CSCI 567: Machine Learning

Vatsal Sharan

Fall 2022

Lecture 4, Sep 15

USC University of Southern California

Administrivia

- HW2 will be released tonight, due in about 2 weeks.
- We will post some practice problems for the quiz by early next week.

Recap

Ensuring generalization

Theorem. *Let $\mathcal{F}$ be a function class with size $|\mathcal{F}|$. Let $y = f^*(\boldsymbol{x})$ for some $f^* \in \mathcal{F}$. Suppose we get a training set $S = \{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$ of size n with each datapoint drawn i.i.d. from the data distribution D. Let*

$$f_S^{ERM} = \operatorname*{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i).$$

For any constants $\epsilon, \delta \in (0,1)$, if $n \geq \frac{\ln(|\mathcal{F}|/\delta)}{\epsilon}$, then with probability $(1-\delta)$ over $\{(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)\}$, $R(f_S^{ERM}) < \epsilon$.

A useful rule of thumb: to guarantee generalization, make sure that your training data set size n is at least linear in the number d of free parameters in the function that you're trying to learn.

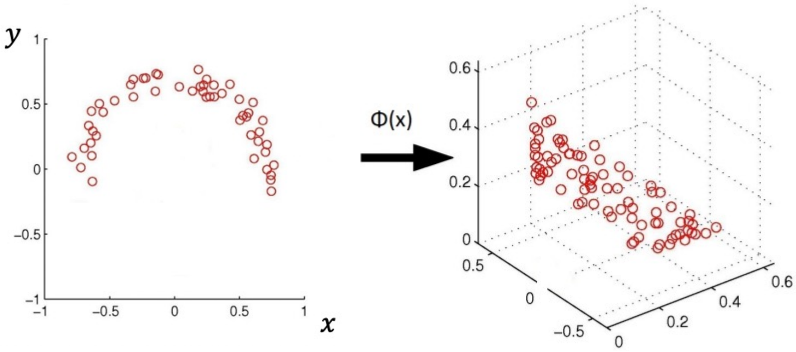
Beyond linear models: nonlinearly transformed features

1. Use a nonlinear mapping

$$\boldsymbol{\phi}(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^d \rightarrow \boldsymbol{z} \in \mathbb{R}^M$$

to transform the data to a more complicated feature space

2. Then apply linear regression (hope: linear model is a better fit for the new feature space).



Polynomial basis functions

Polynomial basis functions for $d = 1$

$$\phi(x) = \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^M \end{bmatrix} \quad \Rightarrow \quad f(x) = w_0 + \sum_{m=1}^{M} w_m x^m$$

Learning a linear model in the new space
$=$ learning an *M-degree polynomial model* in the original space

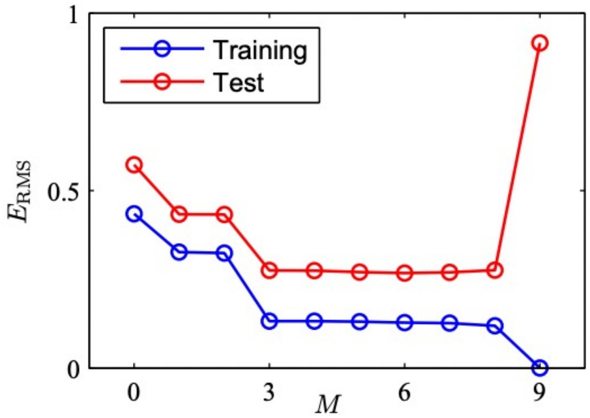
Underfitting and overfitting

$M \leq 2$ is *underfitting* the data

- large training error
- large test error

$M \geq 9$ is *overfitting* the data

- small training error
- **large test error**



More complicated models $\Rightarrow$ larger gap between training and test error

How to prevent overfitting?

See Colab notebook

Preventing overfitting: Regularization

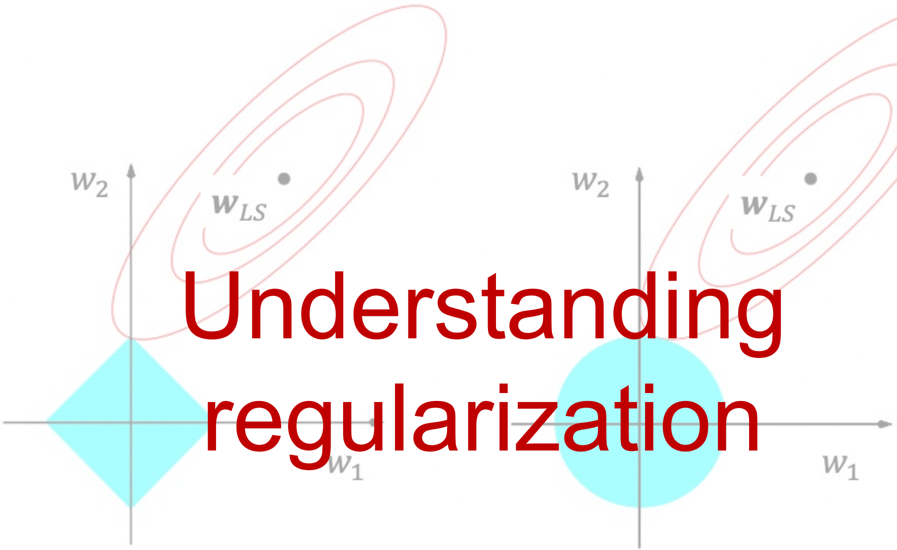
Regularized linear regression: new objective

$$G(\boldsymbol{w}) = \mathrm{RSS}(\boldsymbol{w}) + \lambda\psi(\boldsymbol{w})$$

