

Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 14 - RKHS, Non-parametric Regression

Sequential Minimal Optimization Algorithm for Solving SVR

Solving the SVR Dual Optimization Problem

- It can be shown that the objective:

$$\max_{\alpha_i, \alpha_i^*} - \frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\mathbf{x}_i, \mathbf{x}_j) \\ - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*)$$

- can be written as:

$$\max_{\beta_i} - \frac{1}{2} \sum_i \sum_j \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i \beta_i \\ \text{s.t.}$$

¹https://en.wikipedia.org/wiki/Quadratic_programming#Solvers_and_scripting_.28programming.29_languages

Solving the SVR Dual Optimization Problem

- It can be shown that the objective:

$$\max_{\alpha_i, \alpha_i^*} -\frac{1}{2} \sum_i \sum_j (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) K(\mathbf{x}_i, \mathbf{x}_j) \\ - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i (\alpha_i - \alpha_i^*)$$

- can be written as:

$$\max_{\beta_i} -\frac{1}{2} \sum_i \sum_j \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i \beta_i \\ \text{s.t.}$$

- $\sum_i \beta_i = 0$
- $\beta_i \in [-C, C], \forall i$

$$\left. \begin{array}{l} \max_{\beta_1, \beta_2} a\beta_1\beta_2 + b\beta_1^2 + c\beta_2^2 \\ \quad + d\beta_1 + e\beta_2 \\ \beta_1 + \beta_2 = C \\ \beta_1, \beta_2 \in [0, C] \end{array} \right\}$$

- Even for this form, standard QP (LCQP) solvers¹ can be used
- Question: How about (iteratively) solving for two β_i 's at a time?
 - This is the idea of the Sequential Minimal Optimization (SMO) algorithm

¹https://en.wikipedia.org/wiki/Quadratic_programming#Solvers_and_scripting_.28programming.29_languages

Sequential Minimal Optimization (SMO) for SVR

- Consider:

$$\max_{\beta_i} -\frac{1}{2} \sum_i \sum_j \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i \beta_i$$

s.t.

- $\sum_i \beta_i = 0$
- $\beta_i \in [-C, C], \forall i$

- The SMO subroutine can be defined as:

β_1^{new}
 β_2^{new} } $\arg \max_{\beta_1, \beta_2} O(\beta_1, \beta_2)$
Fix all β_i 's except β_1 & β_2 (wlog)
 $\beta_1 + \beta_2 = \text{const}$
 $\beta_1, \beta_2 \in [-C, C]$ } Solve for convergence
Check if with old β_i 's & $\beta_1^{\text{new}}, \beta_2^{\text{new}}$ KKT violated. If yes, pick next "most" violating β_i, β_j

Sequential Minimal Optimization (SMO) for SVR

- Consider:

$$\max_{\beta_i} -\frac{1}{2} \sum_i \sum_j \beta_i \beta_j K(\mathbf{x}_i, \mathbf{x}_j) - \epsilon \sum_i |\beta_i| + \sum_i y_i \beta_i$$

s.t.

- $\sum_i \beta_i = 0$
- $\beta_i \in [-C, C], \forall i$

- The SMO subroutine can be defined as:

- 1 Initialise β_1, \dots, β_n to some value $\in [-C, C]$
- 2 Pick β_i, β_j to estimate closed form expression for next iterate (i.e. $\beta_i^{new}, \beta_j^{new}$)
- 3 Check if the KKT conditions are satisfied
 - If not, choose β_i and β_j that worst violate the KKT conditions and reiterate

if $d_i = 0, \alpha_i^* = 0 \Rightarrow \xi_i, \xi_i^* = 0$ (could be violated)
if $\alpha_i = C$ or $\alpha_i^* = C \Rightarrow \xi_i > 0$ or $\xi_i^* > 0 \dots$

convergence proof use this

Lasso and Dual Form



$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}} \|\phi \mathbf{w} - \mathbf{y}\|^2 \quad \text{s.t.} \quad \underbrace{\| \mathbf{w} \|_1}_{\text{Penalized}} \leq \eta, \quad (1)$$

