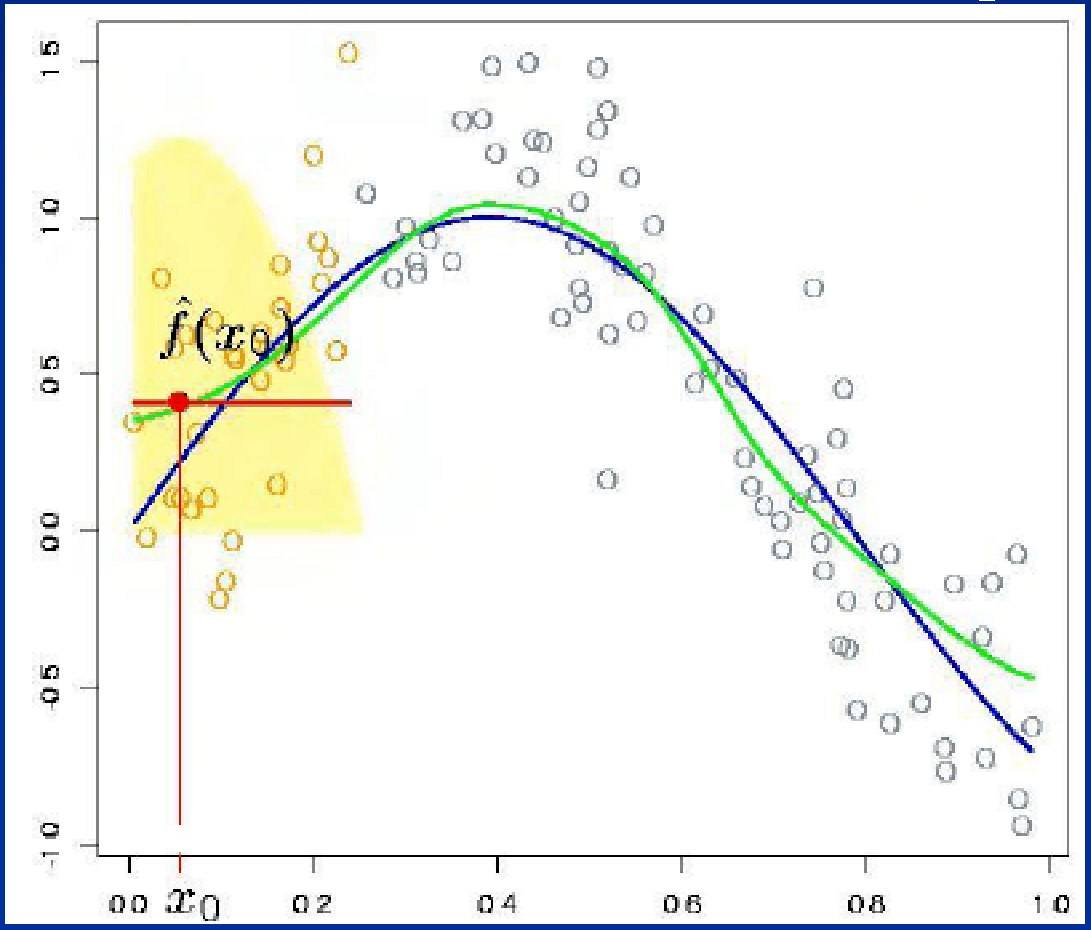
- 问题：加权核回归在训练数据中靠近边界的点的估计很差
  - 核在边界区域不对称，局部加权平均在边界区域上出现严重偏差 → 局部线性回归

