Lecture 16,17: Kernels

Recap: Kernels are much more efficient to compute than features

- As illustrating examples, consider polynomial features of degree exactly k
  - $\phi(x) = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ for $k = 1$ and $d = 2$, then $K(x, x') = x_1 x_1' + x_2 x_2'$
  - $\phi(x) = \begin{bmatrix} x_1^2 \\ x_2^2 \\ x_1 x_2 \\ x_2 x_1 \end{bmatrix}$ for $k = 2$ and $d = 2$, then $K(x, x') = (x^T x')^2$
- Note that for a data point x_i, **explicitly** computing the feature $\phi(x_i)$ takes memory/time $p = d^k$
- For a data point x_i, if we can make predictions by only computing the kernel, then computing $\{K(x_i, x_j)\}_{j=1}^n$ takes memory/time dn
  - The features are **implicit** and accessed only via kernels, making it efficient

Examples of popular Kernels

- Polynomials of degree exactly k

$$K(x,x') = (x^T x')^k$$

- Polynomials of degree up to k

$$K(x,x') = (1 + x^T x')^k$$

- Gaussian (squared exponential) kernel
(a.k.a RBF kernel for Radial Basis Function)

$$K(x,x') = \exp\Big(-\frac{\|x-x'\|_2^2}{2\sigma^2}\Big)$$

- Sigmoid

$$K(x,x') = \tanh(\gamma x^T x' + r)$$

- All these kernels are efficient to compute, but the corresponding features are in high-dimensions

Ridge Linear Regression as Kernels

- Consider Ridge regression: $\hat{w} = \arg\min_{w \in \mathbb{R}^d} \|\mathbf{y} - \mathbf{X}w\|_2^2 + \lambda\|w\|_2^2$
- We will represent prediction with $\widehat{w}$ using linear kernel defined as
  $K(x, x') = x^T x'$ (corresponding feature is x itself and hence $d = p$)
- Training: $\widehat{w} = \begin{cases} (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{d\times d})^{-1}\mathbf{X}^T\mathbf{y} & \text{(when } n > d) \\ \mathbf{X}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_{n\times n})^{-1}\mathbf{y} & \text{(when } n < d \text{ via linear algebra)} \end{cases}$
- Prediction: $x_{\text{new}} \in \mathbb{R}^d$

$$\begin{aligned} \widehat{y}_{\text{new}} &= \widehat{w}^T x_{\text{new}} \\ &= \mathbf{y}^T(\mathbf{X}\mathbf{X}^T + \lambda\mathbf{I}_{n\times n})^{-1}\mathbf{X}x_{\text{new}} \end{aligned}$$

- Hence, to make prediction on any future data points, all we need to know is
  - $\mathbf{X}x_{\text{new}} = \begin{bmatrix} x_1^T x_{\text{new}} \\ \vdots \\ x_n^T x_{\text{new}} \end{bmatrix} = \begin{bmatrix} K(x_1, x_{\text{new}}) \\ \vdots \\ K(x_n, x_{\text{new}}) \end{bmatrix} \in \mathbb{R}^n$, and $\mathbf{X}\mathbf{X}^T = \begin{bmatrix} K(x_1, x_1) & K(x_1, x_2) & \cdots \\ \vdots & \vdots & \\ K(x_n, x_1) & K(x_n, x_2) & \cdots \end{bmatrix} \in \mathbb{R}^{n\times n}$
- For **ridge regression**, even if we run on feature map $\phi(x) \in \mathbb{R}^p$, we only need to access the features via kernel $K(x_i, x_j)$ and $K(x_i, x_{\text{new}})$ and not the features $\phi(x_i)$

Example: feature vs. kernel

- Ridge regression with feature map $\phi(\cdot) \in \mathbb{R}^p$
  - Solve for $\widehat{w} = \arg\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - w^T\phi(x_i)\right)^2 + \lambda \|w\|_2^2$
  - Slow when $p \gg d$

- Ridge regression with kernel $K(\cdot, \cdot)$ corresponding to the feature map $\phi(\cdot)$
  - Finds the optimal solution of the above problem, but
  - only accesses the data via kernel $\{K(x_i, x_j)\}$, which is independent of p and only depends on n, if kernel is efficient to compute
    (which is true for all kernels we looked at and all kernels people use in practice)

The Kernel Trick

- Given data $\{(x_i, y_i)\}_{i=1}^n$, pick a kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$

1. For a choice of a loss, use a linear predictor of the form

$$\boxed{\widehat{w} = \sum_{i=1}^n \alpha_i x_i} \quad \text{for some } \alpha = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{bmatrix} \in \mathbb{R}^n \text{ to be learned}$$

Prediction is $\widehat{y}_{\text{new}} = \widehat{w}^T x_{\text{new}} = \sum_{i=1}^n \alpha_i \, x_i^T x_{\text{new}}$

2. Design an algorithm that finds α while accessing the data only via $\{x_i^T x_j\}$
3. Substitute $x_i^T x_j$ with $K(x_i, x_j)$, and find α using the above algorithm from step 2.
4. Make prediction with $\widehat{y}_{\text{new}} = \sum_{i=1}^n \alpha_i K(x_i, x_{\text{new}})$

