Lecture 16: Kernel Ridge Regression

In the last lecture we introduced the idea of a kernel & the associated RKHS.

- A function $k: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a kernel. If it is non-negative definite & symm. then we have

$$k(\underline{x}, \underline{x}') = \sum_{j=0}^{\infty} F_j(\underline{x}) F_j(\underline{x}')$$

where $F_j : \mathbb{R}^N \to \mathbb{R}$ are the features of k.

- The space $\mathcal{H}_k$, RKHS of k is defined as all $f: \mathbb{R}^d \to \mathbb{R}$ of the form $f(\underline{x}) = \sum_{j=0}^{\infty} c_j F_j(\underline{x})$ s.t $\sum_{j=0}^{\infty} c_j < +\infty$.
- Equivalently such f has the form $f(\underline{x}) = \sum_{j=0}^{\infty} a_j k(\underline{x}_j, \underline{x})$ for some a_j & $\underline{x}_j$.

we also looked at a simple kernel interpolation problem. The goal of this lecture is to extend the kernel methods to ridge regression.

16.1 Kernel Ridge Regression

Recall our original approach to ridge

was to solve problems of the form

(RR) $$\underset{\beta}{\text{minimize}} \; \|A\beta - Y\|^2 + \lambda \|\beta\|^2$$

where A was our feature matrix often of the form

(Prev. we used ψ_j instead of F_j in context of Ridge reg.)

$$A = \begin{bmatrix} F_0(\underline{x}_0) & F_1(\underline{x}_0) & \cdots \\ F_0(\underline{x}_1) & F_1(\underline{x}_1) & \cdots \\ \vdots & & \\ F_0(\underline{x}_{N-1}) & F_1(\underline{x}_{N-1}) & \cdots \end{bmatrix}$$

* essentially this meant that we would pick the features $F_j : \mathbb{R}^d \to \mathbb{R}$ leading to a model of the form

(*) $$f(\underline{x}) = \sum_{j=0}^{J-1} \beta_j F_j(\underline{x}) \approx y(\underline{x})$$

Our knowledge of Kernels now reveals to us that underneath this model is the kernel

(**) $$K(\underline{x}, \underline{x}') = \sum_{j=0}^{J-1} F_j(\underline{x}) F_j(\underline{x}')$$

with features $\psi_j : \mathbb{R}^d \to \mathbb{R}$ & whose RKHS,

$$\mathcal{H}_K := \{ f : \mathbb{R}^d \to \mathbb{R} \mid f(\underline{x}) \text{ has form } (*) \}$$

and

$$\|f\|_{\mathcal{H}_K}^2 = \sum_{j=0}^{J-1} \beta_j^2 = \|\beta\|^2$$

In otherwords (RR) is equivalent to the problem

(KR) $\underset{f \in \mathcal{H}_K}{\text{minimize}} \; \|f(X) - Y\|^2 + \lambda \|f\|^2_{\mathcal{H}_K}$

with $\mathcal{H}_K$ induced by the kernel K above.

Observe, this (KR) problem is now defined for functions $f \in \mathcal{H}_K$ & not just vectors of parameters β.

Problem (KR) is the abstract formulation of Kernel Ridge (KR) Regression.

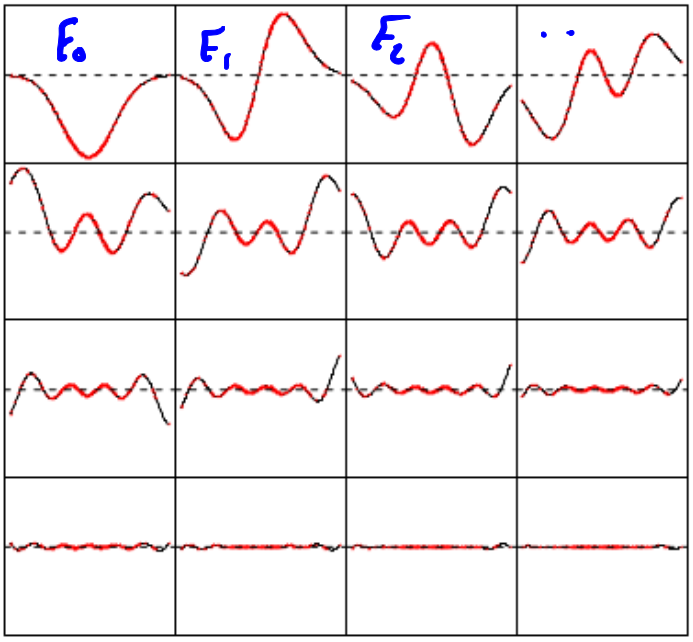
Let us now consider kernels that, unlike (***) do not have finitely many features.