- 局部线性回归：在每一个将要被预测的点x处解一个单独的加权最小二乘问题，找到使下述表达式最小的 $\beta(x)$

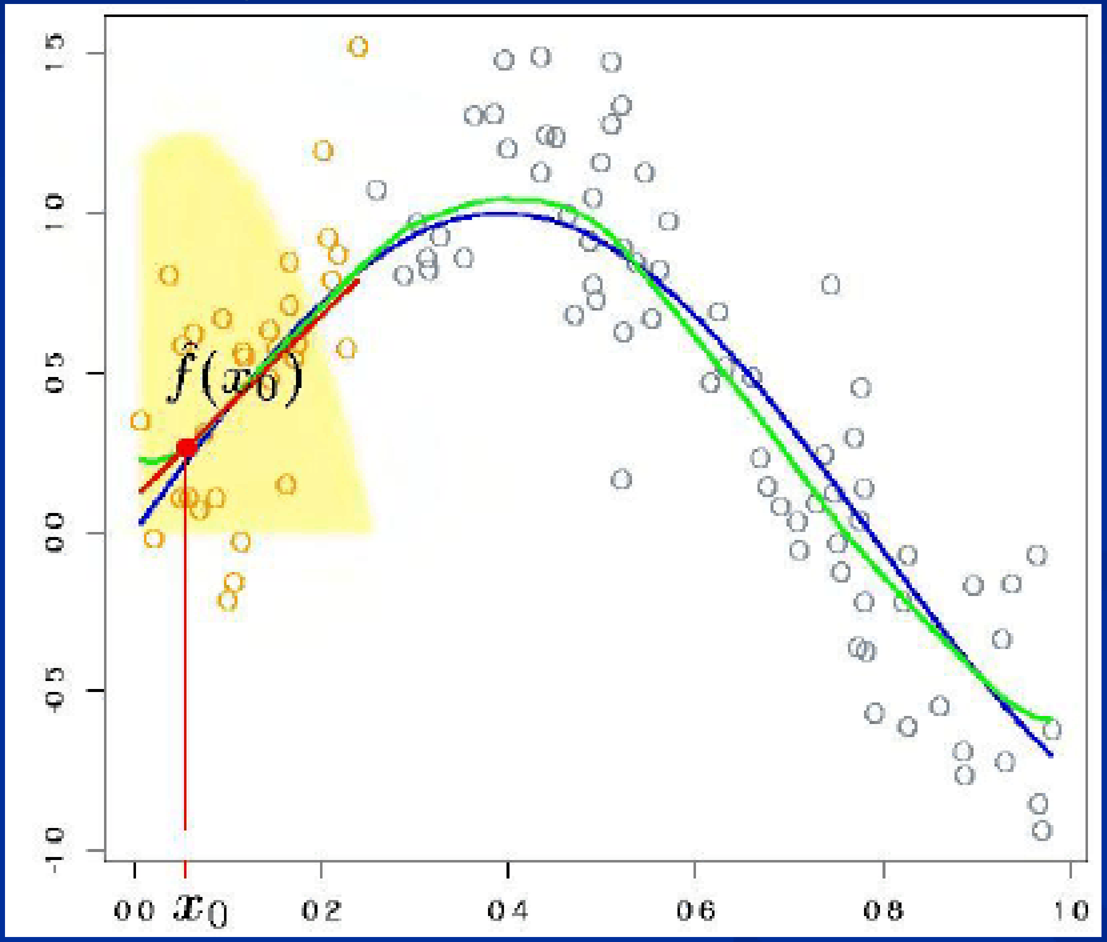
$$\sum_{i=1}^{n} K_h(x, x_i)\left[y_i - x_i\beta(x)\right]^2$$

局部(júbù)线性回归

$$Y = \sin(X) + \varepsilon,\ X \sim Uniform[0,1],\ \varepsilon \sim N(0,1/3)$$



边界上的N-W核：
核在边界不对称→偏差大



边界上的局部线性回归：
将偏差降至一阶

蓝色曲线：真实情况
绿色曲线：估计值
黄色区域：x_0的局部区域

核回归(huíguī)：局部线性回归(huíguī)

- 则估计为：

$$\hat{r}(x)=x\hat{\beta}(x)$$
$$=x^T\left(\mathbf{X}^T\mathbf{W}(x)\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}(x)\mathbf{y}$$
$$=\sum_{i=1}^{n}w_i(x)y_i$$

- 其中$W(x)$是一个 $n\times n$的对角矩阵且第i个对角元素是 $K_h(x,x_i)$
- 估计在y_i上是线性的，因为权重项 $w_i(x)$不涉及y_i ，可被认为是等价核