   (replacing $x_i^T x_{\text{new}}$ with $K(x_i, x_{\text{new}})$)

The Kernel Trick for regularized least squares

$$\widehat{w} = \arg\min_w \sum_{i=1}^n (y_i - w^T x_i)^2 + \lambda \|w\|_2^2$$

There exists an $\alpha \in \mathbb{R}^n$: $\widehat{w} = \sum_{i=1}^n \alpha_i x_i$ (Step 1. We will prove it later)

$$\widehat{\alpha} = \arg\min_\alpha \sum_{i=1}^n (y_i - \sum_{j=1}^n \alpha_j \langle x_j, x_i \rangle)^2 + \lambda \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \langle x_i, x_j \rangle$$

(Step 2. Write an algorithm in terms of $\widehat{\alpha}$)

$$\widehat{\alpha}_{\text{kernel}} = \arg\min_\alpha \sum_{i=1}^n (y_i - \sum_{j=1}^n \alpha_j K(x_i, x_j))^2 + \lambda \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j)$$

(Step 3. Switch inner product with kernel)

$$= \arg\min_\alpha ||\mathbf{y} - \mathbf{K}\alpha||_2^2 + \lambda \alpha^T \mathbf{K} \alpha$$

Where $\mathbf{K}_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$

(Solve for $\widehat{\alpha}_{\text{kernel}}$)

Thus, $\widehat{\alpha}_{\text{kernel}} = (\mathbf{K} + \lambda \mathbf{I}_{n \times n})^{-1} \mathbf{y}$

Why do we need regularization when using kernels?

- Typically, $p \gg d$ and $\mathbf{K} \succ 0$. Why?

- So $\mathbf{K}$ is invertible and $\widehat{\alpha} = (\mathbf{K} + \lambda \mathbf{I}_{n \times n})^{-1}\mathbf{y}$ is well defined.
- What if $\lambda = 0$? What goes wrong?

$$\arg\min_{\alpha} ||\mathbf{y} - \mathbf{K}\alpha||_2^2$$

The Kernel Trick for SVMs

$$\widehat{w} = \arg\min_{w,b} \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(b + w^T x_i)\} + \lambda \|w\|_2^2$$

There exists an $\alpha \in \mathbb{R}^n$: $\widehat{w} = \sum_{i=1}^{n} \alpha_i x_i$ (Step 1. We will prove it later)

$$\widehat{\alpha}, \widehat{b} = \arg\min_{\alpha \in \mathbb{R}^n, b} \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(b + \sum_{j=1}^{n} \alpha_j x_j^T x_i)\} + \lambda \sum_{i=1,j=1}^{n} \alpha_i \alpha_j x_i^T x_j$$

(Step 2. Write an algorithm in terms of $\widehat{\alpha}$)

$$\widehat{\alpha}_{\text{kernel}}, \widehat{b}_{\text{kernel}} = \arg\min_{\alpha \in \mathbb{R}^n, b} \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(b + \sum_{j=1}^{n} \alpha_j K(x_j, x_i))\} + \lambda \sum_{i=1,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$$

(Step 3. Switch inner product with kernel)

$$= \arg\min_{\alpha \in \mathbb{R}^n, b} \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(b + \mathbf{K}\alpha)\} + \lambda \alpha^T \mathbf{K} \alpha$$

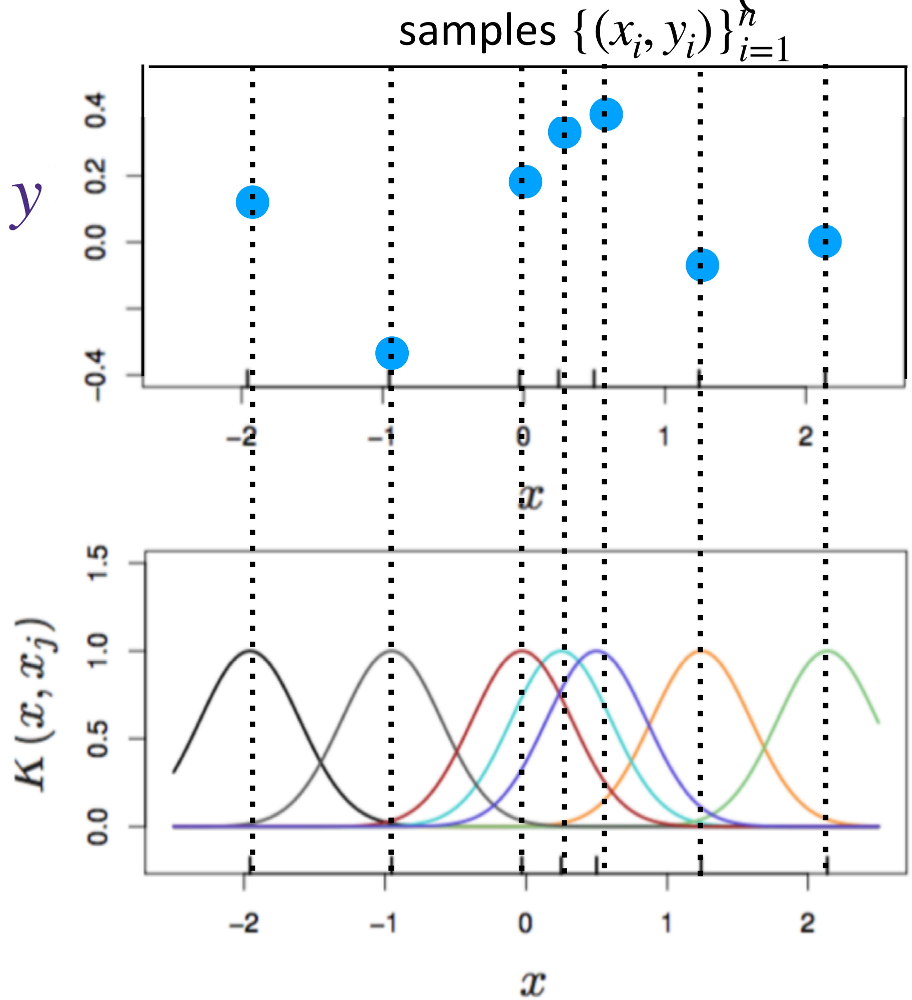
Where $\mathbf{K}_{ij} = K(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle$