where

$$\| \mathbf{w} \|_1 = \left(\sum_{i=1}^n |w_i| \right) \quad (2)$$

- Since $\| \mathbf{w} \|_1$ is not differentiable, one can express (2) as a set of constraints placeholders for $|w_i|$

$$\sum_{i=1}^n \xi_i \leq \eta, \quad w_i \leq \xi_i, \quad -w_i \leq \xi_i$$

- The resulting problem is a linearly constrained Quadratic optimization problem (LCQP):

$$\mathbf{w}^* = \underset{\mathbf{w}, \xi_i}{\operatorname{argmin}} \|\phi \mathbf{w} - \mathbf{y}\|^2 \quad \text{s.t.} \quad \boxed{\sum_{i=1}^n \xi_i \leq \eta, \quad \mathbf{w}_i \leq \xi_i, \quad -\mathbf{w}_i \leq \xi_i} \quad (3)$$

No $\| \mathbf{w} \|_1$ component

$$\|w\|_1 \leq \eta$$

$$|||$$

$$\underbrace{|w_i| \leq \xi_i} \text{ \& } \sum_i \xi_i \leq \eta$$

$$w_i \leq \xi_i \text{ \& } -w_i \leq \xi_i$$

$$\min_{w, \xi} \frac{1}{2} \|\phi w - y\|^2 \quad \text{s.t.} \quad w_i \leq \xi_i, -w_i \leq \xi_i, \sum \xi_i \leq \eta$$

Case 1: w^* without constraint $\|w\|_1 \leq \eta$ satisfies

Case 2: $\|w^*\|_1 = \eta \equiv \xi_i = |w_i|$
(obvious)

$\|w^*\|_1 \leq \eta$ then any ξ_i s.t.
 $\sum \xi_i = \eta$ \& $\xi_i \geq |w_i|$ is OK

Non Parametric Regression

Basis function expansion and the Kernel trick: Additional Discussion 1

Consider regression function $f(\mathbf{x}) = \sum_{j=1}^p w_j \phi_j(\mathbf{x})$ with weight vector \mathbf{w} estimated as

$$\mathbf{w}_{Pen} = \underset{\mathbf{w}}{\operatorname{argmin}} \mathcal{L}(\phi, \mathbf{w}, \mathbf{y}) + \lambda \Omega(\mathbf{w})$$

It can be shown that for $p \in [0, \infty)$, under certain conditions on K , the following can be equivalent representations

-

$$f(\mathbf{x}) = \sum_{j=1}^p w_j \phi_j(\mathbf{x})$$


- And²

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

²Section 5.8.1 of Tibshi.

The Representer Theorem & Reproducing Kernel Hilbert Space (RKHS)

- ① The solution $f^* \in \mathcal{H}$ (Hilbert space) to the following problem

$$f^* = \arg \min_{f \in \mathcal{H}} \sum_{i=1}^m \mathbf{E} \left(f \left(\mathbf{x}^{(i)} \right), y^{(i)} \right) + \Omega(\|f\|_K)$$


can be always written as $f^*(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$, provided $\Omega(\|f\|_K)$ is a

- ② More specifically, if $f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b$ and $K(\mathbf{x}', \mathbf{x}) = \phi^T(\mathbf{x})\phi(\mathbf{x}')$ then the solution $\mathbf{w}^* \in \mathbb{R}^n$ to the following problem

$$(\mathbf{w}^*, b^*) = \arg \min_{\mathbf{w}, b} \sum_{i=1}^m \mathbf{E} \left(f \left(\mathbf{x}^{(i)} \right), y^{(i)} \right) + \Omega(\|\mathbf{w}\|_2)$$

can be always written as $\phi^T(\mathbf{x})\mathbf{w}^* + b = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}^{(i)})$, provided $\Omega(\|\mathbf{w}\|_2)$ is a monotonically increasing function of $\|\mathbf{w}\|_2$. \mathbb{R}^n is the Hilbert space and $K(., \mathbf{x}) : \mathcal{X} \rightarrow \mathbb{R}$ is the **Reproducing (RKHS) Kernel**