局部(júbù)线性回归

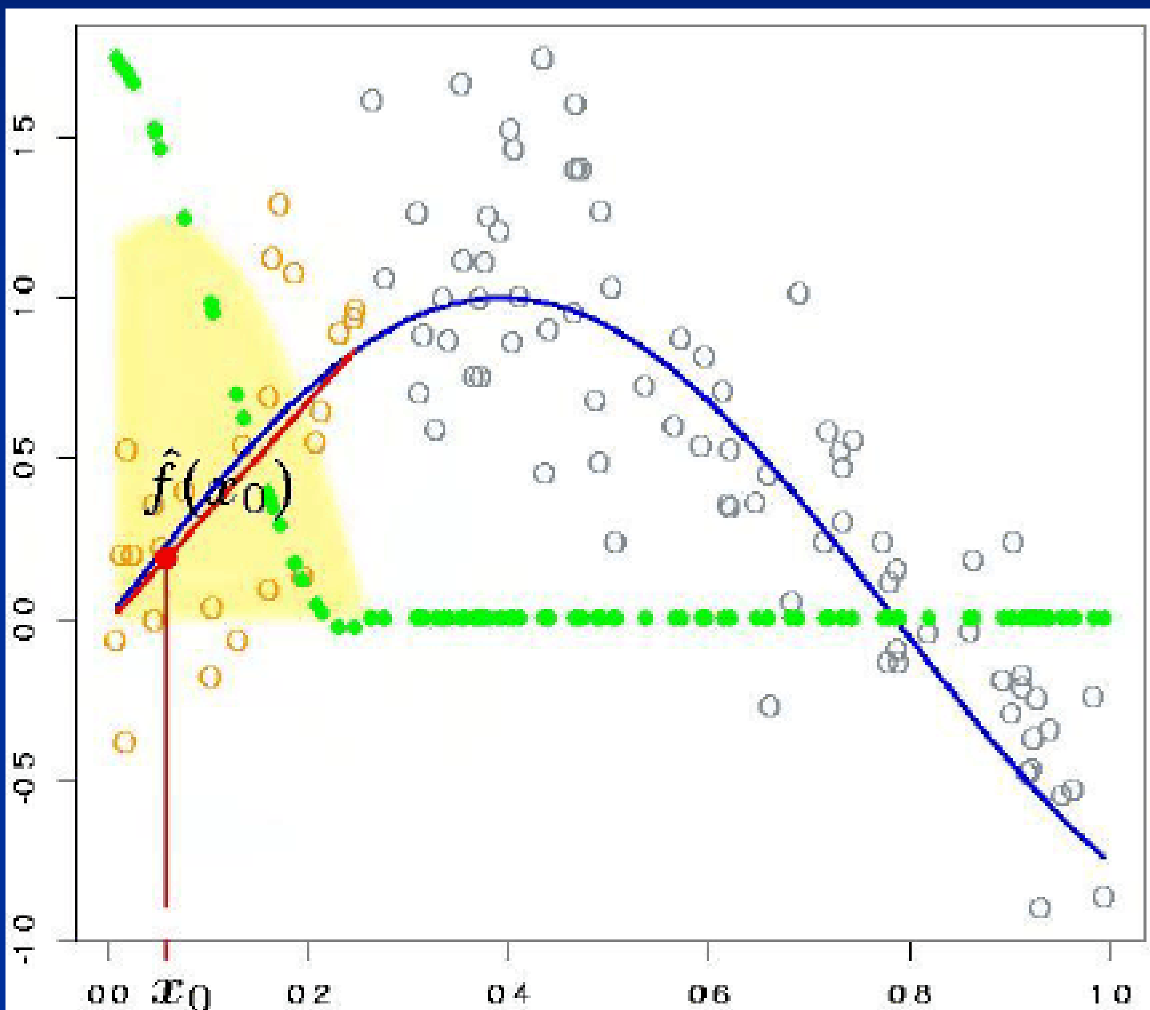
$$r(x_i) \approx r(x_0) + r'(x_0)(x_i - x_0) + \frac{r''(x_0)}{2}(x_i - x_0)^2$$

- 局部线性回归通过自动修改核，将偏差降至一阶

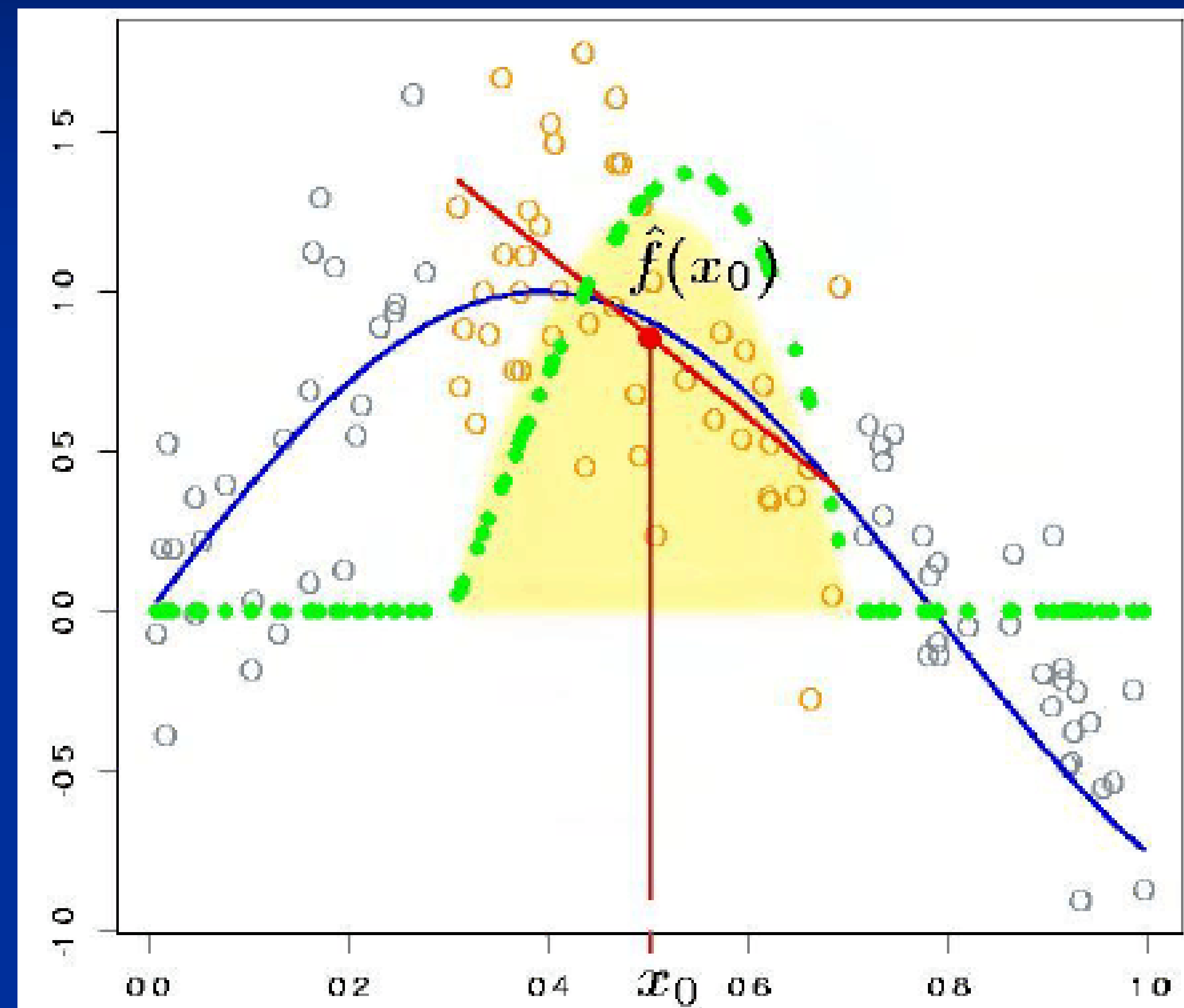
$$\mathbb{E}(\hat{r}(x_0)) = \sum_{i=1}^{n} w_i(x_0) r(x_i)$$
$$\approx r(x_0)\sum_{i=1}^{n} w_i(x_0) + r'(x_0)\sum_{i=1}^{n}(x_i - x_0) w_i(x_0)$$
$$+\frac{r''(x_0)}{2}\sum_{i=1}^{n}(x_i - x_0)^2 w_i(x_0)$$