Goal: find $\boldsymbol{w}^* = \mathrm{argmin}_w\, G(\boldsymbol{w})$

- $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^+$ is the *regularizer*
  - measure how complex the model $\boldsymbol{w}$ is, penalize complex models
  - common choices: $\|\boldsymbol{w}\|_2^2$, $\|\boldsymbol{w}\|_1$, etc.
- $\lambda > 0$ is the *regularization coefficient*
  - $\lambda = 0$, no regularization
  - $\lambda \rightarrow +\infty$, $\boldsymbol{w} \rightarrow \mathrm{argmin}_w\, \psi(\boldsymbol{w})$
  - i.e. control **trade-off** between training error and complexity

Understanding regularization



ℓ_2 **regularization**: penalizing large weights

ℓ_2 regularization, $\psi(w) = \|w\|_2^2$:

$$G(\boldsymbol{w}) = \mathrm{RSS}(\boldsymbol{w}) + \lambda\|\boldsymbol{w}\|_2^2 = \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$$

$$\nabla G(\boldsymbol{w}) = 2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{w} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}) + 2\lambda\boldsymbol{w} = 0$$
$$\Rightarrow \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I}\right)\boldsymbol{w} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$
$$\Rightarrow \boldsymbol{w}^* = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$$

Linear regression with ℓ_2 regularization is also known as **ridge regression**.

With a Bayesian viewpoint, corresponds to a Gaussian prior for $\boldsymbol{w}$.

Encouraging sparsity: ℓ_0 regularization

Sparsity of w: Number of non-zero coefficients in $\boldsymbol{w}$. Same as $||\mathbf{w}||_0$

Advantage:

- Sparse models are a natural inductive bias in many settings. In many applications we have numerous possible features, only some of which may have any relationship with the label.
- Sparse models may also be more **interpretable**. They could narrow down a small number of features which carry a lot of signal.
- Data required to learn sparse model maybe significantly less than to learn dense model.

We'll see more on the third point next.

ℓ_0 regularization: The good, the bad and the ugly

Choose $\psi(\boldsymbol{w}) = \|\boldsymbol{w}\|_0$.

$$G(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 + \lambda \|\boldsymbol{w}\|_0.$$

Good : "Information-theoretically" great ! (Need less data to learn).

Suppose weights in w are in $\{-W, -W+1, \ldots 0, \ldots, W\}$

How many such s-sparse vectors are there in dimensions?

Answer: $\binom{d}{s} \cdot (2W)^s$ possibilities.

ℓ_0 regularization: The good, the bad and the ugly

How much data to learn?

$$\left(\binom{d}{s} \simeq \left(\frac{d}{s}\right)^s\right)$$

About $\log(|F|)$ samples to learn. (using the Theorem from last time, note that we're ignoring ε, δ here)

$$\rightarrow \log\left(\binom{d}{s}(2w)^s\right) = \log\left(\left(\frac{d}{s}\right)^s\right) + \log\left((2w)^s\right)$$

$$= s\log\left(\frac{d}{s}\right) + s\log(2w)$$

How many free parameters?

every bit is like a parameter you need to learn

→ Choose s co-ordinates : need $\log d$ bits per co-ordinate

⇒ $s\log d$ in total.

→ choose the value for non-zero coordinates : fix s values $\simeq s\log W$ in total.

ℓ_0 regularization: The good, the bad and the ugly

In contrast, without s-sparsity need about $\gtrsim d$ samples in d-dimensions.

$\therefore$ If $s \ll d$, need much less data to learn!!

Bad: $\|w\|_0$ is non-convex :(

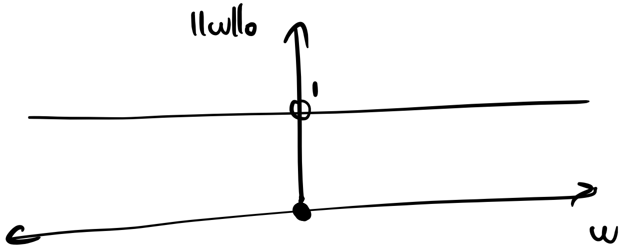
$\|w\|_p$, $p<1$ is non-convex.

minimizing $G(w) = \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|_0$

is NP-Hard :(

ℓ_0 regularization: The good, the bad and the ugly

Ugly: $\|w\|_0$ is highly-discontinuous.



GD has no hope!!

ℓ_1 regularization as a proxy for ℓ_0 regularization

Choose $\psi(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$.

$$G(\boldsymbol{w}) = \sum_{i=1}^{n} (\boldsymbol{w}^T \boldsymbol{x}_i - y_i)^2 + \lambda \|\boldsymbol{w}\|_1.$$

$\|w\|_1$ is convex :) Can use GD / SGD to solve!

Minimizing $\|w\|_1$ often suffices to minimize $\|w\|_0$!