The Reproducing Kernel Hilbert Space (RKHS)

$$\mathcal{K} = \{f(\cdot) = K(\cdot, x) \mid x \in \mathcal{X}\} = \{K_x(\cdot) \mid K_x(\cdot) = K(\cdot, x)\}$$

family of fns, one fn for each x fixed

Consider the set of functions $\mathcal{K} = \{K(\cdot, x) \mid x \in \mathcal{X}\}$ and let \mathcal{H} be the set of all functions that are **finite** linear combinations of functions in \mathcal{K} . That is, any function

$h \in \mathcal{H}$ can be written as $h(\cdot) = \sum_{t=1}^T \alpha_t K(\cdot, x_t)$ for some T and $x_t \in \mathcal{X}, \alpha_t \in \mathbb{R}$. One

can easily verify that \mathcal{H} is a vector space³ with an inner product.

$$\mathcal{H} = \left\{ \sum_{t=1}^T \alpha_t K_{x_t}(\cdot) \right\} = \text{Space of linear combinations of kernel fn applied to a subset of pts (as second arg)}$$

SVR searches for some good $\{\alpha_1, \alpha_2, \dots, \alpha_T\}$ in space \mathcal{H}

³Try it yourself. Prove that \mathcal{H} is closed under vector addition and (real) scalar multiplication.

Inner Product over RKHS \mathcal{H}

For any $g(\cdot) = \sum_{s=1}^S \beta_s K(\cdot, \mathbf{x}'_s) \in \mathcal{H}$ and $h(\cdot) = \sum_{t=1}^T \alpha_t K(\cdot, \mathbf{x}_t) \in \mathcal{H}$, define the inner product⁴

one such element
of \mathcal{H}

another element
of \mathcal{H}

$$\langle g(\cdot), h(\cdot) \rangle = \text{scalar } \dagger = \langle h(\cdot), g(\cdot) \rangle \quad (\text{for reals})$$

$$\langle \theta_1 g_1(\cdot) + \theta_2 g_2(\cdot), h(\cdot) \rangle = \theta_1 \langle g_1(\cdot), h(\cdot) \rangle + \theta_2 \langle g_2(\cdot), h(\cdot) \rangle$$

⁴Again, you can verify that $\langle f, g \rangle$ is indeed an inner product following properties such as symmetry, linearity in the first argument and positive-definiteness:

$$\langle g(\cdot), g(\cdot) \rangle \geq 0$$

4

$$\langle g(\cdot), h(\cdot) \rangle^2 \leq \langle g(\cdot), g(\cdot) \rangle \langle h(\cdot), h(\cdot) \rangle$$

$\langle a, a \rangle$ $\langle b, b \rangle$

$$|a^T b|^2 \leq \|a\|^2 \|b\|^2$$

equality holds iff
a & b are parallel

Ans:

$$\langle g(\cdot), h(\cdot) \rangle = \sum_{t=1}^T \sum_{s=1}^S \beta_s \alpha_t K(x'_s, x_t)$$

Inner Product over RKHS \mathcal{H}

For any $g(\cdot) = \sum_{t=1}^S \beta_s K(\cdot, \mathbf{x}'_s) \in \mathcal{H}$ and $h(\cdot) = \sum_{t=1}^T \alpha_t K(\cdot, \mathbf{x}_t) \in \mathcal{H}$, define the inner product⁴

$$\langle h, g \rangle = \sum_{s=1}^S \beta_s \sum_{t=1}^T \alpha_t K(\mathbf{x}'_s, \mathbf{x}_t)$$

Can be verified to be an inner product (4)

Further simplifying (4),

$$\langle h, g \rangle = \sum_{s=1}^S \beta_s \sum_{t=1}^T \alpha_t K(\mathbf{x}'_s, \mathbf{x}_t) = \sum_{s=1}^S \beta_s h(\mathbf{x}'_s) = \sum_{t=1}^T \alpha_t g(\mathbf{x}_t) \quad (5)$$