- 由于 $\sum_{i=1}^{n} w_i(x_0) = 1$, $\sum_{i=1}^{n}(x_i - x_0) w_i(x_0) = 0$
- 偏差 为 $\mathbb{E}(\hat{r}(x_0)) - r(x_0) \approx \dfrac{r''(x_0)}{2}\sum_{i=1}^{n}(x_i - x_0)^2 w_i(x_0)$

局部(júbù)线性回归



边界上的局部等价核
（绿色点）



内部区域的局部等价核
（绿色点）

局部(júbù)多项式回归

- 局部多项式回归：用d次多项式回归代替线性回归
  - 可以考虑任意阶的多项式，但有一个偏差和方差的折中
  - 通常认为：超过线性的话，会增大方差，但对偏差的减少不大，因为局部线性回归能处理大多数的边界偏差，

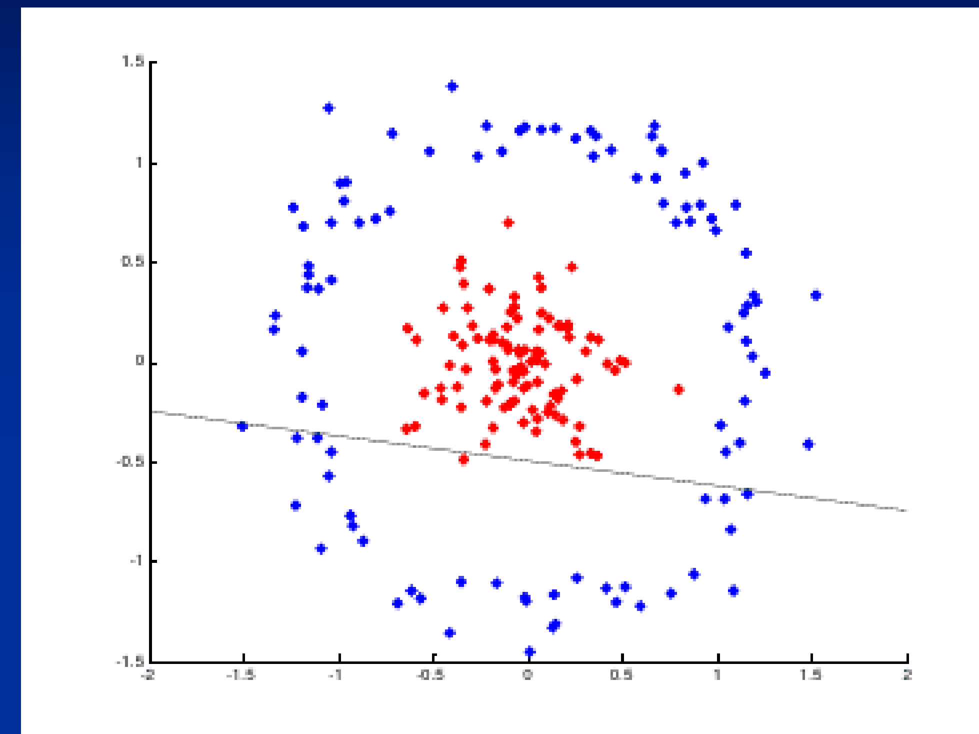
可变宽度(kuāndù)核

- 可变宽度核：如使每一个训练点的带宽与它的第k个近邻的距离成反比
  - 在实际应用中很好用，虽然尚未有理论支持怎样选择参数
  - 不会改变收敛速度，但在有限样本时表现更好
- 注意：上述这些扩展（包括局部线性/局部多项式）都可应用到核密度估计中