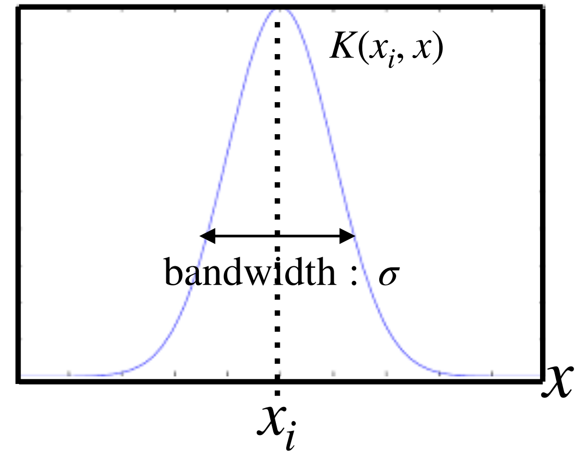
Prediction for x_{new}:

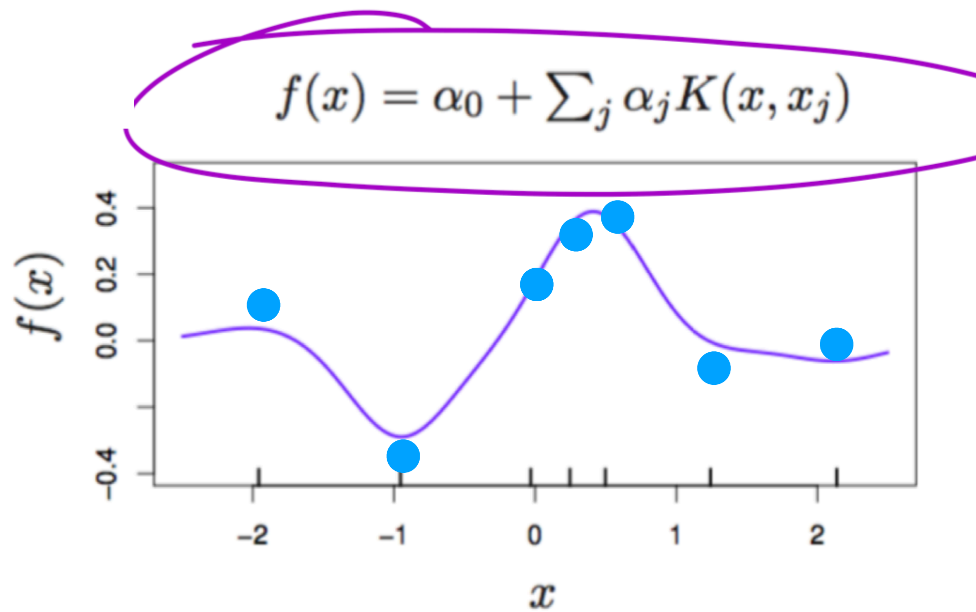
(Solve for $\widehat{\alpha}_{\text{kernel}}, \widehat{b}_{\text{kernel}}$ using optimization)

$$\widehat{y} = \text{sign}\left(\sum_{i=1}^{n} \widehat{\alpha}_{\text{kernel},i} K(x_i, x_{\text{new}}) + \widehat{b}_{\text{kernel}} \right)$$

RBF kernel $k(x_i, x) = \exp\left\{ -\dfrac{\|x_i - x\|_2^2}{2\sigma^2} \right\}$







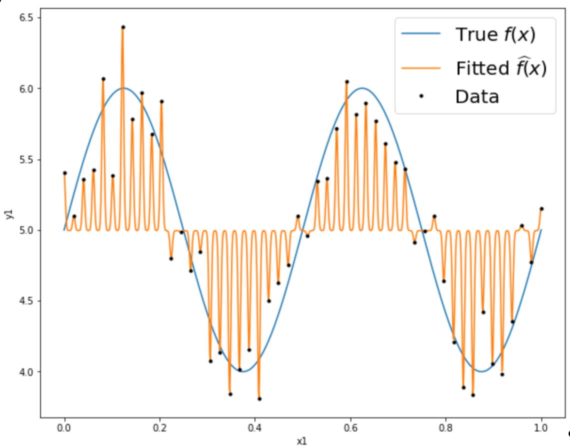
- predictor $f(x) = \sum_{i=1}^{n} \alpha_i K(x_i, x)$ is taking weighted sum of n kernel functions centered at each sample points