ex. the Gaussian / RBF / Squared exponential kernel

$$k(\underline{x}, \underline{x}') = \exp\left(-\gamma \|\underline{x} - \underline{x}'\|^2\right)$$

$$= \exp\left(-\frac{\|\underline{x} - \underline{x}'\|^2}{2\sigma^2}\right)$$

$$= \sum_{j=0}^{\infty} F_j(\underline{x}) F_j(\underline{x}')$$



The same ideas that were applied to the interpolation problem last week still hold in the setting of KR. ie, the problem (KR) can be solved via a simple finite dimensional problem, even when K has infinitely many features.

This is a fundamental result in Kernel methods called a "Representer theorem".

Thm (Representer)

A function $\hat{f} \in \mathcal{H}_K$ is a minimizer of

$$\underset{f \in \mathcal{H}_K}{\text{minimize}} \; \|f(X) - Y\|^2 + \lambda \|f\|_{\mathcal{H}_K}^2$$

if & only if

$$\hat{f} = \sum_{n=0}^{N-1} \hat{a}_n K(\underline{x}_n, \underline{x})$$

with $\hat{\underline{a}} = (a_0, \dots, a_{N-1}) \in \mathbb{R}^N$ being a minimizer of

$$\underset{\underline{a} \in \mathbb{R}^N}{\text{minimize}} \; \|\Theta \underline{a} - Y\|^2 + \lambda \underline{a}^T \Theta \underline{a}$$

where we recall $\Theta_{ij} = K(\underline{x}_i, \underline{x}_j)$

Similar to the case of penalized least squares this problem has an exact solution

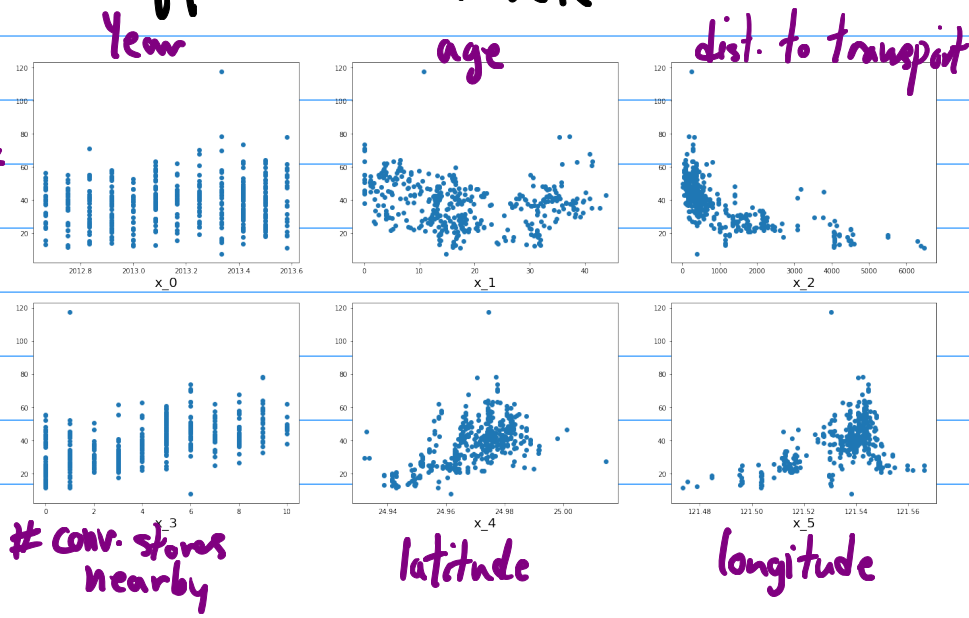
$$\hat{\underline{a}} = (\Theta + \lambda I)^{-1} Y.$$

- The representer thm is a profound result & it is the reason why kernel methods are so prevalent.
- Even though K might have infinitely many features (ex Gaussian kernel) Rep. Thm. tells us that the minimizer $\hat{f}$ is still identified by a finite-dimensional problem for $\hat{\underline{a}} \in \mathbb{R}^N$.