核方法(fāngfǎ)

- 为什么要用核方法？
  - 得到更丰富的模型，但仍然采用同样的方法
    - 如岭回归方法→核岭回归
- 内容
  - Kernel trick
  - 再生Hilbert空间

线性模型(móxíng)



- 线性模型：
  - 方便、应用广泛
  - 有很强的理论保证
  - 但还是有局限性

- 可以通过扩展特征空间增强线性模型的表示能力
  - 如

$$\phi(x_1,x_2)=\left(1,\sqrt{2}x_1,\sqrt{2}x_2,\sqrt{2}x_1x_2,x_1^2,x_2^2\right)$$

  - 特征空间为R^6而不是R^2特
  - 该特征空间的线性预测器为

$$y=\begin{bmatrix}1 & 0 & 0 & 0 & -1 & -1\end{bmatrix}\times\phi(x)=1-x_1^2-x_2^2$$

岭回归(huíguī)

- 对给定的 $\lambda > 0$
- 最小化正则化的残差

$$RSS^{ridge}(\beta,\lambda) = \left(\mathbf{y}-\mathbf{X}\beta\right)^T\left(\mathbf{y}-\mathbf{X}\beta\right)+\lambda\beta^T\beta$$

$$= \lambda\|\beta\|^2 + \|\mathbf{y}-\mathbf{X}\beta\|^2$$

- 则最优解为

$$\frac{RSS^{ridge}(\beta,\lambda))}{\partial\beta} = 2\lambda\beta - 2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\beta = 0$$

$$\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p\right)\beta = \mathbf{X}^T\mathbf{y}$$

$$\Rightarrow \quad \hat{\beta} = \left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{y}$$

需$O(p^3)$运算

对偶表示

$$\boxed{\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}_p\right)\beta=\mathbf{X}^T\mathbf{y}}$$

- 一种(yī zhǒng)对偶表示为：

$$\left(\mathbf{X}^T\mathbf{X}+\lambda\mathbf{I}\right)\beta=\mathbf{X}^T\mathbf{y} \quad\Rightarrow\quad \beta=\lambda^{-1}\left(\mathbf{X}^T\mathbf{y}-\mathbf{X}^T\mathbf{X}\beta\right)$$

$$\Rightarrow\quad \beta=\lambda^{-1}\mathbf{X}^T\left(\mathbf{y}-\mathbf{X}\beta\right)=\mathbf{X}^T\alpha$$

$$\alpha=\lambda^{-1}\left(\mathbf{y}-\mathbf{X}\beta\right)$$

$$\Rightarrow\lambda\alpha=\left(\mathbf{y}-\mathbf{X}\beta\right)=\left(\mathbf{y}-\mathbf{X}\mathbf{X}^T\alpha\right)$$

$$\Rightarrow\mathbf{X}\mathbf{X}^T\alpha+\lambda\alpha=\mathbf{y}$$

$$\Rightarrow\alpha=\left(\mathbf{G}+\lambda\mathbf{I}_n\right)^{-1}\mathbf{y}$$

需$O(n^3)$运算

- 其中 $\mathbf{G}=\mathbf{X}\mathbf{X}^T$

对偶(duì ǒu)岭回归

$$\beta = \mathbf{X}^T\alpha,\quad \alpha = \left(\mathbf{G} + \lambda\mathbf{I}_n\right)^{-1}\mathbf{y}$$

- 为了预测一个新的点

$$f(x) = \langle \beta, x \rangle = \left\langle \sum_{i=1}^{n} \alpha_i x_i, x \right\rangle = \mathbf{y}^T \left(\mathbf{G} + \lambda\mathbf{I}\right)^{-1} z$$

- 其中 $z = \langle x_i, x \rangle$

- 此时只需计算Gram矩阵G

$$\mathbf{G} = \mathbf{X}\mathbf{X}^T,\quad G_{ij} = \langle x_i, x_j \rangle$$

岭回归只需计算数据点的内积

特征空间(kōngjiān)中的线性回归

- 基本思想：
  - 将数据映射到高维空间（特征空间）
  - 然后在高维空间中用线性方法
- 嵌入式特征映射：

$$\phi : x \in R^p \to F \subseteq R^P \quad P >> p$$

核函数(hánshù)

- 则核函数为

$$K\langle x,u\rangle = \langle \phi(x),\phi(u)\rangle_F$$

- 其中 ϕ 为将数据映射到高维空间的映射

- 有许多可能的核函数
- 最简单的为核

$$K\langle x,u\rangle = \langle x,u\rangle$$

特征(tèzhēng)空间中的岭回归

$$\boxed{\beta = \mathbf{X}^T\alpha, \quad \alpha = \left(\mathbf{G} + \lambda \mathbf{I}_n\right)^{-1}\mathbf{y}}$$