RBF kernel $k(x_i, x) = \exp\left\{ -\frac{\|x_i - x\|_2^2}{2\sigma^2} \right\}$

- $\mathscr{L}(\alpha) \;=\; \|\mathbf{K}\alpha - \mathbf{y}\|_2^2 \;+\; \lambda\|w\|_2^2$
- The bandwidth σ^2 of the kernel regularizes the predictor, and the regularization coefficient λ also regularizes the predictor

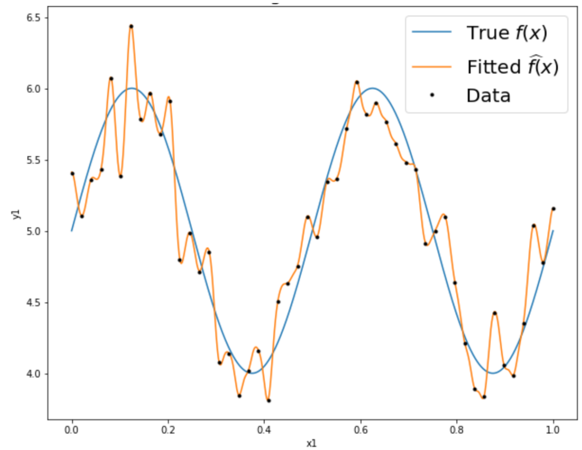
y

$\sigma = 10^{-3}\ \lambda = 10^{-4}$

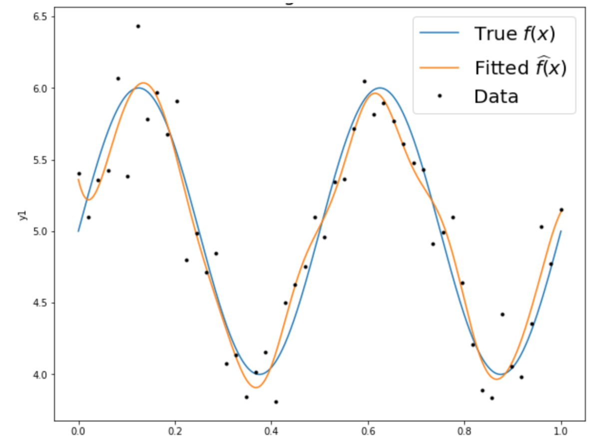


x

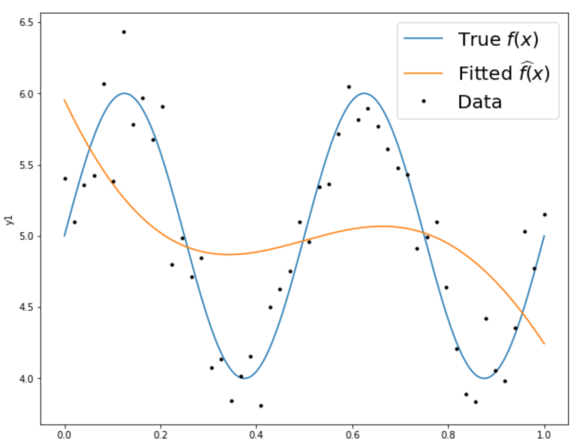
$\sigma = 10^{-2}\ \ \lambda = 10^{-4}$



$\sigma = 10^{-1}\ \ \lambda = 10^{-4}$

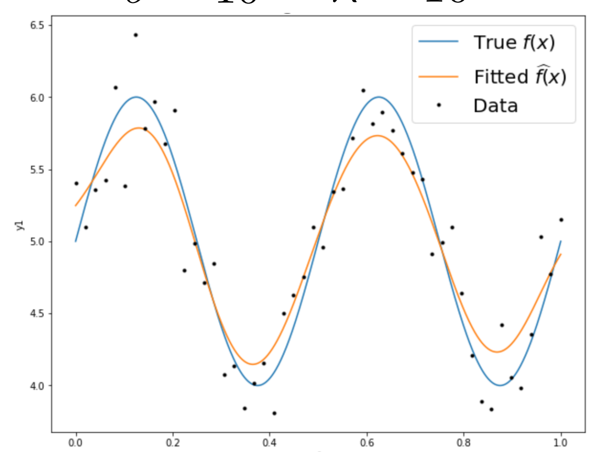


$\sigma = 10^{-0}\ \lambda = 10^{-4}$



$$\widehat{f}(x) = \sum_{i=1}^{n} \widehat{\alpha}_i K(x_i, x)$$

$\sigma = 10^{-1}\ \ \lambda = 10^{-0}$



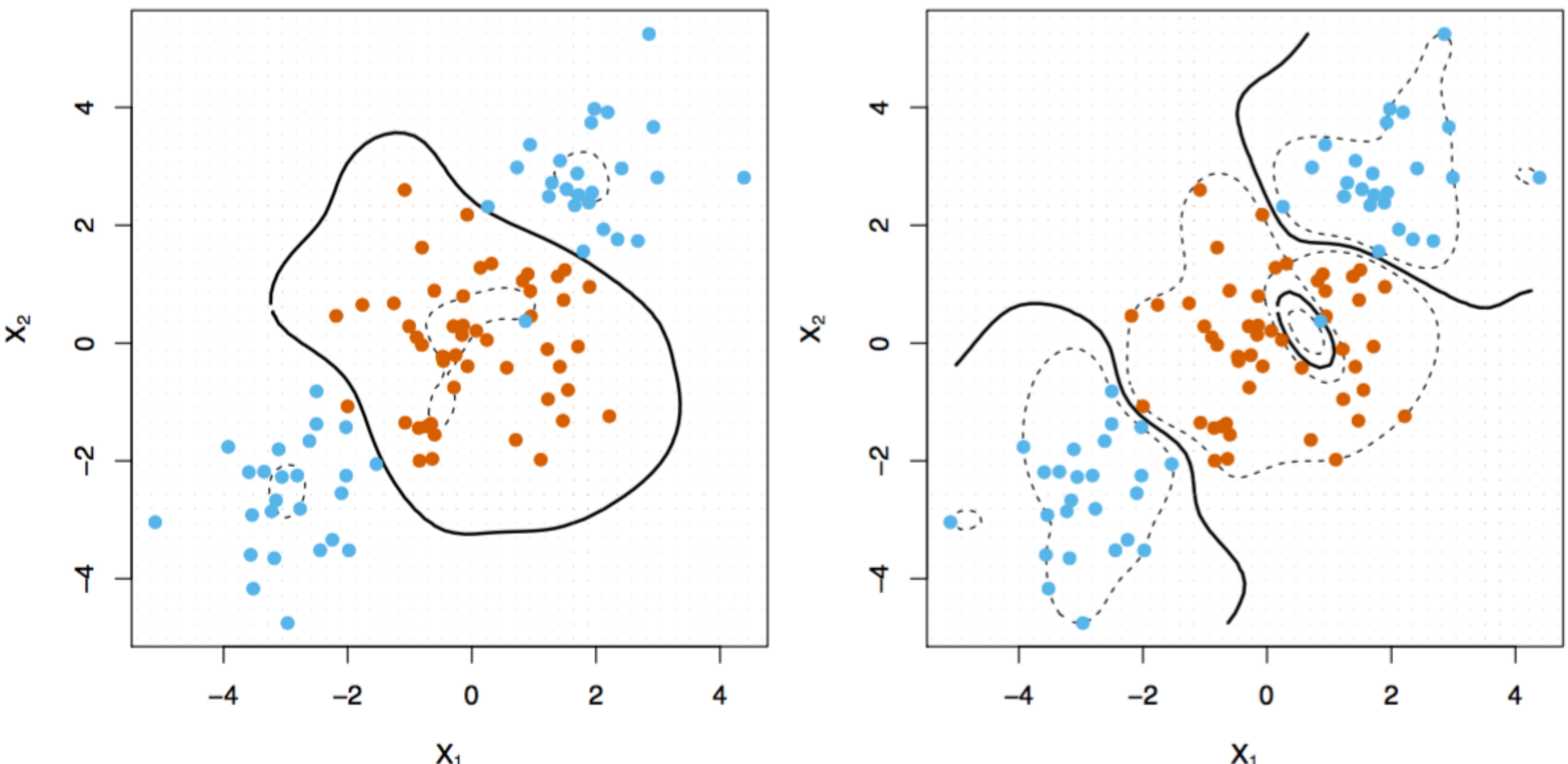
RBF kernel and random features

$$\widehat{w} = \arg\min_{w,b} \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(b + w^T x_i)\} + \lambda \|w\|_2^2$$

