# VOLTERRA SERIES & KERNEL REGRESSION

Prof. Alexander Huth
10/26/2017

# NONLINEAR PROBLEM

| | | | | |
|---|---|---|---|---|
| **x1** | 0 | 1 | 1 | 0 |
| **x2** | 0 | 0 | 1 | 1 |
| **y** | 0 | 0 | 1 | 0 |

$$y = f(x_1, x_2)$$

what is $f$?

# VOLTERRA SERIES

y = \sum_{n=1}^P\sum_{\tau_1
^p\cdots \sum_{\tau_n=1}^p
h_n(\tau_1, \dots, \tau_n) \prod
^n x_j

* For a finite Volterra series of order $P$, consider every nonlinear combination of up to $P$ variables

$$y = \sum_{n=1}^{P} \sum_{\tau_1=1}^{p} \cdots \sum_{\tau_n=1}^{p} h_n(\tau_1, \dots, \tau_n) \prod_{j=1}^{n} x_j$$

# VOLTERRA SERIES

y = h_{1,0} x_1 + h_{0,1} x_2 +
x_1 x_2 + h_{2,0} x_1^2 + h_{
x_2^2 + h_{2,2} x_1^2 x_2^2

* A finite Volterra series of order $P$ considers every nonlinear combination of up to $P$ variables

$$y = h_{1,0}x_1 + h_{0,1}x_2 + h_{1,1}x_1x_2 + h_{2,0}x_1^2 + h_{0,2}x_2^2 + h_{2,2}x_1^2x_2^2 + \dots$$

# VOLTERRA SOLUTION!

| | | | | |
|---|---|---|---|---|
| **x1** | 0 | 1 | 1 | 0 |
| **x2** | 0 | 0 | 1 | 1 |
| **y** | 0 | 0 | 1 | 0 |

$$y = f(x_1, x_2)$$

$$y = h_{1,0}x_1 + h_{0,1}x_2 + h_{1,1}x_1x_2 + h_{2,0}x_1^2 + h_{0,2}x_2^2 + h_{2,2}x_1^2x_2^2 + \ldots$$

$$h_{1,1} = 1, h_{i,j} = 0 \text{ for all other } i, j$$

# VOLTERRA SERIES

* (btw, Volterra series is just a different linearized model…)

* (but it’s one that can capture any nonlinear function!)

# VOLTERRA SERIES

* Volterra series have ***nightmarish*** numbers of parameters

* Suppose X's are 16x16 image patches (i.e. p=256)

* How many coefficients (h's) are there in a 5th-order Volterra model? (~1 billion!)

# KERNEL REGRESSION

***FORGET FEATURES,***
***USE SAMPLES!***

* *Please do not actually forget features*

# KERNEL REGRESSION

* Let’s say the y for a new sample is some a combination of the y’s from old samples
* *Example:* image patches

# KERNEL REGRESSION

* **Kernel function:** $k(a, b) = \phi(a)^\top \phi(b)$

  tells you how similar a and b are in some "Reproducing kernel Hilbert space", $H$

# KERNEL REGRESSION

\hat{f} = \underset{f\in \mathcal{H}}{\mbox{argmin}}\left[ ||Y-f(X)||_2^2 + \lambda ||f||_\mathcal{H}^2 \right]

* **Representer theorem:**

$$\hat{f} = \underset{f\in \mathcal{H}}{\mbox{argmin}}\left[ ||Y-f(X)||_2^2 + \lambda ||f||_\mathcal{H}^2 \right]$$

then: $$\hat{f}(z) = \sum_{i=1}^{n} \alpha_i k(z, X_i)$$

i.e. the function value for a new datapoint, z, is a linear combination (with weights alpha) of the kernel similarities between z and existing datapoints in X

# KERNEL REGRESSION

\hat{\alpha} = \underset{\alpha}{\mbox{argmin}} \left[ ||Y - K \alpha||_2^2 + \lambda \alpha^\top K \alpha \right]

* How do we find the alphas?

$$\hat{\alpha} = \underset{\alpha}{\mbox{argmin}} \left[ ||Y - K \alpha||_2^2 + \lambda \alpha^\top K \alpha \right]$$

where: $K_{ij} = k(X_i, X_j)$

# KERNEL REGRESSION

* How do we find the alphas?

