ỨNG DỤNG KERNEL TRONG REGRESSION

KERNEL RIDGE REGRESSION

- Hồi quy sườn núi (KRR) kết hợp hồi quy sườn núi (bình phương nhỏ nhất tuyến tính với chính quy chuẩn l2) với thủ thuật kernel . Do đó, một hàm tuyến tính trong không gian được tạo ra bởi hạt nhân và dữ liệu tương ứng.

THIẾT LẬP CÔNG THỨC

- Từ công thức hồi quy sườn núi: $\widehat{\beta}_{\text{ridge}} = (X'X + kI_p)^{-1}X'y$
- Với $\mathbf{\Phi} = [\varphi(\mathbf{x}_1), \ldots, \varphi(\mathbf{x}_N)]^\top$ là ma trận N*K ta có $\hat{\boldsymbol{\beta}} = (\lambda \mathbf{I}_K + \mathbf{\Phi}^\top \mathbf{\Phi})^{-1} \mathbf{\Phi}^\top \mathbf{y}.$
- Nhưng với k có thể là vô cùng , nên ta sử dụng Woodbury matrix identity để viết lại:

$$\hat{\boldsymbol{\beta}} = \mathbf{\Phi}(\lambda \mathbf{I}_N + \mathbf{\Phi}\mathbf{\Phi}^\top)^{-1} \mathbf{y}.$$

- Do đó ta có thể cung cấp 1 điểm mới x là điểm y là dự đoan:

$$y \equiv [\varphi(\mathbf{x})]^\top \boldsymbol{\beta} = [\varphi(\mathbf{x})]^\top (\lambda \mathbf{I}_N + \mathbf{\Phi}\mathbf{\Phi}^\top)^{-1} \mathbf{y}.$$

- Ta có thể viết lại là:

$$y = [\mathbf{k}(\mathbf{x})]^\top (\lambda \mathbf{I}_N + \mathbf{K})^{-1} \mathbf{y}$$

- Sử dụng “ kernel trick” đê tìm và đánh giá k qua dữ liệu đầu vào

MỘT SỐ ƯU ĐIỂM SO VỚI LEAST SQUARES

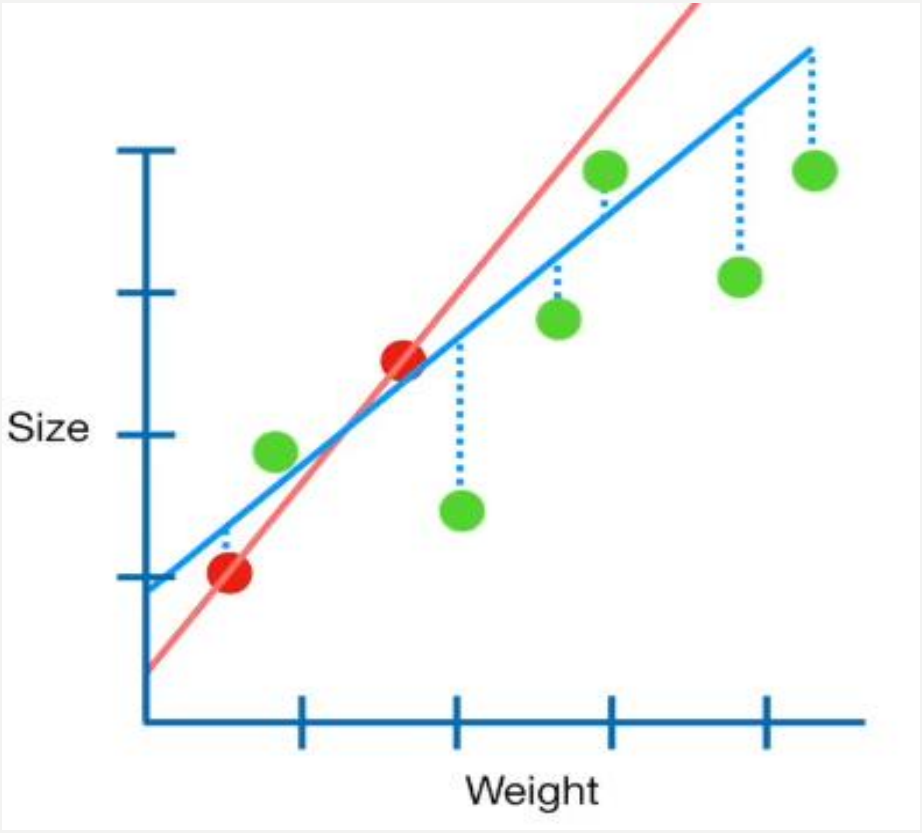
- Hồi quy bình phương tối thiểu hoàn toàn không được xác định khi số lượng yếu tố dự đoán vượt quá số lượng quan sát.
- Nó không phân biệt các yếu tố dự báo “quan trọng” với các yếu tố dự báo “ít quan trọng hơn” trong một mô hình, vì vậy nó bao gồm tất cả chúng. Quy hồi bình phương cũng có các vấn đề liên quan đến đa cộng tuyến trong dữ liệu. **Hồi quy Ridge tránh được tất cả những vấn đề này.**