ℓ_1 regularization as a proxy for ℓ_0 regularization

Theorem. *Given n vectors $\{\boldsymbol{x}_i \in \mathbb{R}^d, i \in [n]\}$ drawn i.i.d. from $N(0, \boldsymbol{I})$, let $y_i = {\boldsymbol{w}^*}^T \boldsymbol{x}_i$ for some $\boldsymbol{w}^*$ with $\|\boldsymbol{w}^*\|_0 = s$. Then for some fixed constant $C > 0$, the minimizer of $G(\boldsymbol{w})$ with $\psi(\boldsymbol{w}) = \|\boldsymbol{w}\|_1$ will be $\boldsymbol{w}^*$ as long as $n > C \cdot s \log d$ (with high probability over the randomness in the training datapoints $\boldsymbol{x}_i$).*

[similar result can also be proven under more general conditions].

(the details of this Theorem are not important, just focus on the takeaway)

Why does ℓ_1 regularization encourage sparse solutions?

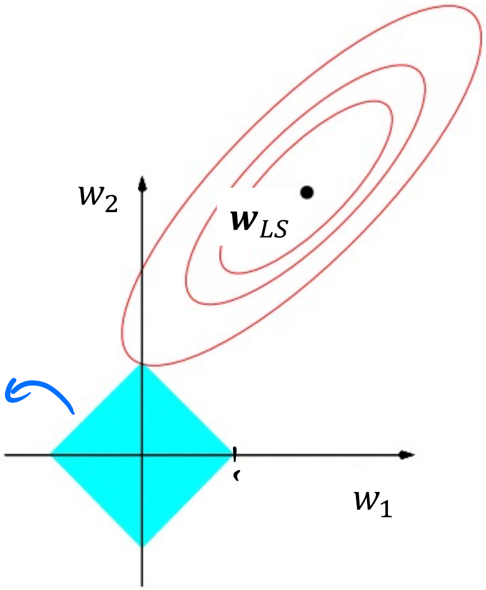
Optimization problem: $\mathrm{argmin}_w \; \mathrm{RSS}(\boldsymbol{w})$, subject to $\psi(\boldsymbol{w}) \leq \beta$

Contour lines: lines along which $RSS(w)$ remains constant

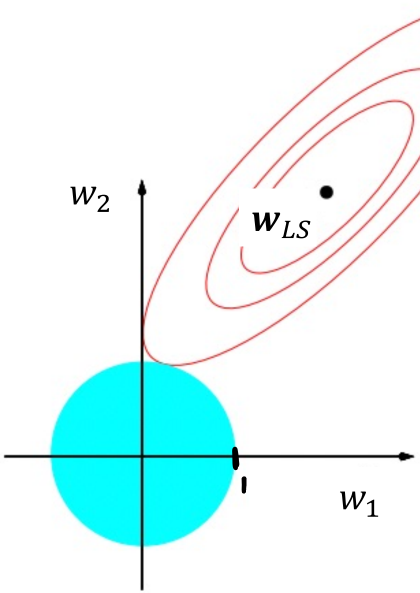
$RSS(w) = \|Xw - y\|_2^2$

$\beta = 1$

$\|w\|_1 \leq 1$



$\Psi(w) = \|w\|_1$



$\Psi(w) = \|w\|_2^2$

Adapted from ESL

Diving deeper: ℓ_1 and ℓ_2 regularization for the "isotropic" case

Isotropic informally means
① all features have mean 0
② all features have variance 1
③ features are uncorrelated

Isotropic assumption : $X^T X = I$

① $\psi(w) = \|w\|_2^2$

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

$$w^* = (X^T X + \lambda I)^{-1} X^T y$$

Now, $X^T X = I \Rightarrow w^* = \left(\frac{1}{1+\lambda}\right) X^T y$

jth row of X^T

$$\underbrace{w_j^*}_{\text{jth coordinate of } w^*} = \left(\frac{1}{1+\lambda}\right) \underbrace{x_{(j)}^T y}_{\text{correlation of jth feature with label}}$$

Diving deeper: ℓ_1 and ℓ_2 regularization for the "isotropic" case

l_2 regularization "shrinks" the estimated parameters.

Note: when features have unequal variance, l_2 regularization applies similar shrinkage to all of them

$\therefore$ scaling features can be important.

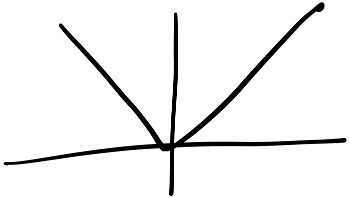
2 $\psi(w) = \|w\|_1$

$$G(w) = \sum_{i=1}^{n} \left(x_i^T w - y_i \right)^2 + \lambda \|w\|_1$$

Let's examine the gradient.

Diving deeper: ℓ_1 and ℓ_2 regularization for the "isotropic" case

What is gradient of $|w|$?



$$\frac{\partial |w|}{\partial w} = \begin{cases} 1 & w > 0 \\ -1 & w < 0 \end{cases}$$

At $w=0$, we have a subgradient, ignore for now.