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

(this is called ***KERNEL RIDGE REGRESSION***)

# KERNEL REGRESSION

* Ok fine. But what the heck is $k$?!?

* **Possibility 1:** linear kernel!

$$k(a, b) = a^\top b$$

# KERNEL REGRESSION

* **Possibility 1:** linear kernel!

*remember:* $\hat{f}(z) = \sum_{i=1}^{n} \alpha_i k(z, X_i)$

\Rightarrow \hat\alpha = (X X^\top
\lambda I)^{-1}Y

\Rightarrow \hat{f}(z) = z X^\top
\hat\alpha = z X^\top (XX^\top +

$$k(a,b) = a^\top b \qquad \Rightarrow K = XX^\top$$

$$\Rightarrow \hat{\alpha} = (XX^\top + \lambda I)^{-1} Y$$

$$\Rightarrow \hat{f}(z) = zX^\top \hat{\alpha} = zX^\top (XX^\top + \lambda I)^{-1} Y$$

# KERNEL REGRESSION

* **Possibility 1:** linear kernel!

*remember:* $$\hat{f}(z) = \sum_{i=1}^{n} \alpha_i k(z, X_i)$$

\Rightarrow \hat\alpha = (X X^\top
\lambda I)^{-1}Y

\Rightarrow \hat{f}(z) = z X^\top
\hat\alpha = z X^\top (XX^\top +

$$k(a,b) = a^\top b \qquad \Rightarrow K = XX^\top$$

$$\Rightarrow \hat{\alpha} = (XX^\top + \lambda I)^{-1} Y$$

$$\Rightarrow \hat{f}(z) = zX^\top \hat{\alpha} = zX^\top (XX^\top + \lambda I)^{-1} Y$$

*what if we just called this part “beta”?*

# KERNEL REGRESSION

\phi_p(x) = (x_1, x_2, x_1 x_2,
x_1^p x_2^p)