16.2 Kernel Ridg. Reg demo

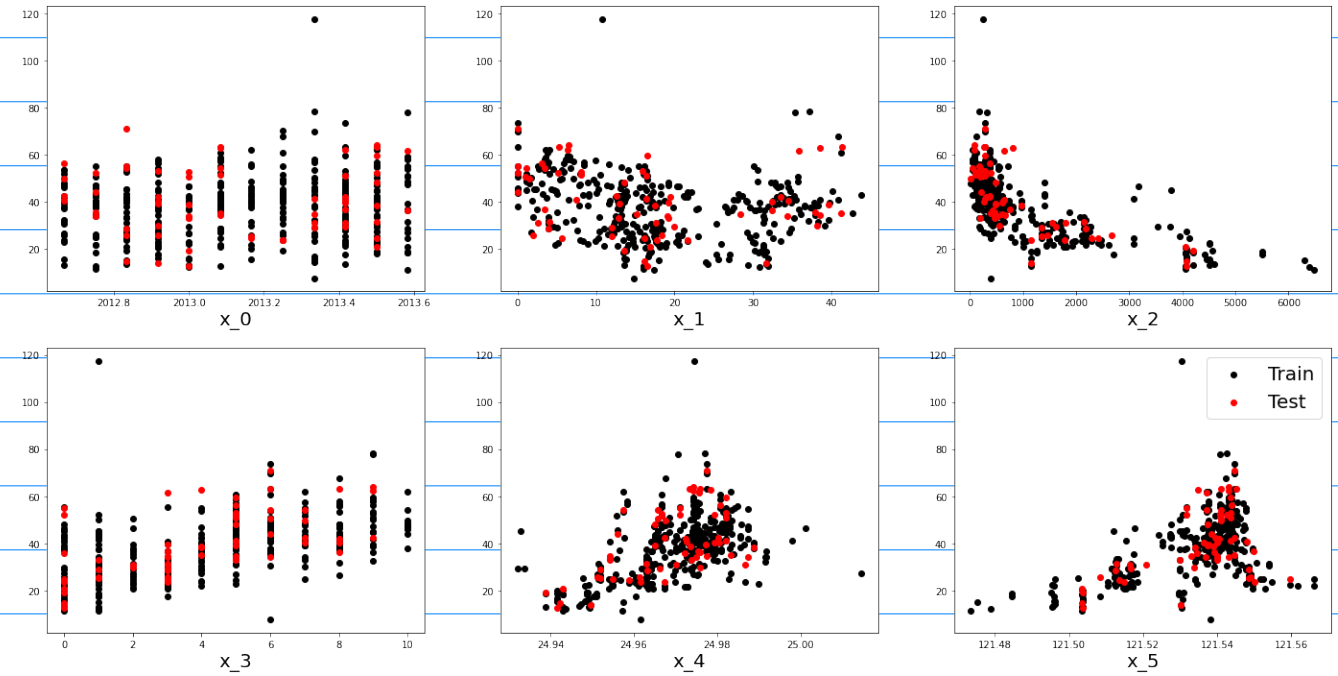
Before proceeding further with theoretical discussions let us consider an application of KR.

Housing prices in Taiwan from UCI Repo. Real state Valuation dataset



Goal: Find $\hat{f}(x_0, \ldots, x_5)$ that predicts the price of a house on the market.