For $\underbrace{w_j}_{\text{jth coordinate of } w} \neq 0$

$$\frac{\partial G(w)}{\partial w_j} = 2 \sum_{i=1}^{n} (x_i^T w - y_i) \underbrace{x_{i,j}}_{\text{jth-coordinate of } x_i} + \lambda \, \text{sign}(w_j)$$

Diving deeper: ℓ_1 and ℓ_2 regularization for the "isotropic" case

$$\frac{\partial L(w)}{\partial w_j} = 2\sum_{i=1}^{n}\left(x_{i,j}\, x_i^T w\right) - 2\sum_{i=1}^{n} x_{i,j}\, y_i + \lambda\, \text{sign}(w_j)$$

$$= 2\sum_{i=1}^{n} x_{i,j} \boxed{\; x_i \;} \boxed{w} - 2\, X_{(i)}^T y + \lambda\, \text{sign}(w_j)$$

This follows as $X^T X = I$

$$= 2 w_j - 2\, X_{(i)}^T y + \lambda\, \text{sign}(w_j)$$

$$\therefore \text{ GD steps: } \quad w_j \leftarrow w_j - \eta\left(2\left(w_j - X_{(i)}^T y\right) + \lambda\, \text{sign}(w_j)\right)$$

Diving deeper: ℓ_1 and ℓ_2 regularization for the "isotropic" case

Let's understand the gradient.

First, without ℓ_1 regularization,

$$w_j \leftarrow w_j - 2\eta\left(w_j - X_{(i)}^T Y\right)$$

With ℓ_1 regularization: GD always has a shift of $-\eta\lambda\,\text{sign}(w_j)$ which pushes towards 0.

Diving deeper: ℓ_1 and ℓ_2 regularization for the "isotropic" case

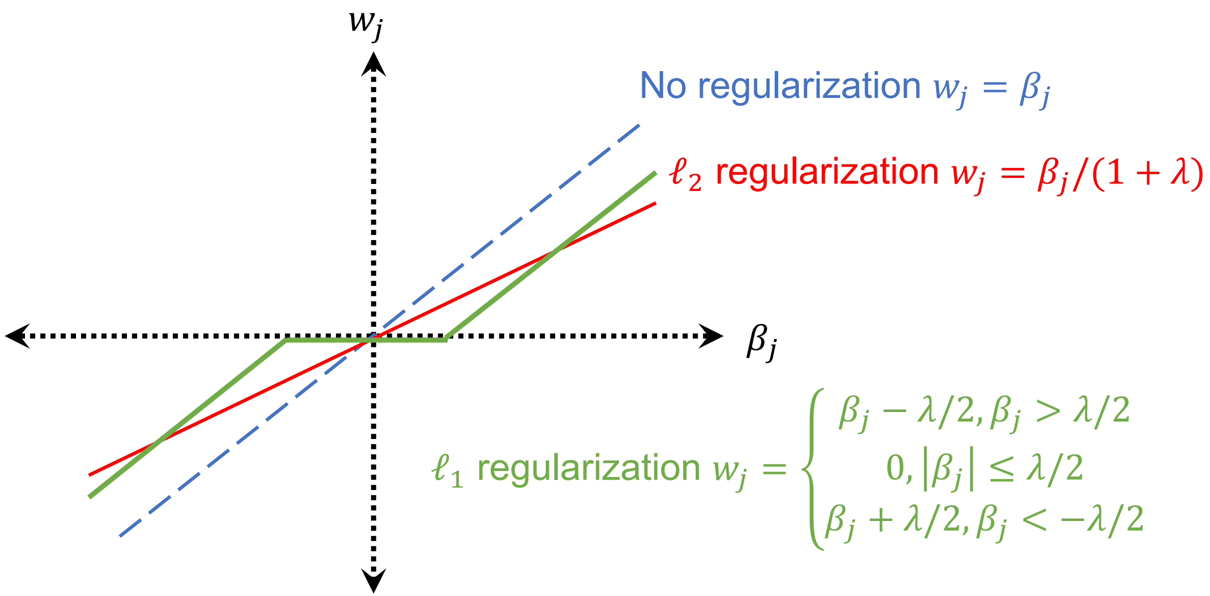
Let $\beta_j = \boldsymbol{X}_{(i)}^T \boldsymbol{y}$

Using subgradients, we can show that for the ℓ_1 regularized case:

$$w_j = \begin{cases} \beta_j - \lambda/2, \beta_j > \lambda/2 \\ 0, |\beta_j| \leq \lambda/2 \\ \beta_j + \lambda/2, \beta_j < -\lambda/2 \end{cases}$$

Diving deeper: ℓ_1 and ℓ_2 regularization for the "isotropic" case

Summary: Isotropic case ($\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}$).
Let $\beta_j = \boldsymbol{X}_{(j)}^T\boldsymbol{y}$



Implicit regularization

So far, we explicitly added a $\psi(\boldsymbol{w})$ term to our objective function to regularize.

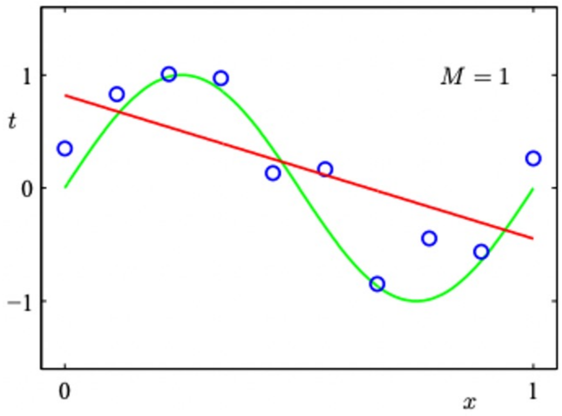
In many cases, the optimization algorithm we use can themselves act as regularizers, favoring some solutions over others.