$$\widehat{\alpha}, \widehat{b} = \arg\min_{\alpha \in \mathbb{R}^n, b} \frac{1}{n}\sum_{i=1}^{n} \max\{0, 1 - y_i(b + \sum_{j=1}^{n} \alpha_j K(x_j, x_i))\} + \lambda \sum_{i=1, j=1}^{n} \alpha_i \alpha_j K(x_i, x_j)$$

Bandwidth σ is large enough

Bandwidth σ is small



Features vs. RBF kernel vs. random features

$$K(\mathbf{u},\mathbf{v}) = \exp\left(-\frac{||\mathbf{u}-\mathbf{v}||_2^2}{2\sigma^2}\right)$$

If n is very large, allocating an n-by-n matrix is tough.

Instead, consider generating random feature maps of the form:

$$\phi(x) = \begin{bmatrix} \sqrt{2}\cos(w_1^T x + b_1) \\ \vdots \\ \sqrt{2}\cos(w_p^T x + b_p) \end{bmatrix} \qquad \begin{aligned} w_k &\sim \mathcal{N}(0, 2\gamma\, I) \\ b_k &\sim \text{uniform}(0, \pi) \end{aligned}$$

with $p \ll n$

One can show that

$$\mathbb{E}_{w,b}\Big[\frac{1}{p}\phi(x)^T\phi(x')\Big] \;=\; \exp(-\gamma\|x-x'\|_2^2)$$

So this choice of random features approximate the desired RBF kernel with $\gamma = \dfrac{1}{2\sigma^2}$