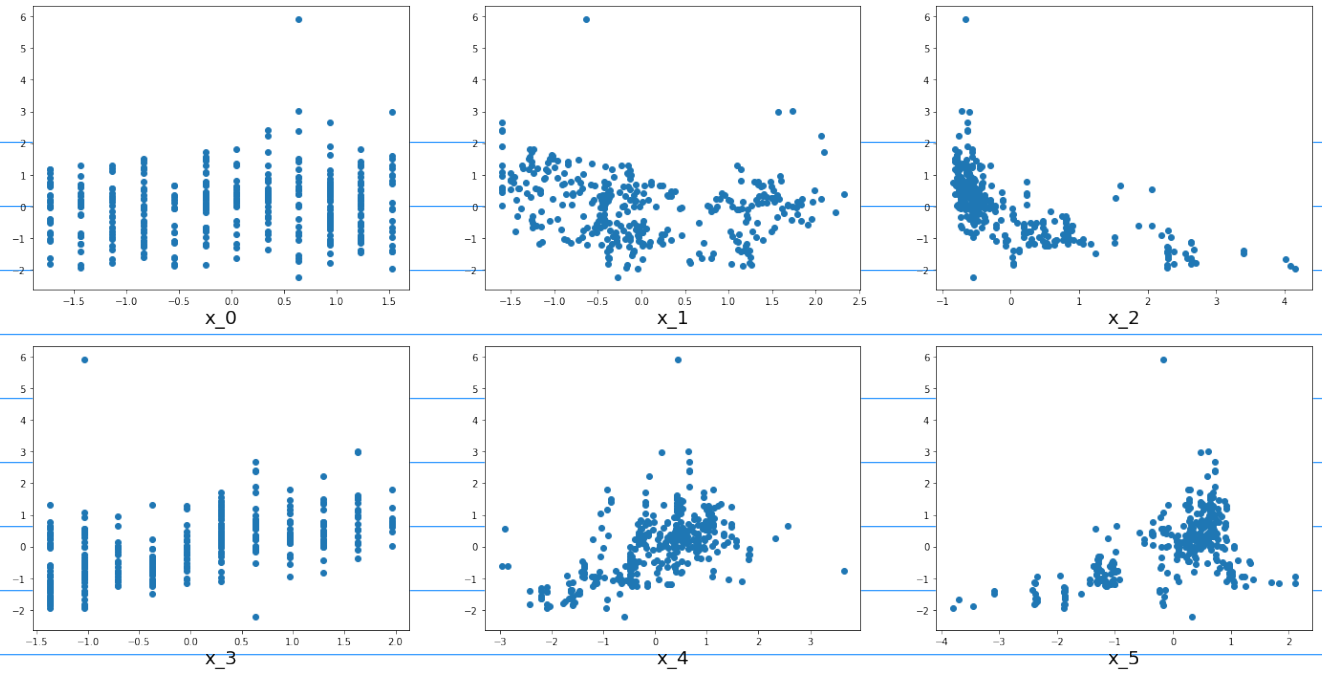
First thing we do is to split the data set ($N = 414$ instances) into train & test sets. Luckily Sklearn has a function that does this for us.



Only the black pts will be used for train (finding $\hat{\underline{a}}$).

Next we normalize & center the training data set so that each column of X, ie, $\underline{x}_j$ have mean 0 & standard deviation 1. This is not crucial in theory but is an important step in practice!

Note, for ex., that $x_0 \approx 2000$'s while $x_3 \in (0, 10)$ while $y \in (0, 120)$. This makes it hard to equally weight the $\underline{x}_j$.



We are now ready to train our KR modeling our normalized training set.

We have some parameters to choose:

(a) kernel $K(\underline{x}, \underline{x}')$ & its possible parameters.

(b) The penalty param. λ.

SKlearn accomodates various kernels:

'linear' $K(\underline{x}, \underline{x}') = \underline{x}^T \underline{x}'$

"poly" $K(\underline{x}, \underline{x}') = (\gamma \underline{x}^T \underline{x}' + c_0)^d$

"sigmoid" $K(\underline{x}, \underline{x}') = \tanh(\gamma \underline{x}^T \underline{x}' + c_0)$

"rbf" $K(\underline{x}, \underline{x}') = \exp(-\gamma \|\underline{x} - \underline{x}'\|^2)$

"Laplace" $K(\underline{x}, \underline{x}') = \exp(-\gamma \|\underline{x} - \underline{x}'\|)$

You can define your own kernel as well but we will use RBF(Gaussian). We have two parameters to pass to SKlearn, γ & λ.