Currently a very active area of research, you'll see more in the homework.

Bias-variance tradeoff

The phenomenon of underfitting and overfitting is often referred to as the ***bias-variance tradeoff*** in the literature.

A model whose complexity is too *small* for the task will *underfit*. This is a model with a large bias because the model's accuracy will not improve even if we add a lot of training data.

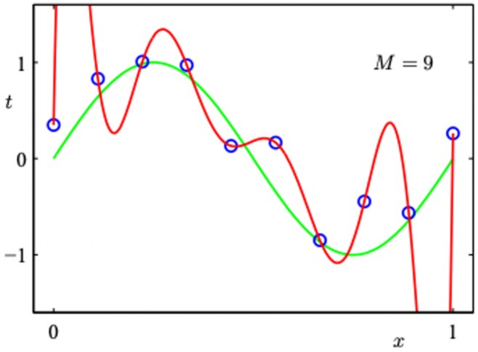


sin(x) fitting example we saw in Lec 3

Bias-variance tradeoff

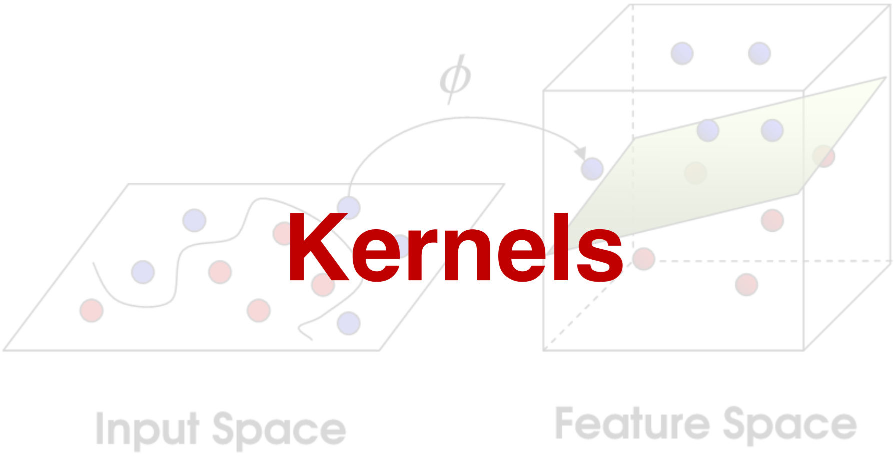
The phenomenon of underfitting and overfitting is often referred to as the ***bias-variance tradeoff*** in the literature.

A model whose complexity is too *large* for the amount of available training data will *overfit*. This is a model with high variance, because the model's predictions will vary a lot with the randomness in the training data (it can even fit any noise in the training data).



sin(x) fitting example we saw in Lec 3

Kernels



Motivation

Recall the nonlinear function map for linear regression:

1. **Use a nonlinear mapping**

$$\boldsymbol{\phi}(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^d \rightarrow \boldsymbol{z} \in \mathbb{R}^M$$

to transform the data to a more complicated feature space

2. **Then apply linear regression** (hope: linear model is a better fit for the new feature space).

Kernel methods give a way to choose and efficiently work with the nonlinear map $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^M$ (for linear regression, and much more broadly).

Regularized least squares

Let's continue with regularized least squares with non-linear basis:

$$\begin{aligned} \boldsymbol{w}^* &= \operatorname*{argmin}_{\boldsymbol{w}} F(\boldsymbol{w}) \\ &= \operatorname*{argmin}_{\boldsymbol{w}} \left(\|\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda \|\boldsymbol{w}\|_2^2 \right) \\ &= \left(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{y} \end{aligned}$$

$$\underset{\in \mathbb{R}^{n \times M}}{\boldsymbol{\Phi}} = \begin{pmatrix} \phi(\boldsymbol{x}_1)^{\mathrm{T}} \\ \phi(\boldsymbol{x}_2)^{\mathrm{T}} \\ \vdots \\ \phi(\boldsymbol{x}_n)^{\mathrm{T}} \end{pmatrix}, \quad \underset{\in \mathbb{R}^{n}}{\boldsymbol{y}} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

This operates in space $\mathbb{R}^M$ and M could be huge (and even infinite).

Regularized least squares solution: Another look

By setting the gradient of $F(\boldsymbol{w}) = \|\boldsymbol{\Phi}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{w}\|_2^2$ to be $\mathbf{0}$:

$$\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{w}^* - \boldsymbol{y}) + \lambda\boldsymbol{w}^* = \mathbf{0}$$

scale by $1/\lambda$

$$w^* = M\left\{\overbrace{\Phi^T}^{n}\right. \alpha$$

we know

$$\boldsymbol{w}^* = \frac{1}{\lambda}\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}^*) = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha} = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)$$

Thus the least square solution is **a linear combination of features of the datapoints**!

This calculation does not show what $\boldsymbol{\alpha}$ should be, but ignore that for now.

Why is this helpful?

Assuming we know $\boldsymbol{\alpha}$, the prediction of $\boldsymbol{w}^*$ on a new example $\boldsymbol{x}$ is

$$\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}) \leftarrow \sum \left(\alpha_i \phi(x_i)\right)^{\mathrm{T}} \phi(x)$$

Therefore, *only inner products in the new feature space matter!*

Kernel methods are exactly about computing inner products *without explicitly computing ϕ*.