[Rahimi, Recht NIPS 2007]
"NIPS Test of Time Award, 2018"

Kernel trick finds the optimal solution for linear models under a feature map $\phi(\,\cdot\,)$

- Once we have chosen to use a feature map $\phi(\,\cdot\,) \in \mathbb{R}^p$, what we want to solve is
$$\widehat{w} = \arg\min_{w \in \mathbb{R}^p} \sum_{i=1}^{n} \ell\left(y_i, w^T\phi(x_i)\right)$$ for some convex loss $\ell(,)$
- Gradient descent update (from initialization $w^{(0)} = 0$) that find the optimal solution is
$$w^{(t+1)} = w^{(t)} - \eta \sum_{i=1}^{n} \ell'(y_i, w^T\phi(x_i))\phi(x_i)$$
- One crucial observation is that all $w^{(t)}$'s (including the optimal solution $w^{(\infty)}$) lie on the subspace spanned by $\{\phi(x_1), \ldots, \phi(x_n)\}$, which is an n-dimensional subspace in $\mathbb{R}^p$
- Hence, it is sufficient to look for a solution that is represented as
$$\widehat{w} = \sum_{i=1}^{n} \alpha_i \phi(x_i)$$ to find the optimal solution
- Kernel trick finds the optimal solution efficiently, by searching over the model that can be represented as $\widehat{w} = \sum_{i=1}^{n} \alpha_i \phi(x_i)$

Fixed Feature V.S. Learned Feature

Can we learn the feature mapping $\phi : \mathbb{R}^d \to \mathbb{R}^p$ from data also?

Questions?

Bootstrap

confidence interval

- suppose you have training data $\{(x_i, y_i)\}_{i=1}^n$ drawn i.i.d. from some true distribution $P_{x,y}$
- we train a kernel ridge regressor, with some choice of a kernel $K : \mathbb{R}^{d\times d} \to \mathbb{R}$

$$\text{minmize}_\alpha \ \|\mathbf{K}\alpha - \mathbf{y}\|_2^2 + \lambda\alpha^T\mathbf{K}\alpha$$

- the resulting predictor is

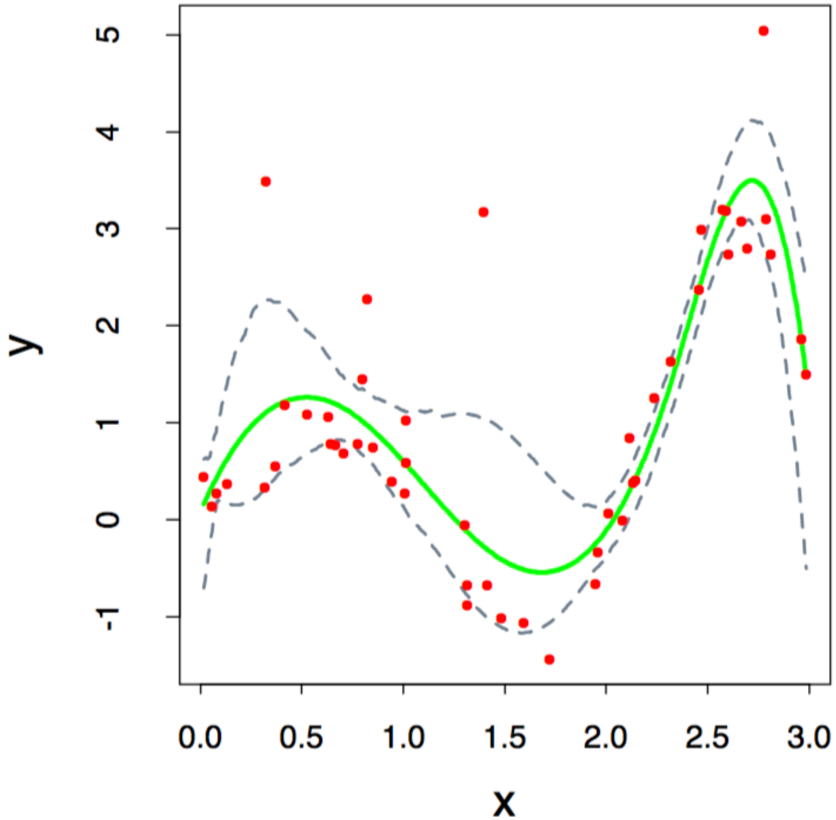
$$f(x) \ = \ \sum_{i=1}^n K(x_i, x)\hat{\alpha}_i,$$

where

$$\hat{\alpha} \ = \ (\mathbf{K} + \lambda\mathbf{I})^{-1}\mathbf{y} \quad \in \mathbb{R}^n$$