- Hồi quy OLS sử dụng công thức sau để ước tính các hệ số

$$\hat{\underline{B}} = (X'X)^{-1}X'\underline{Y}$$

- Hồi quy Ridge thêm một *tham số ridge* (k), của ma trận nhận dạng vào ma trận tích chéo, tạo thành một ma trận mới (X`X + kI)

$$\tilde{\underline{B}} = (X'X + kI)^{-1}X'\underline{Y}$$



GAUSSIAN PROCESS REGRESSION

- Gaussian (GP) là một phương pháp học tập có giám sát chung được thiết kế để giải quyết các vấn đề hồi quy và phân loại theo xác suất .

THIẾT LẬP CÔNG THỨC

- Đối với Gaussian ta quan nhiễu với mô hình $y_n = f(\mathbf{x}_n) + \varepsilon_n,$
- Với $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$ và f là một biến ngẫu nhiên ta có:

$$NS \sim \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x})).$$

- m là một hàm trung binh và k là một hạt nhân xác định tích cực,nếu ta sử dụng f là $[f(\mathbf{x}_1), \ldots, f(\mathbf{x}_N)]^\top$ sau đó ta có mô hình GP đầy đủ:

$$\begin{bmatrix} \mathbf{f}_* \\ \mathbf{f} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{X}_*) \\ m(\mathbf{X}) \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}_*, \mathbf{X}_*) & K(\mathbf{X}_*, \mathbf{X}) \\ K(\mathbf{X}, \mathbf{X}_*) & \sigma^2\mathbf{I} + K(\mathbf{X}, \mathbf{X}) \end{bmatrix} \right).$$

Sử dụng thuộc tinh của phân phối Gaussian ta dễ dàng tìm được giá trị trung bình và phương sai của f.

$$\begin{aligned} \mathbb{E}[\mathbf{f}_*] &= K(\mathbf{X}_*, \mathbf{X})[\sigma^2\mathbf{I} + K(\mathbf{X}, \mathbf{X})]^{-1}\mathbf{y} \\ \mathrm{Cov}(\mathbf{f}_*) &= K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})[\sigma^2\mathbf{I} + K(\mathbf{X}, \mathbf{X})]^{-1}K(\mathbf{X}, \mathbf{X}_*)). \end{aligned}$$

MỘI SỐ ƯU ĐIỂM

- Dự đoán nội suy các quan sát (ít nhất là đối với các hạt nhân thông thường).
- Dự đoán mang tính xác suất (Gaussian) để người ta có thể tính toán khoảng tin cậy theo kinh nghiệm và quyết định dựa trên những khoảng đó liệu người ta có nên điều chỉnh lại (phù hợp trực tuyến, phù hợp thích ứng) dự đoán trong một số khu vực quan tâm hay không.
- Đa năng: có thể chỉ định các nhân khác nhau . Các nhân chung được cung cấp, nhưng cũng có thể chỉ định các nhân tùy chỉnh.