But we need to figure out what $\boldsymbol{\alpha}$ is first!

Solving for α, Step 1: Kernel matrix

Plugging in $\boldsymbol{w} = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha}$ into $F(\boldsymbol{w})$ gives

$= (\Phi^{T}\alpha)^{T}(\Phi^{T}\alpha) = \alpha^{T}\Phi\Phi^{T}\alpha$

$$\begin{aligned} H(\boldsymbol{\alpha}) &= F(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha}) \\ &= \|\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda\|\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha}\|_2^2 \\ &= \|\boldsymbol{K}\boldsymbol{\alpha} - \boldsymbol{y}\|_2^2 + \lambda\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}\boldsymbol{\alpha} \quad (\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} \in \mathbb{R}^{n\times n}) \end{aligned}$$

$\boldsymbol{K}$ is called **Gram matrix** or **kernel matrix** where the (i,j)-th entry is

$$\boldsymbol{K}_{(i,j)} = \phi(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)$$

Kernel matrix: Example

$$\phi(x_1) = \begin{pmatrix} 1 \\ -1 \\ 1 \\ -1 \end{pmatrix} \quad \phi(x_2) = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \quad \phi(x_3) = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$$

Gram/Kernel matrix

$$\begin{aligned} \boldsymbol{K} &= \begin{pmatrix} \phi(x_1)^{\mathrm{T}}\phi(x_1) & \phi(x_1)^{\mathrm{T}}\phi(x_2) & \phi(x_1)^{\mathrm{T}}\phi(x_3) \\ \phi(x_2)^{\mathrm{T}}\phi(x_1) & \phi(x_2)^{\mathrm{T}}\phi(x_2) & \phi(x_2)^{\mathrm{T}}\phi(x_3) \\ \phi(x_3)^{\mathrm{T}}\phi(x_1) & \phi(x_3)^{\mathrm{T}}\phi(x_2) & \phi(x_3)^{\mathrm{T}}\phi(x_3) \end{pmatrix} \\ &= \begin{pmatrix} 4 & 1 & 0 \\ 1 & 1 & 1 \\ 0 & 1 & 4 \end{pmatrix} \end{aligned}$$

Kernel matrix vs Covariance matrix

	dimensions	entry (i, j)	property
$\Phi\Phi^{\mathrm{T}}$	$n \times n$	$\phi(x_i)^T \phi(x_j)$	both are symmetric & positive semidefinite (psd)
$\Phi^{\mathrm{T}}\Phi$	$M \times M$	$\sum_{k=1}^{n} \underbrace{\phi(x_k)_i}_{\text{ith co-ordinate of feature}} \phi(x_k)_j$	

Why are they psd?

Any matrix $A = UU^T$ is psd.

$$(x^T A x = x^T U U^T x = \|U^T x\|_2^2 \geq 0)$$

Solving for α, Step 2: Minimize the dual

Minimize (the so-called *dual formulation*)

$$H(\boldsymbol{\alpha}) = \|\boldsymbol{K\alpha} - \boldsymbol{y}\|_2^2 + \lambda\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K\alpha}$$

Setting the derivative to $\mathbf{0}$ we have

$$\mathbf{0} = (\boldsymbol{K}^2 + \lambda\boldsymbol{K})\boldsymbol{\alpha} - \boldsymbol{Ky} = \boldsymbol{K}\left((\boldsymbol{K} + \lambda\boldsymbol{I})\boldsymbol{\alpha} - \boldsymbol{y}\right)$$

Thus $\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$ **is a minimizer** and we obtain

$$\boldsymbol{w}^* = \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\alpha} = \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{K} + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}$$

Exercise: *are there other minimizers? and are there other $\boldsymbol{w}^*$'s?*

Comparing two solutions

covariance

Minimizing $F(\boldsymbol{w})$ **gives** $\boldsymbol{w}^* = (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y}$

Minimizing $H(\boldsymbol{\alpha})$ **gives** $\boldsymbol{w}^* = \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$ Kernel

Note $\boldsymbol{I}$ has different dimensions in these two formulas.

Natural question: *are the two solutions the same or different?*

They have to be the same because $F(\boldsymbol{w})$ has a unique minimizer!

And they are:

$$
\begin{aligned}
&(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{y} \\
&= (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda \boldsymbol{I})(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y} \\
&= (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda \boldsymbol{\Phi}^{\mathrm{T}})(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y} \\
&= (\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Phi} + \lambda \boldsymbol{I})\boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y} \\
&= \boldsymbol{\Phi}^{\mathrm{T}}(\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}
\end{aligned}
$$

The kernel trick

If the solutions are the same, then what is the difference?

First, computing $(\overbrace{\underbrace{\mathbf{\Phi\Phi}^{\mathrm{T}}}_{n\times n \text{ dimensional}} + \lambda \boldsymbol{I}}^{\text{can solve in } O(d^3) \text{ time}})^{-1}$ can be more efficient than computing $(\overbrace{\underbrace{\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi}}_{M\times M \text{ dimensional}} + \lambda \boldsymbol{I}}^{O(M^3) \text{ time}})^{-1}$ when $n \leq M$.