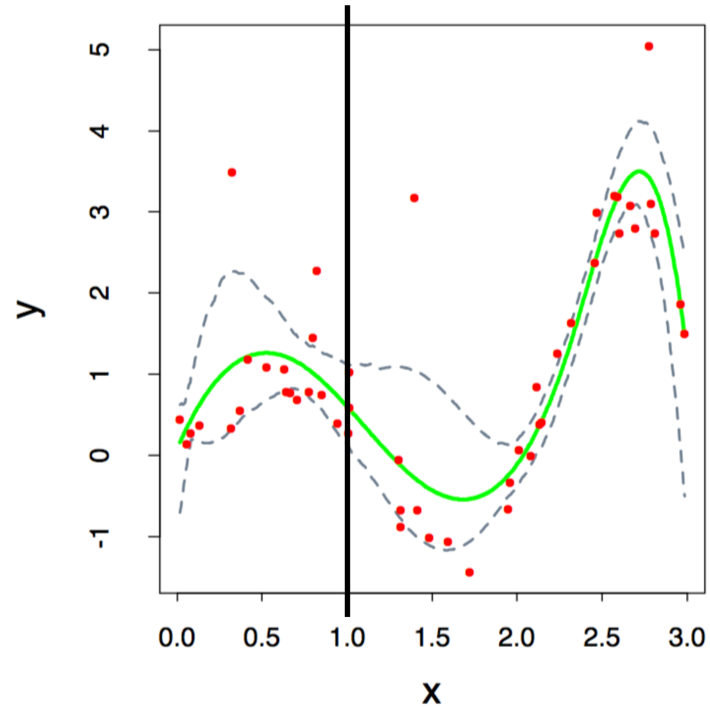
- we wish to build a confidence interval for our predictor $f(x)$, using 5% and 95% percentiles

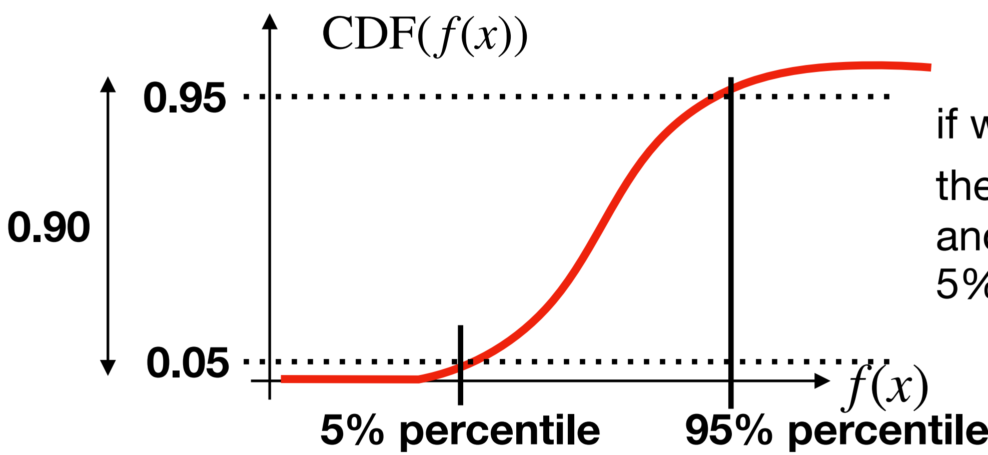
Example of 5% and 95% percentile curves for predictor f(x)



confidence interval

- let's focus on a single $x \in \mathbb{R}^d$
- note that our predictor $f(x)$ is a random variable, whose randomness comes from the training data $S_{\text{train}} = \{(x_i, y_i)\}_{i=1}^n$
- if we know the statistics (in particular the CDF of the random variable $f(x)$) of the predictor, then the **confidence interval** with **confidence level 90%** is defined as





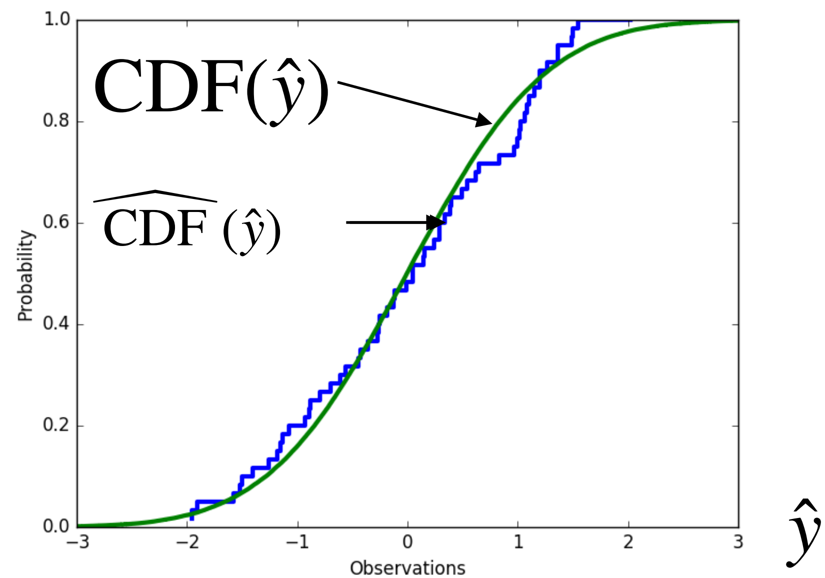
if we know the distribution of our predictor $f(x)$, the green line is the expectation $\mathbb{E}[f(x)]$ and the black dashed lines are the 5% and 95% percentiles in the figure above

- as we do not have the cumulative distribution function (CDF), we need to approximate them

confidence interval

- hypothetically, if we can sample as many times as we want, then we can train $B \in \mathbb{Z}^+$ i.i.d. predictors, each trained on n fresh samples to get **empirical estimate of the CDF of** $\hat{y} \ = \ f(x)$