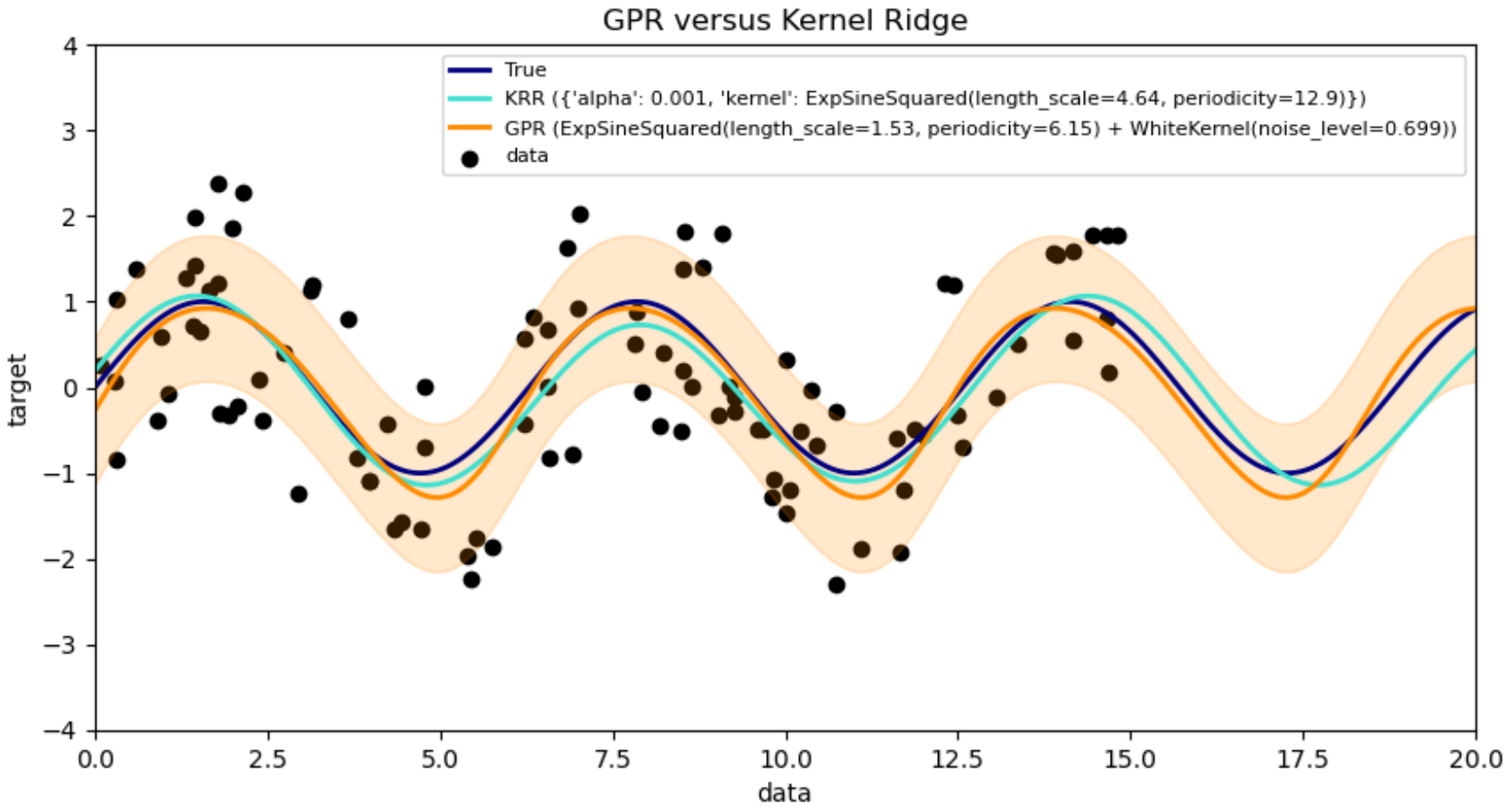
NHỮNG BẤT LỢI

- Chúng không thưa thớt, tức là chúng sử dụng toàn bộ thông tin về mẫu / tính năng để thực hiện dự đoán.
- Chúng mất hiệu quả trong không gian chiều cao - cụ thể là khi số lượng đối tượng vượt quá vài chục.

SO SÁNH KERNEL RIDGE VÀ GAUSSIAN PROCESS REGRESSION

- Cả hai đều sử dụng “ kernel trick”
- GPR có thể chọn siêu tham số của kernel dựa trên độ dốc đi lên của hàm khả năng cận biên trong khi KRR cần thực hiện tìm kiếm trên một hàm mất mát được xác nhận chéo
- GPR học một mô hình chung, xác suất của hàm mục tiêu và do đó có thể cung cấp khoảng tin cậy có ý nghĩa và mẫu hậu cùng với các dự đoán trong khi KRR chỉ cung cấp các dự đoán.

Ví dụ này minh họa cả hai phương pháp trên tập dữ liệu nhân tạo, bao gồm hàm đích hình sin và nhiễu mạnh



```
Time for KRR fitting: 4.001
Time for GPR fitting: 0.094
Time for KRR prediction: 0.030
Time for GPR prediction: 0.054
Time for GPR prediction with standard-deviation: 0.045
```

- Sự khác nhau chủ yếu về thời gian điều chỉnh và dự đoán

➢ Trong việc điều chỉnh KRR về nguyên tắc là nhanh, thì việc tìm kiếm theo lưới để tối ưu hóa siêu tham số sẽ tăng theo cấp số nhân với số siêu tham số.

➢ Việc tối ưu hóa dựa trên độ dốc của các tham số trong GPR không bị ảnh hưởng bởi tỷ lệ mũ này và do đó nhanh hơn đáng kể trong ví dụ này với không gian siêu tham số 3 chiều.

➢ Thời gian dự đoán tương tự nhau.