ECE595 / STAT598: Machine Learning I
Lecture 3.2: Regression with Kernels - Examples

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

Mathematical Background

- Lecture 1: Linear regression: A basic data analytic tool
- Lecture 2: Regularization: Constraining the solution
- Lecture 3: Kernel Method: Enabling nonlinearity

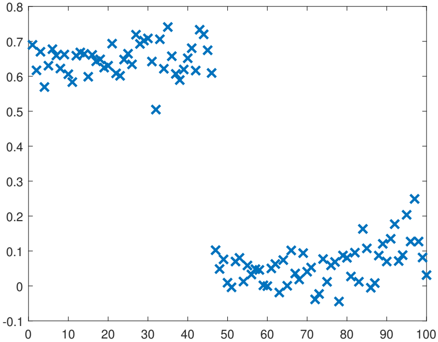
Lecture 3: Kernel Method

- Kernel Method
  - Dual Form
  - Kernel Trick
  - Algorithm
- Examples
  - Radial Basis Function (RBF)
  - Regression using RBF
  - Kernel Methods in Classification

Example

Goal: Use the kernel method to fit the data points shown as follows.

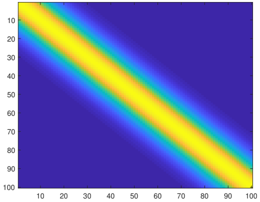
- What is the input feature vector $\boldsymbol{x}^n$? $\boldsymbol{x}^n = t_n$: The time stamps.
- What is the output y_n? y^n is the height.
- Which kernel to choose? Let us consider the RBF.



Example (using RBF)

- Define the fitted function as $g_{\boldsymbol{\theta}}(t)$. [Here, $\boldsymbol{\theta}$ refers to $\boldsymbol{\alpha}$.]
- The RBF is defined as $k(t_i, t_j) = \exp\{-(t_i - t_j)^2/2\sigma^2\}$, for some σ.
- The matrix $\boldsymbol{K}$ looks something below

$$[\boldsymbol{K}]_{ij} = \exp\{-(t_i - t_j)^2/2\sigma^2\}.$$



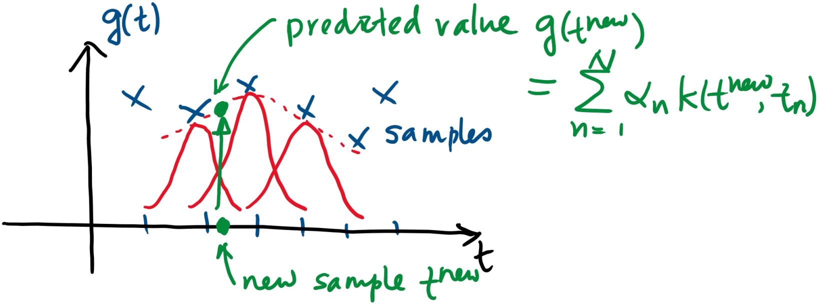
- $\boldsymbol{K}$ is a banded diagonal matrix. (Why?)
- The coefficient vector is $\alpha_n = [(\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}]_n$.

Example (using RBF)

- Using the RBF, the predicted value is given by

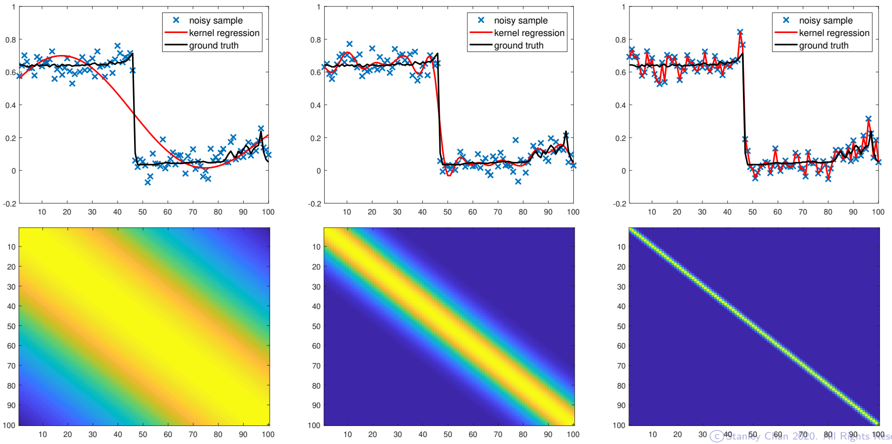
$$g_{\theta}(t^{\text{new}}) = \sum_{n=1}^{N} \alpha_n k(t^{\text{new}}, t_n) = \sum_{n=1}^{N} \alpha_n e^{-\frac{(t^{\text{new}} - t_n)^2}{2\sigma^2}}.$$

- Pictorially, the predicted function g_{θ} can be viewed as the linear combination of the Gaussian kernels.



Effect of σ

- Large σ: Flat kernel. Over-smoothing.
- Small σ: Narrow kernel. Under-smoothing.
- Below shows an example of the fitting and the kernel matrix $\boldsymbol{K}$.