One immediately observes that in the special case that $g(\cdot) = K(\cdot, \mathbf{x})$, ($S=1, \beta=1$)

$$\langle h, g \rangle = \beta h(\mathbf{x}) = h(\mathbf{x}) = \sum_{t=1}^T \alpha_t K(\mathbf{x}, \mathbf{x}_t)$$

⁴Again, you can verify that $\langle f, g \rangle$ is indeed an inner product following properties such as symmetry, linearity in the first argument and positive-definiteness:

Inner Product over RKHS \mathcal{H}

For any $g(\cdot) = \sum_{s=1}^S \beta_s K(\cdot, \mathbf{x}'_s) \in \mathcal{H}$ and $h(\cdot) = \sum_{t=1}^T \alpha_t K(\cdot, \mathbf{x}_t) \in \mathcal{H}$, define the inner product⁴

$$\langle h, g \rangle = \sum_{s=1}^S \beta_s \sum_{t=1}^T \alpha_t K(\mathbf{x}'_s, \mathbf{x}_t) \quad (4)$$

Further simplifying (4),

$$\langle h, g \rangle = \sum_{s=1}^S \beta_s \sum_{t=1}^T \alpha_t K(\mathbf{x}'_s, \mathbf{x}_t) = \sum_{s=1}^S \beta_s f(\mathbf{x}_s) \quad (5)$$

One immediately observes that in the special case that $g(\cdot) = K(\cdot, \mathbf{x})$,

[x behaves like a fix pt]

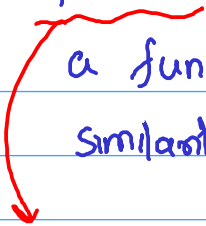
$$\langle h, K(\cdot, \mathbf{x}) \rangle = h(\mathbf{x}) \quad \langle K(\cdot, \mathbf{x}), g \rangle = g(\mathbf{x}) \quad (6)$$

⁴Again, you can verify that $\langle f, g \rangle$ is indeed an inner product following properties such as symmetry, linearity in the first argument and positive-definiteness:

https://en.wikipedia.org/wiki/Inner_product_space

$$\langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{x}') \rangle = K(\mathbf{x}, \mathbf{x}')$$

RKHS lets you represent
a function in a space of
similarity with other pts



Reproducing because $\langle K(\cdot, x), g(\cdot) \rangle = g(x)$

$$\langle h(\cdot), K(\cdot, x) \rangle = h(x)$$

Representer theorem gives us equivalence in
searching in this space!!

Orthogonal Decomposition

Since $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\} \subseteq \mathcal{X}$ and $\mathcal{K} = \{K(\cdot, \mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$ with \mathcal{H} being the set of all finite linear combinations of function in \mathcal{K} , we also have that

vector space \swarrow $\text{lin_span} \{K(\cdot, \mathbf{x}^{(1)}), K(\cdot, \mathbf{x}^{(2)}), \dots, K(\cdot, \mathbf{x}^{(m)})\} \subseteq \mathcal{H}$

Thus, we can use orthogonal projection to

$$f(\cdot) = f_H^{\parallel}(\cdot) + f_H^{\perp}(\cdot)$$

Orthogonal Decomposition

Since $\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots, \mathbf{x}^{(m)}\} \subseteq \mathcal{X}$ and $\mathcal{K} = \{K(., \mathbf{x}) \mid \mathbf{x} \in \mathcal{X}\}$ with \mathcal{H} being the set of all finite linear combinations of function in \mathcal{K} , we also have that

$$\text{lin_span} \left\{ K(., \mathbf{x}^{(1)}), K(., \mathbf{x}^{(2)}), \dots, K(., \mathbf{x}^{(m)}) \right\} \subseteq \mathcal{H}$$

Thus, we can use orthogonal projection to decompose any $h \in \mathcal{H}$ into a sum of two functions, one lying in $\text{lin_span} \{K(., \mathbf{x}^{(1)}), K(., \mathbf{x}^{(2)}), \dots, K(., \mathbf{x}^{(m)})\}$, and the other lying in the orthogonal complement:

$$h = \underbrace{h^{\parallel}} + h^{\perp} = \sum_{i=1}^m \alpha_i K(., \mathbf{x}^{(i)}) + h^{\perp} \quad (7)$$

where $\langle K(., \mathbf{x}^{(i)}), h^{\perp} \rangle = 0$, for each $i = [1..m]$.