More importantly, computing $\boldsymbol{\alpha} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}\boldsymbol{y}$ also *only requires computing inner products in the new feature space!*

Now we can conclude that the exact form of $\boldsymbol{\phi}(\cdot)$ is not essential; *all we need to do is know the inner products $\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}')$.*

For some $\boldsymbol{\phi}$ it is indeed possible to compute $\boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}')$ without computing/knowing $\boldsymbol{\phi}$. This is the *kernel trick*.

The kernel trick: Example 1

Consider the following polynomial basis $\boldsymbol{\phi} : \mathbb{R}^2 \rightarrow \mathbb{R}^3$:

$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$$\boldsymbol{\phi}(\boldsymbol{x}) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$$

What is the inner product between $\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{\phi}(\boldsymbol{x}')$?

$$\begin{aligned} \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}') &= {x_1}^2{x_1'}^2 + 2x_1x_2x_1'x_2' + {x_2}^2{x_2'}^2 \\ &= (x_1x_1' + x_2x_2')^2 = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}')^2 \end{aligned}$$

Therefore, *the inner product in the new space is simply a function of the inner product in the original space*.

The kernel trick: Example 2

$\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^{2d}$ is parameterized by θ:

$$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{pmatrix} \qquad \boldsymbol{\phi}_\theta(\boldsymbol{x}) = \begin{pmatrix} \cos(\theta x_1) \\ \sin(\theta x_1) \\ \vdots \\ \cos(\theta x_m) \\ \sin(\theta x_m) \end{pmatrix}$$

What is the inner product between $\boldsymbol{\phi}_\theta(\boldsymbol{x})$ and $\boldsymbol{\phi}_\theta(\boldsymbol{x}')$?

$$\begin{aligned} \boldsymbol{\phi}_\theta(\boldsymbol{x})^{\mathrm{T}}\boldsymbol{\phi}_\theta(\boldsymbol{x}') &= \sum_{m=1}^{d} \cos(\theta x_m)\cos(\theta x'_m) + \sin(\theta x_m)\sin(\theta x'_m) \\ &= \sum_{m=1}^{d} \cos(\theta(x_m - x'_m)) \qquad \text{(trigonometric identity)} \end{aligned}$$

Once again, *the inner product in the new space is a simple function of the features in the original space.*

The kernel trick: Example 3

Based on ϕ_θ, define $\phi_L : \mathbb{R}^d \to \mathbb{R}^{2d(L+1)}$ for some integer L:

$$\phi_L(\boldsymbol{x}) = \begin{pmatrix} \phi_0(\boldsymbol{x}) \\ \phi_{\frac{2\pi}{L}}(\boldsymbol{x}) \\ \phi_{2\frac{2\pi}{L}}(\boldsymbol{x}) \\ \vdots \\ \phi_{L\frac{2\pi}{L}}(\boldsymbol{x}) \end{pmatrix}$$

θ varies from $(0, \frac{2\pi}{L}, \ldots, 2\pi)$

What is the inner product between $\phi_L(\boldsymbol{x})$ and $\phi_L(\boldsymbol{x}')$?

$$\phi_L(\boldsymbol{x})^{\mathrm{T}}\phi_L(\boldsymbol{x}') = \sum_{\ell=0}^{L} \phi_{\frac{2\pi\ell}{L}}(\boldsymbol{x})^{\mathrm{T}}\phi_{\frac{2\pi\ell}{L}}(\boldsymbol{x}')$$

$$= \sum_{\ell=0}^{L}\sum_{m=1}^{d} \cos\left(\frac{2\pi\ell}{L}(x_m - x'_m)\right)$$

The kernel trick: Example 4

When $L \to \infty$, even if we cannot compute $\phi(x)$ (since it's a vector of *infinite dimension*), we can still compute inner product:

change order of summation & integral

$$\phi_\infty(\boldsymbol{x})^{\mathrm{T}}\phi_\infty(\boldsymbol{x}') = \int_0^{2\pi} \sum_{m=1}^{d} \cos(\theta(x_m - x'_m))\, d\theta$$

$$= \sum_{m=1}^{d} \frac{\sin(2\pi(x_m - x'_m))}{x_m - x'_m}$$

Again, a simple function of the original features.

Note that when using this mapping in linear regression, we are *learning a weight $\boldsymbol{w}^*$ with infinite dimension!*

Kernel functions

Definition: a function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is called a *kernel function* if there exists a function $\boldsymbol{\phi} : \mathbb{R}^d \to \mathbb{R}^M$ so that for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$,

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}')$$

Examples we have seen

$$k(\boldsymbol{x}, \boldsymbol{x}') = (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}')^2$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sum_{m=1}^{d} \frac{\sin(2\pi(x_m - x'_m))}{x_m - x'_m}$$

Using kernel functions

Choosing a nonlinear basis ϕ becomes equivalent to choosing a kernel function.

As long as computing the kernel function is more efficient, we should apply the kernel trick.

Gram/kernel matrix becomes:

$\phi(x_1)^T\phi(x_1)$ $\phi(x_1)^T\phi(x_n)$

$$\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} = \begin{pmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & k(\boldsymbol{x}_1, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ k(\boldsymbol{x}_2, \boldsymbol{x}_1) & k(\boldsymbol{x}_2, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_2, \boldsymbol{x}_n) \\ \vdots & \vdots & \vdots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & k(\boldsymbol{x}_n, \boldsymbol{x}_2) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{pmatrix}$$

In fact, k is a kernel if and only if $\boldsymbol{K}$ is positive semidefinite for *any* n *and any* $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n$ (**Mercer theorem**).