- 为了预测一个新的点

$$f(x) = \langle \beta, \phi(x) \rangle = \left\langle \sum_{i=1}^{n} \alpha_i \phi(x_i), \phi(x) \right\rangle = \mathbf{y}^T \left(\mathbf{G} + \lambda \mathbf{I}\right)^{-1} z$$

- 其中 $z = \langle \phi(x_i), \phi(x) \rangle$

- 计算Gram矩阵G

$$\mathbf{G} = \phi(\mathbf{X})\left(\phi(\mathbf{X})\right)^T, \quad G_{ij} = \left\langle \phi(x_i), \phi(x_j) \right\rangle = K(x_i, x_j)$$

利用核函数计算内积

另一种(yī zhǒng)对偶表示推导方式

- 线性岭回归最小化：

$$\left(\sum_{i=1}^{n}\left(y_i - r(x_i)\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right),\ \ r(x_i) = \sum_{j=1}^{p} x_{ij}\beta_j$$

- 等价于

$$\min_{\beta_j}\left(\sum_{i=1}^{n}\xi_i^2 + \lambda\sum_{j=1}^{p}\beta_j^2\right)$$

- 满足约束

$$\xi_i = y_i - r(x_i) = y_i - \sum_{j=1}^{p} x_{ij}\beta_j$$

- 则拉格朗日函数为

$$L(\alpha,\beta,\xi) = \sum_{i=1}^{n}\xi_i^2 + \lambda\sum_{j=1}^{p}\beta_j^2 + \sum_{i=1}^{n}\alpha_i\left(\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right) - \xi_i\right)$$

Wolfe对偶(duì ǒu)问题

- 转化为其对偶问题： $Q(\alpha)=\min_{\beta,\xi} L(\alpha,\beta,\xi)$
- 对L求偏导并置为0，得到

$$\frac{\partial L}{\partial \beta_j}=2\lambda\beta_j-\sum_{i=1}^{n}\alpha_i x_{ij}=0 \Rightarrow \beta_j=\frac{1}{2\lambda}\sum_{i=1}^{n}\alpha_i x_{ij}$$

$$\frac{\partial L}{\partial \xi_i}=2\xi_i-\alpha_i=0 \Rightarrow \xi_i=\frac{1}{2}\alpha_i$$

$$\boxed{L(\alpha,\beta,\xi)=\sum_{i=1}^{n}\xi_i^2+\lambda\sum_{j=1}^{p}\beta_j^2+\sum_{i=1}^{n}\alpha_i\left(\left(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\right)-\xi_i\right)}$$

Wolfe对偶(duì ǒu)问题

- 将 $\beta_j = \frac{1}{2\lambda}\sum_{i=1}^{n}\alpha_i x_{ij}$ 和 $\xi_i = \frac{1}{2}\alpha_i$ 代入拉格朗日函数
- 原目标函数

$$L(\alpha,\beta,\xi) = \sum_{i=1}^{n}\xi_i^2 + \lambda\sum_{j=1}^{p}\beta_j^2 + \sum_{i=1}^{n}\alpha_i\left(\left(y_i - \sum_{j=1}^{p}x_{ij}\beta_j\right) - \xi_i\right)$$

- 转化为

$$Q(\alpha) = \frac{1}{4}\sum_{i=1}^{n}\alpha_i^2 + \frac{1}{4\lambda}\sum_{j=1}^{p}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k x_{ij}x_{kj}$$

$$+\sum_{i=1}^{n}\alpha_i\left(y_i - \sum_{j=1}^{p}x_{ij}\frac{1}{2\lambda}\sum_{k=1}^{n}\alpha_k x_{kj} - \frac{1}{2}\alpha_i\right)$$

$$= -\frac{1}{4}\sum_{i=1}^{n}\alpha_i^2 - \frac{1}{4\lambda}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k\sum_{j=1}^{p}x_{ij}x_{kj} + \sum_{i=1}^{n}\alpha_i y_i$$

最优解

$$Q(\alpha)=\sum_{i=1}^{n}\alpha_i y_i-\frac{1}{4}\sum_{i=1}^{n}\alpha_i^2-\frac{1}{4\lambda}\sum_{i=1}^{n}\sum_{k=1}^{n}\alpha_i\alpha_k\sum_{j=1}^{p}x_{ij}x_{kj}$$

- 写成矩阵形式(xíngshì)为：

$$Q(\alpha)=\alpha^T y-\frac{1}{4}\alpha^T\alpha-\frac{1}{4\lambda}\alpha^T\mathbf{G}\alpha,\ \text{where } G_{ij}=\underline{\langle x_i,x_j\rangle}$$