For a specific training point $\mathbf{x}^{(j)}$, substituting from (7) into (6) for any $h \in \mathcal{H}$, using the fact that $\langle K(., \mathbf{x}^{(i)}), h^\perp \rangle = 0$

$$\underline{h(\mathbf{x}^{(j)})} = \left\langle \underbrace{\sum_{i=1}^m \alpha_i K(., \mathbf{x}^{(i)})}_{\text{is } 0} + h^\perp, K(., \mathbf{x}^{(j)}) \right\rangle = \sum_{i=1}^m \alpha_i \langle K(., \mathbf{x}^{(i)}), K(., \mathbf{x}^{(j)}) \rangle = \underline{\sum_{i=1}^m \alpha_i K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})} \quad (8)$$

which we observe is independent of h^\perp .

Representing $h(\mathbf{x}^{(j)})$ purely in terms
of components in lin-span

Basis function expansion & Kernel: Additional Discussion 2

Consider regression function $f(\mathbf{x}) = \sum_{j=1}^p w_j \phi_j(\mathbf{x})$ with weight vector \mathbf{w} estimated as

$$\mathbf{w}_{Pen} = \underset{\mathbf{w}}{\operatorname{argmin}} \mathcal{L}(\phi, \mathbf{w}, \mathbf{y}) + \lambda \Omega(\mathbf{w})$$

It can be shown that for $p \in [0, \infty)$, under certain conditions on K , the following can be equivalent representations

•

$$f(\mathbf{x}) = \sum_{j=1}^p w_j \phi_j(\mathbf{x})$$

• And⁵

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i)$$

Representing f in terms of its similarity alone

How about starting with this? [Nonparametric?]

⁵Section 5.8.1 of Tibshi.

Basis function expansion & Kernel: Additional Discussion 2

- We could also begin with (Eg: Nadaraya-Watson kernel regression)

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i \underbrace{K(\mathbf{x}, \mathbf{x}_i)} = \frac{\sum_{i=1}^m y_i k_n(\|\mathbf{x} - \mathbf{x}_i\|)}{\sum_{i=1}^m k_n(\|\mathbf{x} - \mathbf{x}_i\|)}$$

A non-parametric kernel k_n is a non-negative real-valued integrable function

satisfying the following two requirements: $\int_{-\infty}^{+\infty} \underbrace{k_n(u) du}_{\text{behaves like pdf}} = 1$ and $\underbrace{k_n(-u) = k_n(u)}_{\text{symmetric}}$
for all values of u

$$\alpha_i = \frac{y_i}{\sum_i k_n(\|\mathbf{x} - \mathbf{x}_i\|)}$$

$$\underbrace{K(\mathbf{x}, \mathbf{x}_i)} = k_n(\|\mathbf{x} - \mathbf{x}_i\|)$$

Should be independent of all \mathbf{x}_j except \mathbf{x}_i & \mathbf{x}

Basis function expansion & Kernel: Additional Discussion 2


- We could also begin with (Eg: Nadaraya-Watson kernel regression)

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i) = \frac{\sum_{i=1}^m y_i k_n(\|\mathbf{x} - \mathbf{x}_i\|)}{\sum_{i=1}^m k_n(\|\mathbf{x} - \mathbf{x}_i\|)}$$

For KNN $k_n(\cdot)$ this is k

A non-parametric kernel k_n is a non-negative real-valued integrable function satisfying the following two requirements: $\int_{-\infty}^{+\infty} k_n(u) du = 1$ and $k_n(-u) = k_n(u)$ for all values of u