- useful for proving that a function is not a kernel

Examples which are not kernels

Function

$$k(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|_2^2$$

is *not a kernel*, why?

If it is a kernel, the kernel matrix for two data points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$: this entry is $\|x - x\|_2^2 = 0$

$$\boldsymbol{K} = \begin{pmatrix} 0 & \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 \\ \|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2^2 & 0 \end{pmatrix}$$

must be positive semidefinite, *but is it?*

$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$ is not psd. Why?

$$\begin{pmatrix} 1 & -1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = -2$$

Properties of kernels

For any function $f : \mathbb{R}^d \to \mathbb{R}$, $k(\boldsymbol{x}, \boldsymbol{x}') = f(\boldsymbol{x})f(\boldsymbol{x}')$ is a kernel. → What is ϕ?

$\phi : \mathbb{R}^d \to \mathbb{R}$, $\phi(x) = f(x)$

If $k_1(\cdot, \cdot)$ and $k_2(\cdot, \cdot)$ are kernels, then the following are also kernels:

- **conical combination**: $\alpha k_1(\cdot, \cdot) + \beta k_2(\cdot, \cdot)$ if $\alpha, \beta \geq 0$ → what is ϕ?
- **product**: $k_1(\cdot, \cdot) k_2(\cdot, \cdot)$ → HW2
- **exponential**: $e^{k(\cdot, \cdot)}$
- $\cdots$

ϕ_1 : map for k_1

ϕ_2 : map for k_2

ϕ' : map for $\alpha k_1(\cdot,\cdot) + \beta k_2(\cdot,\cdot)$

Exercise: find ϕ'

Verify using the definition of kernel!

Popular kernels

Polynomial kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \underbrace{(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{x}' + c)}_{\text{polynomial in original i.p.}}{}^{M}$$

for $c \geq 0$ and M is a positive integer.

What is the corresponding ϕ?

$c=0$, $M=2$, we saw earlier $\qquad \phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{pmatrix}$

The case of larger M can be obtained by applying this repeatedly.

Popular kernels

Gaussian kernel or Radial basis function (RBF) kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2\sigma^2}\right) \quad \text{for some } \sigma > 0.$$

What is the corresponding ϕ?

$$k(x, x') = \exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right) \exp\left(-\frac{\|x'\|_2^2}{2\sigma^2}\right) \exp\left(\frac{x^T x'}{\sigma^2}\right)$$

focus on this

$$k(x, x') = f(x) f(x')$$

where $f(x) = \exp\left(-\frac{\|x\|_2^2}{2\sigma^2}\right)$

transformation for the product.

Popular kernels

Gaussian kernel or Radial basis function (RBF) kernel

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2\sigma^2}\right) \quad \text{for some } \sigma > 0.$$

What is the corresponding ϕ?

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ldots$$

$$\exp\left(\frac{x^T x'}{\sigma^2}\right) = 1 + \frac{x^T x'}{\sigma^2} + \underbrace{\frac{1}{2!}\frac{(x^T x')^2}{(\sigma^2)^2}}_{} + \frac{1}{3!}\left(\frac{x^T x'}{\sigma^2}\right)^3 + \ldots$$

each of these is a polynomial kernel

∞ dimensional feature space !

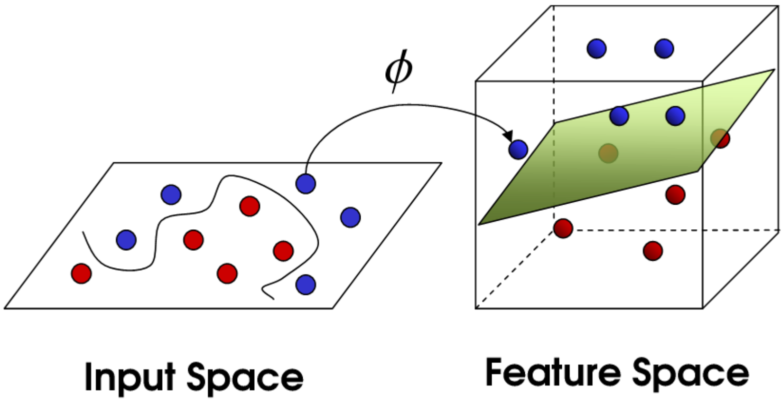
Prediction with kernels

As long as $\boldsymbol{w}^* = \sum_{i=1}^n \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)$, prediction on a new example $\boldsymbol{x}$ becomes

$$\boldsymbol{w}^{*\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i \boldsymbol{\phi}(\boldsymbol{x}_i)^{\mathrm{T}} \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{n} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x}).$$

This is known as a **non-parametric method**. Informally speaking, this means that there is no fixed set of parameters that the model is trying to learn (remember $\boldsymbol{w}^*$ could be infinite). Nearest-neighbors is another non-parametric method we have seen.

Classification with kernels



Similar ideas extend to the classification case, and we can predict using $\text{sign}(\boldsymbol{w}^T \boldsymbol{\phi})$.
Data may become linearly separable in the feature space!