点积 $=\sum_{j=1}^{p}x_{ij}x_{kj}=x_i^T x_k$

- 得到解：

$$\frac{\partial Q}{\partial\alpha}=y-\frac{1}{2}\alpha-\frac{1}{2\lambda}\mathbf{G}\alpha=0\ \Rightarrow\ \alpha=2\lambda(\mathbf{G}+\lambda\mathbf{I})^{-1}y$$

- 相应的回归方程为：

$$\beta_j=\frac{1}{2\lambda}\sum_{i=1}^{n}\alpha_i x_{ij}$$

$$\hat{r}(x)=\langle x,\beta\rangle=y^T(\mathbf{G}+\lambda I)^{-1}z,\ \text{where } z=\langle x,x_i\rangle$$

核化岭回归(huíguī)

- 将点积 $G_{ij} = \langle x_i, x_j \rangle$ 换成核函数 $K_{ij} = K(x_i, x_j)$
  - Kernel trick
- 就实现了对线性岭回归的核化，在空间统计学中称为Kriging算法。

核方法(fāngfǎ)

- 通过将输入空间映射到高维空间（特征空间），然后在高维空间中用线性方法
  - 高维：维数灾难
  - 通过核技巧，避免维数灾难

Kernel Trick

- 将问题(wèntí)变为其对偶问题(wèntí)：只需计算点积，与特征的维数无关，如在线性岭回归中，最大化下列目标函数

$$Q(\alpha)=\alpha^T y-\frac{1}{4}\alpha^T\alpha-\frac{1}{4\lambda}\alpha^T\mathbf{G}\alpha,\ \text{where } G_{ij}=\langle x_i,x_j\rangle$$

- 在高维空间中的点积可写成核(*kernel*)的形式，
- 如果选定核函数，这无需计算映射 $\phi(x)$ 可以计算点积

$$Q(\alpha)=\alpha^T y-\frac{1}{4}\alpha^T\alpha-\frac{1}{4\lambda}\alpha^T\mathbf{K}\alpha,\ \text{where } K_{ij}=\langle\phi(x_i),\phi(x_j)\rangle$$

Kernel Trick

- 总之，这些被称为核技巧(*kernel trick*)，寻找一个映射：$\phi:\mathcal{X}\to\mathcal{F}$和一个学习方法，使得
  - F的维数比X高，因此模型更丰富
  - 算法只需要计算点积
  - 存在一个核函数(hánshù)，使得

$$\langle\phi(x),\phi(\tilde{x})\rangle\equiv K(x,\tilde{x})$$

点积核

  - 在算法中任何出现项$\langle x,\tilde{x}\rangle$的地方，用$K(x,\tilde{x})$代替

亦称为原方法的核化(*kernelizing* the original method).

什么样的函数(hánshù)可以作为核函数(hánshù)?

- Mercer's 定理给出了连续对称函数k可作为核函数的充要条件：半正定

- 半正定核：
  - 对称：$k\left(x,\tilde{x}\right)=k\left(\tilde{x},x\right)$
  - 且对任意训练样本点 $x_1,\ldots,x_n$
  - 和任意 $\alpha_1,\ldots,\alpha_n\in R$
  - 满足

$$\sum_{i,j}\alpha_i\alpha_j K_{ij}\geq 0,\ \ K_{ij}=k\left(x_i,x_j\right)$$

矩阵形式：

$$\alpha^T\mathbf{K}\alpha\geq 0$$

- K被称为$Gram$矩阵或核矩阵。

半正定(zhèng dìng)核的性质

- 对称

$$K(x,\tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle = \langle \phi(\tilde{x}), \phi(x) \rangle = K(\tilde{x}, x)$$

- Cauchy-Schwarz不等式

$$K(x,\tilde{x})^2 = \langle \phi(x), \phi(\tilde{x}) \rangle^2 \le \|\phi(x)\|^2 \|\phi(\tilde{x})\|^2 = K(x,x)K(\tilde{x},\tilde{x})$$

Mercer's Theorem

- 当且仅当一个函数K满足半正定形式(xíngshì)时，函数K可以写成

$$K(x,\tilde{x}) = \langle \phi(x), \phi(\tilde{x}) \rangle = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(\tilde{x})$$

- 其中 $\phi(x)$ 为特征映射： $x \to \phi(x) \in F$

- 该核定义了一个函数集合 $\mathcal{H}_K$ 其中每个元素 可以写成

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$$

Mercer核

- 因此某些核对应无限个预测变量的变换

RKHS：再生(zàishēng)Hilbert空间
—Reproducing Kernel Hilbert Spaces

- 为了证明上述定理，构造一个特殊的特征空间

$$\phi : x \to \phi(x) = k(x,.)$$

映射到一个函数空间

<u>定义函数空间</u>

$$f(.) = \sum_{i=1}^{n} \alpha_i k(x_i,.), \ \forall \ n, x_i$$

$$< f, g >= \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \beta_j k(x_i, x_j)$$

$$< f, f >= \sum_{i=1}^{n} \sum_{j=1}^{n'} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