- E.g.: $k_n(x_i - x) = I(\|x_i - x\| \leq \|x_{(k)} - x\|)$ where $x_{(k)}$ is the training observation ranked k^{th} in distance from x and $I(S)$ is the indicator of the set S
- This is precisely the Nearest Neighbor Regression model
- Kernel regression and density models are other examples of such *local regression* methods⁶

This x_i will have $k_n(x_i - x) = 0$  k nearest nbgs of x
This x_i will have $k_n(x_i - x) = 1$

Basis function expansion & Kernel: Additional Discussion 2

- We could also begin with (Eg: Nadaraya-Watson kernel regression)

$$f(\mathbf{x}) = \sum_{i=1}^m \alpha_i K(\mathbf{x}, \mathbf{x}_i) = \frac{\sum_{i=1}^m y_i k_n(\|\mathbf{x} - \mathbf{x}_i\|)}{\sum_{i=1}^m k_n(\|\mathbf{x} - \mathbf{x}_i\|)}$$

A non-parametric kernel k_n is a non-negative real-valued integrable function satisfying the following two requirements: $\int_{-\infty}^{+\infty} k_n(u) du = 1$ and $k_n(-u) = k_n(u)$ for all values of u

- E.g.: $k_n(x_i - x) = I(\|x_i - x\| \leq \|x_{(k)} - x\|)$ where $x_{(k)}$ is the training observation ranked k^{th} in distance from x and $I(S)$ is the indicator of the set S
- This is precisely the Nearest Neighbor Regression model
- Kernel regression and density models are other examples of such *local regression* methods⁶
- The broader class - **Non-Parametric Regression: $y = g(\mathbf{x}) + \epsilon$** where functional form of $g(\mathbf{x})$ is not fixed

⁶Section 2.8.2 of Tibshirani

Non-parametric Kernel weighted (Local Linear) Regression: Tut 5, Prob 3

$(w', b') = \text{fn of query pt } x'$

Given $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_i, y_i), \dots, (\mathbf{x}_n, y_n)\}$, predict $f(\mathbf{x}') = (w'^T \phi(\mathbf{x}') + b)$ for each test (or query point) \mathbf{x}' as:

$$(w', b') = \underset{w, b}{\operatorname{argmin}} \sum_{i=1}^n K(\mathbf{x}', \mathbf{x}_i) \left(y_i - (w^T \phi(\mathbf{x}_i) + b) \right)^2$$

\downarrow
query pt

\swarrow
 $w'_i = w$ for i^{th} kg pt

\swarrow
 x'_i

- 1 If there is a closed form expression for (w', b') and therefore for $f(x')$ in terms of the known quantities, derive it.
 \swarrow in terms of x'
- 2 How does this model compare with linear regression and k -nearest neighbor regression? What are the relative advantages and disadvantages of this model?
- 3 In the one dimensional case (that is when $\phi(x) \in \mathbb{R}$), graphically try and interpret what this regression model would look like, say when $K(.,.)$ is the linear kernel⁷.

⁷Hint: What would the regression function look like at each training data point?

Answer to Question 1

The weighing factor $r_i^{x'}$ of each training data point (\mathbf{x}_i, y_i) is now also a function of the query or test data point $(\mathbf{x}', ?)$, so that we write it as $r_i^{x'} = K(\mathbf{x}', \mathbf{x}_i)$ for $i = 1, \dots, m$. Let $r_{m+1}^{x'} = 1$ and let R be an $(m+1) \times (m+1)$ diagonal matrix of $r_1^{x'}, r_2^{x'}, \dots, r_{m+1}^{x'}$.

$$R' = \begin{bmatrix} r_1^{x'} & 0 & \dots & 0 \\ 0 & r_2^{x'} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & r_{m+1}^{x'} \end{bmatrix}$$

$(w', b') = \operatorname{argmin}_i \sum_i r_i^{x'} (w'^T \phi(x_i) + b' - y_i)^2$

$$\Phi = \begin{bmatrix} \phi_1(x_1) & \dots & \phi_p(x_1) \\ \dots & \dots & \dots \\ \phi_1(x_m) & \dots & \phi_p(x_m) \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