γ: lengthscale of kernel ↑

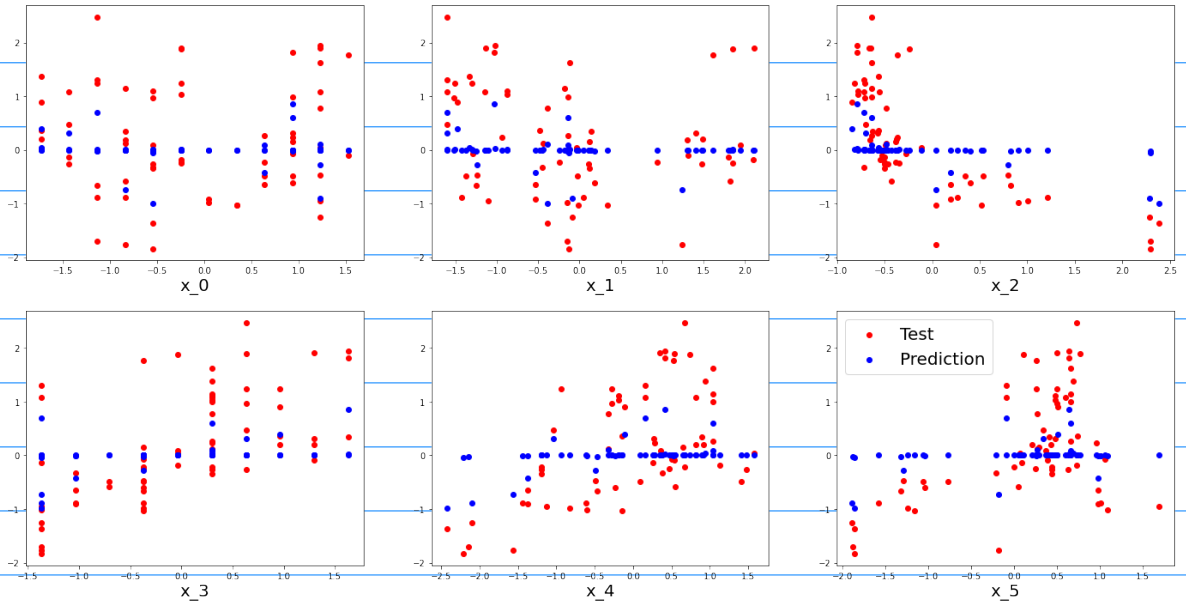
λ: reg. parameter.

Instead of working with γ I prefer to define $\sigma = \sqrt{2\gamma}$ so that the kernel looks like $\exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$, a Gaussian.

The choice of σ & λ has a profound impact!

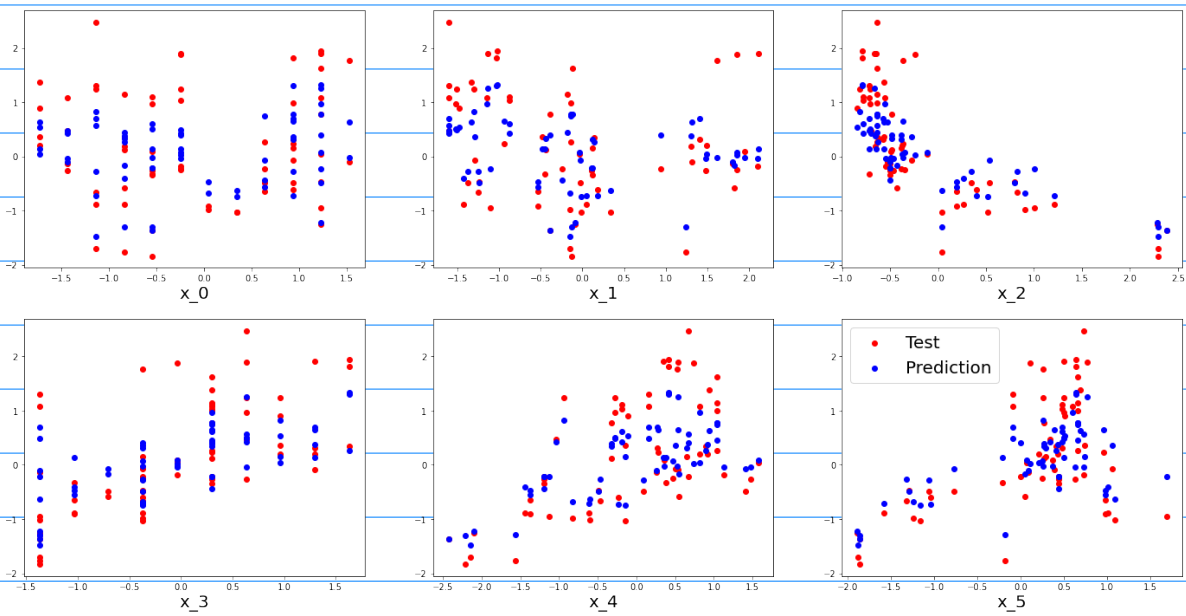
$\sigma = 0.1$

Bad results



$\sigma = 0.5$

Good!



So, we need a strategy to pick σ & λ. We will use 5-fold CV. But since we have two parameters, σ & λ there are no automatic functions to give us the optimal choice. This is where your job as a data scientist is to make an informed decision.

optimal region