有限、半正定

<u>再生性质</u>

$$\left\langle f, \phi(x) \right\rangle = \left\langle f, k(x,.) \right\rangle =$$

$$\left\langle \sum_{i=1}^{n} \alpha_i k(x_i,.), k(x,.) \right\rangle =$$

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

$$\Rightarrow \boxed{\left\langle \phi(x), \phi(\tilde{x}) \right\rangle = k(x, \tilde{x})}$$

Mercer's Theorem

- 粗略地说，如果K对可积函数 g

$$\int g^2(x)\,dx < \infty$$

- 是正定的，即

$$\int K(x,\tilde{x})g(x)g(\tilde{x})\,dx\,d\tilde{x} \geq 0,\ \forall g \in L_2$$

- 则对K存在(cúnzài)对应ϕ的

$$K(x,\tilde{x}) = \sum_{j=1}^{\infty} \gamma_j \phi_j(x)\phi_j(\tilde{x})$$

- 因此K是一个合适的核

Mercer 核

- 一些常用(cháng yòng)的核函数满足上述性质：

高斯核：$$K(x,\tilde{x}) = \exp\left(-\frac{(x-\tilde{x})^2}{2a^2}\right)$$

多项式核：$$K(x,\tilde{x}) = \left(\langle x,\ \tilde{x}\rangle + a\right)^b$$

sigmoid核：$$K(x,\tilde{x}) = \tanh\left(\langle x,\ \tilde{x}\rangle + a\right)$$

- 对字符串、图等对象，也可以构造核函数

RKHS：点积空间(kōngjiān)

- 定义该函数空间的点积

$$\left\langle \sum_{i=1}^{\infty} c_i \phi_i(x), \sum_{i=1}^{\infty} d_i \phi_i(x) \right\rangle = \sum_{i=1}^{\infty} \frac{c_i d_i}{\lambda_i}$$

- Mercer定理隐含

$$\|f\|_{\mathrm{H}_K}^2 = \sum_{i=1}^{\infty} c_i^2 \big/ \lambda_i < \infty$$

正则(zhèng zé)化和RKHS

- 一种通用的正则化的形式为

$$\min_{f \in H} \left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f) \right]$$

- 假设 f 在RKHS中，则

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y)$$

$$f(x) = \sum_{i=1}^{\infty} c_i \phi_i(x) = \sum_{i=1}^{n} \alpha_i K(x, x_i)$$

$$J(f) = \|f\|_{\mathrm{H}_K}^2 = \sum_{i=1}^{n} \sum_{i'=1}^{n} \alpha_i \alpha_i' K\left(x_i, x_i'\right)$$

正则(zhèng zé)化和RKHS

- 则求解

$$\min_{f\in H}\left[\sum_{i=1}^{N} L(y_i, f(x_i)) + \lambda J(f)\right]$$

- 转化为求解下述“简单”问题

$$\min_{\alpha}\left[L(y, K\alpha) + \lambda\alpha^T K\alpha\right]$$

例：岭回归(huíguī)

- 当回归分析取平方误差损失时，

$$\hat{\alpha} = (K + \lambda I)^{-1} y \quad , \hat{f}(x) = \sum_{i=1}^{n} \hat{\alpha}_i K(x, x_i)$$

- 因此

$$\hat{f} = K\hat{\alpha} = K(K + \lambda I)^{-1} y = (I + \lambda K^{-1})^{-1} y$$

正则(zhèng zé)化的贝叶斯解释

- $$\min_{f\in\mathcal{H}}\left[\sum_{i=1}^{n}L\left(y_i,f\left(x_i\right)\right)+\lambda\|f\|_{\mathcal{H}_K}^2\right]$$
  $$-\log\mathbb{P}\left(f\mid\left(x_1,y_1,\ldots,x_n,y_n\right)\right)$$
- 为贝叶斯MAP估计
- 其中先验为 $\exp\left(-\lambda\|f\|_{\mathcal{H}_K}^2\right)$
- 似然为 $\exp\left(-\sum_{i=1}^{n}L\left(y_i,f\left(x_i\right)\right)\right)$
  - 损失函数取L_2时，高斯分布： $\exp\left(-\left(y_i-f\left(x_i\right)\right)^2\right)$
  - 损失函数取L_1时，为Laplace分布： $\exp\left(-\left|y_i-f\left(x_i\right)\right|\right)$

其他与核方法(fāngfǎ)相关的一些论题

- 高斯过程
- SVM
- …

- 关于核方法一本较好的参考书：
  - 支持向量机导论（An Introduction to Support Vector Machines and Other Kernel-based Learning Methods）
  - Nello Cristianini, John Shawe-Taylor著，李国正，王猛， 曾华军译， 电子工业出版社，北京，2004
  - Bernhard Schölkopf: Introduction to Kernel Methods, Analysis of Patterns Workshop, Erice, Italy, 2005
  - Schölkopf& Smola: Learning with Kernels, MIT Press, 2002

下节课内容(nèiróng)

- 模型选择: [ESL] Chp7