* **Possibility 2:** $\phi_p(x) = (x_1, x_2, x_1 x_2, \ldots, x_1^p x_2^p)$

*remember:* $k(a, b) = \phi(a)^\top \phi(b)$

# KERNEL REGRESSION

* **Possibility 2:** $\phi_p(x) = (x_1, x_2, x_1 x_2, \ldots, x_1^p x_2^p)$

*remember:* $k(a,b) = \phi(a)^\top \phi(b)$

***Volterra series model!***
***But with only n parameters!***

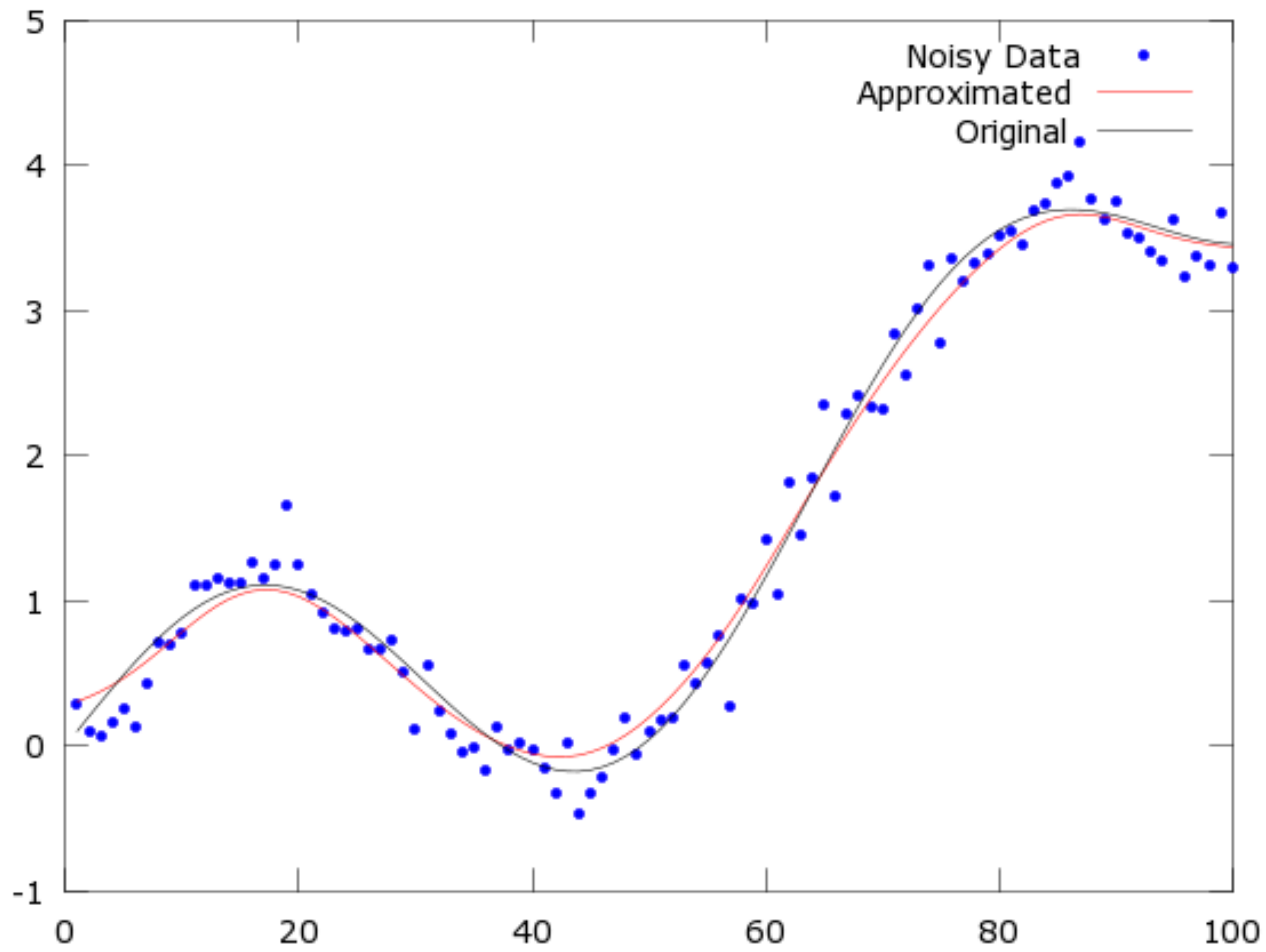
# KERNEL REGRESSION

* **Possibility 3:** Radial basis function (RBF)

$$k(a,b) = e^{-||a-b||_2^2/(2\sigma^2)}$$

# KERNEL REGRESSION

* **Possibility 3:** Radial basis function (RBF)



# KERNEL EFFICIENCY

* What’s the cost of kernel regression?

* Compare kernel ridge vs. non-kernel ridge

# PROJECT!

* Goal is to apply or explore something / anything we’ve talked about in this class

  * could be using real data (e.g. fit some kind of model to this dataset)

  * could be theory/methods (e.g. find a better way to do something)

# PROJECT!

* Proposal due **next Thursday** (11/2):
  * ~1-2 paragraphs describing what you plan to do. Email to huth@cs.utexas.edu before class
* Writeup (3-4 pages explaining background & what you did) due Dec. 5
* In-class presentations (5-10 minutes) Dec. 5 & 7

# PROJECT!

* Sources of data (among many more):
  * CRCNS: https://crcns.org/data-sets
  * Allen Inst.: http://www.brain-map.org
  * Study Forrest: http://studyforrest.org/

# NEXT TIME

* Neural networks! (well, at least perceptrons)