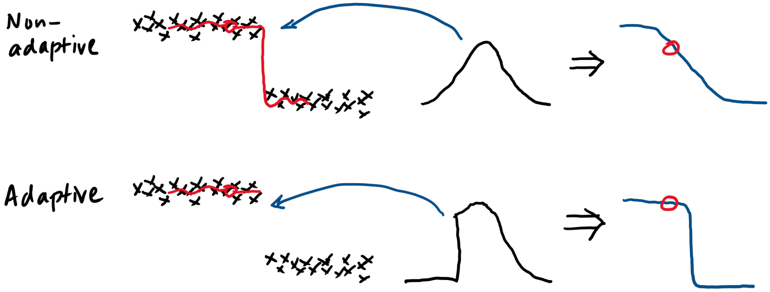
Too large | About right | Too small

Any Improvement?

- We can improve the above kernel by considering $\boldsymbol{x}^n = [y_n,\ t_n]^T$.
- Define the kernel as

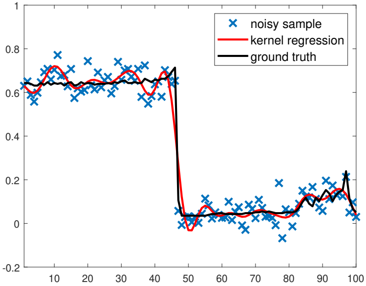
$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \exp\left\{ -\left(\frac{(y_i - y_j)^2}{2\sigma_r^2} + \frac{(t_i - t_j)^2}{2\sigma_s^2} \right) \right\}.$$

- This new kernel is adaptive (**edge-aware**).

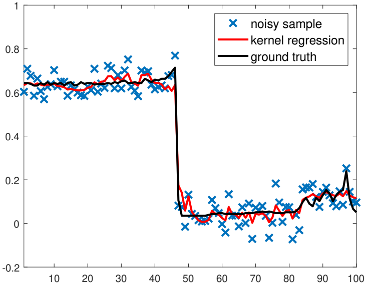


Any Improvement?

Here is a comparison.



without improvement

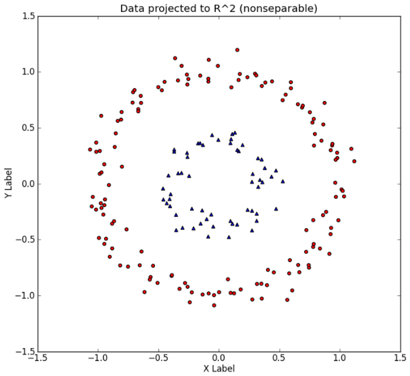


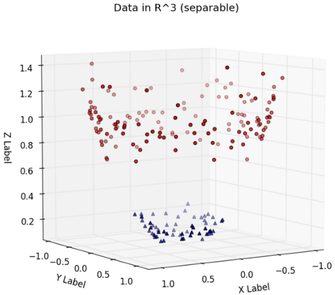
with improvement

- This idea is known as **bilateral filter** in the computer vision literature.
- Can be further extended to 2D image where $\boldsymbol{x}^n = [y_n, \boldsymbol{s}_n]$, for some spatial coordinate $\boldsymbol{s}_n$.
- Many applications. See Reading List.

Kernel Methods in Classification

- The concept of lifting the data to higher dimension is useful for classification. [1]



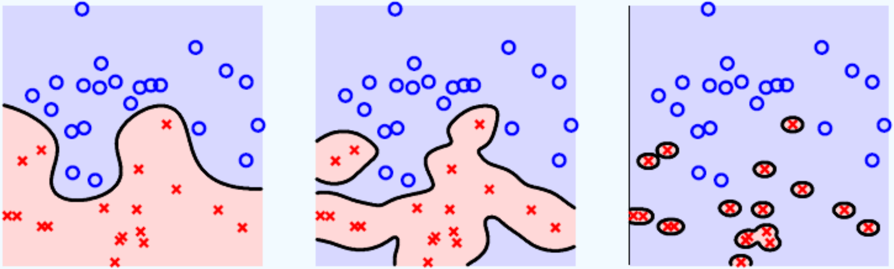


[1] Image source: https://towardsdatascience.com/understanding-the-kernel-trick-e0bc6112ef78

Kernels in Support Vector Machines

Example. RBF for SVM (We will discuss SVM later in the semester.)

- Radial Basis Function is often used in support vector machine.
- Poor choice of parameter can lead to low training loss, but with the risk of over-fit.
- Under-fitted data can sometimes give better generalization.



$\exp(-1\|\mathbf{x}-\mathbf{x}'\|^2)$ $\exp(-10\|\mathbf{x}-\mathbf{x}'\|^2)$ $\exp(-100\|\mathbf{x}-\mathbf{x}'\|^2)$

Reading List

Kernel Method:

- Learning from Data (Chapter 3.4)
  https://work.caltech.edu/telecourse
- CMU 10701 Lecture 4 https://www.cs.cmu.edu/~tom/10701_sp11/slides/Kernels_SVM_04_7_2011-ann.pdf
- Berkeley CS 194 Lecture 7 https://people.eecs.berkeley.edu/~russell/classes/cs194/
- Oxford C19 Lecture 3
  http://www.robots.ox.ac.uk/~az/lectures/ml/lect3.pdf

Kernel Regression in Computer Vision:

- Bilateral Filter https://people.csail.mit.edu/sparis/bf_course/course_notes.pdf
- Takeda and Milanfar, "Kernel regression for image processing and reconstruction", IEEE Trans. Image Process. (2007)
  https://ieeexplore.ieee.org/document/4060955