$(w', b') = \operatorname{argmin} \|R'(\Phi w' - y)\|$

Further, let

and

Answer to Question 1 (contd.)

$$\hat{\mathbf{w}} = \begin{bmatrix} w_1 \\ \vdots \\ w_p \\ b \end{bmatrix}$$

and

$$\mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$$

The sum-square error function then becomes

$$\frac{1}{2} \sum_{i=1}^m r_i (y_i - (\hat{\mathbf{w}}^T \phi(x_i) + b))^2 = \frac{1}{2} \|\sqrt{R}\mathbf{y} - \sqrt{R}\Phi\hat{\mathbf{w}}\|_2^2$$

where \sqrt{R} is a diagonal matrix such that each diagonal element of \sqrt{R} is the square root of the corresponding element of R .

Answer to Question 1 (contd.)

The sum-square error function:

$$\frac{1}{2} \sum_{i=1}^m r_i (y_i - (\hat{\mathbf{w}}^T \phi(x_i) + b))^2 = \frac{1}{2} \|\sqrt{R}\mathbf{y} - \sqrt{R}\Phi\hat{\mathbf{w}}\|_2^2$$

weighing for each prediction y_i & each feature $\phi(x_i)$

This convex function has a global minimum at $\hat{\mathbf{w}}_*^{x'}$ such that

$$\hat{\mathbf{w}}_*^{x'} = (\Phi^T R \Phi)^{-1} \Phi^T R \mathbf{y}$$

This is referred to as local linear regression (Section 6.1.1 of Tibshi).

Non-parametric: Because form of $f(\cdot)$ is not fixed across all query pts
params/shape

Answer to Question 2

Avoids overfitting to an extent
Advantage

closer as kernel
pts $K(x_i, x_j)$

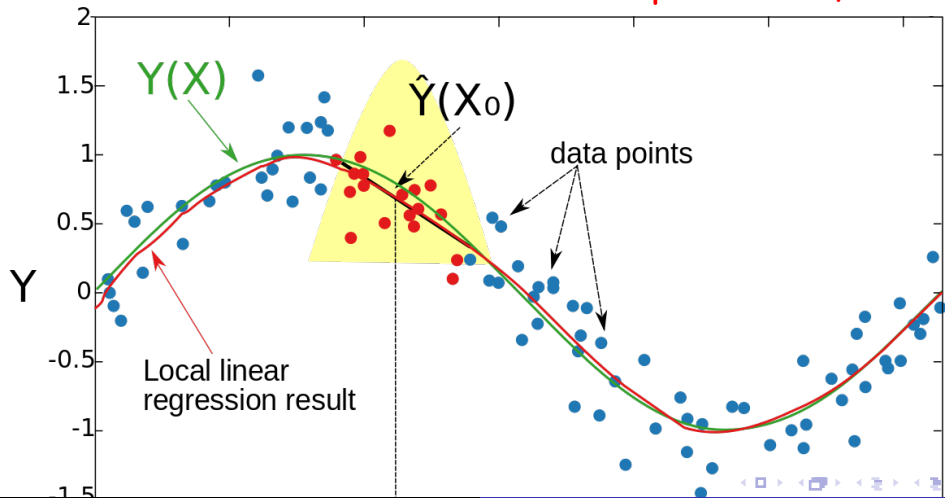
$\phi^T \phi \Rightarrow$ Equal imp to all components
 $(1 + \phi^T \phi)^2 \Rightarrow$ Similar components given more imp.

- 1 Local linear regression gives more importance (than linear regression) to points in \mathcal{D} that are closer/similar to x' and less importance to points that are less similar.
- 2 Important if the regression curve is supposed to take different shapes in different parts of the space.
- 3 Local linear regression comes close to k-nearest neighbor. But unlike k-nearest neighbor, local linear regression gives you a smooth solution

Smother than knn

Answer to Question 3

Curvature of "red" will increase for RBF kernel!



THIS IS NOT PART OF MIDSEM

Gaussian Process Regression

↓
There is a distribution
But yet non-parametric!