- for b=1,…,B
  - draw n fresh samples
  - train a regularized kernel regression $\alpha^{*(b)}$
  - Predict $\hat{y}^{(b)} \ = \ (\alpha^{*(b)})^T h(x)$



- let the empirical CDF of those B predictors $\{\hat{y}^{(b)}\}_{b=1}^{B}$ be $\widehat{\mathrm{CDF}}(\hat{y})$, defined as

$$\widehat{\mathrm{CDF}}(\hat{y}) \ = \ \frac{1}{B}\sum_{b=1}^{B}\mathbf{I}\{\ \hat{y}^{(b)} \leq \hat{y}\ \} \ = \ \frac{1}{B}\sum_{b=1}^{B}\mathbf{I}\{\ (\alpha^{*(b)})^T h(x) \leq \hat{y}\ \}$$

- compute the confidence interval using $\widehat{\mathrm{CDF}}(\hat{y})$

Bootstrap

- as we cannot sample repeatedly (in typical cases), we use **bootstrap samples** instead
- bootstrap is a general tool for assessing statistical accuracy
- we learn it in the context of confidence interval for trained models

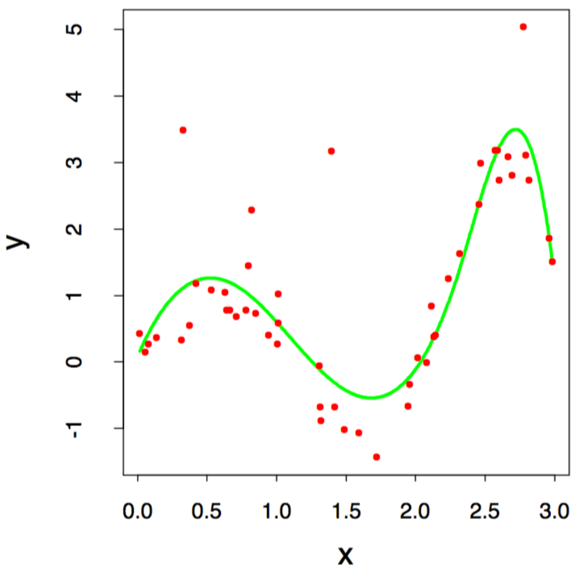
- a **bootstrap dataset** is created from the training dataset by taking n (the same size as the training data) examples uniformly at random **with replacement** from the training data $\{(x_i, y_i)\}_{i=1}^{n}$

- for b=1,…,B
  - create a bootstrap dataset $S^{(b)}_{\text{bootstrap}}$
  - train a regularized kernel regression $\alpha^{*(b)}$
  - predict $(\alpha^{*(b)})^T h(x)$

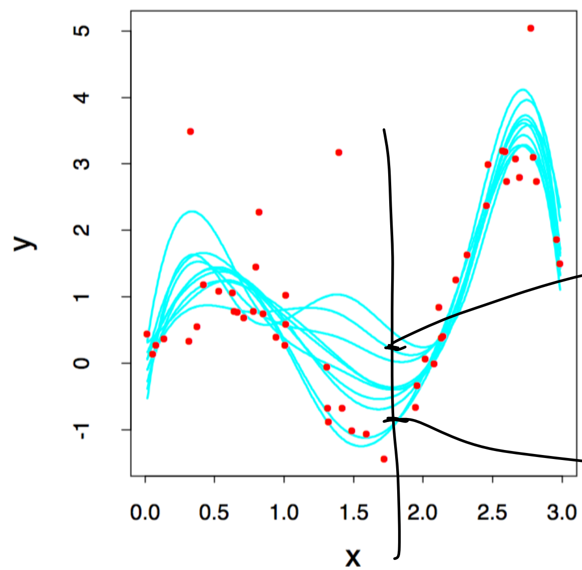
- compute the empirical CDF from the bootstrap datasets, and compute the confidence interval

bootstrap

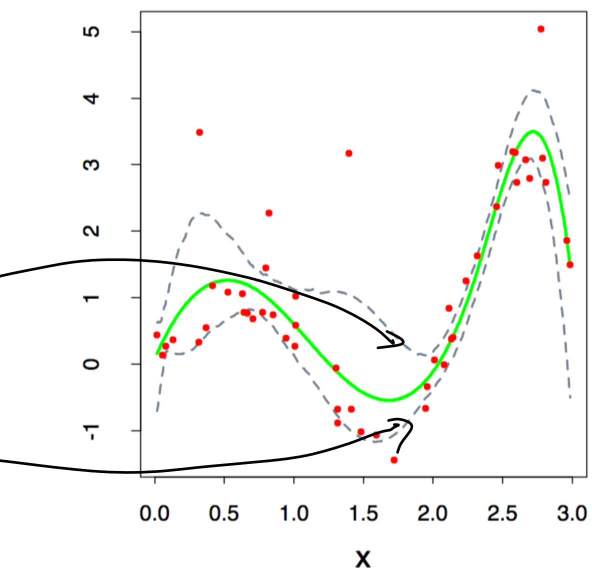
training a single predictor



multiple bootstrapped predictors



90% confidence interval



Figures from Hastie et al

Questions?