DS-GA 1003: Machine Learning

Lab 5 Review
Features & Kernels

Presenter:
Yihuai Hong

Topics: Feature Maps · PSD Matrices · Representer Theorem

Today's Lab

01

High-Level Review: Feature Maps

~15 min

Why beyond $\mathbb{R}^d$ · Feature extraction · Geometric intuition

02

Deep Dive: PSD Matrices

~15 min

Definition · Two equivalent conditions · Connection to kernels

03

Representer Theorem

~20 min

Ridge Regression & SVM examples · General statement

SECTION 01

High-Level Review: Feature Maps

Slides 3–9 of the lecture

Input Space $\mathcal{X}$ — Going Beyond $\mathcal{X} = \mathbb{R}^d$

So far, $\mathcal{X} = \mathbb{R}^d$ for Ridge Regression, Lasso, and SVMs. Our hypothesis space was:

$$\mathcal{H} = \{ x \mapsto w^\top x + b \ : \ w \in \mathbb{R}^d, \ b \in \mathbb{R} \}$$

But what if inputs are NOT natively in $\mathbb{R}^d$?

T

Text Documents

Variable length, symbolic — no natural fixed-dim encoding

♪

Sound Recordings

Time-series of varying length and sample rate

🖼

Image Files

Resolution varies; spatial structure is important

G

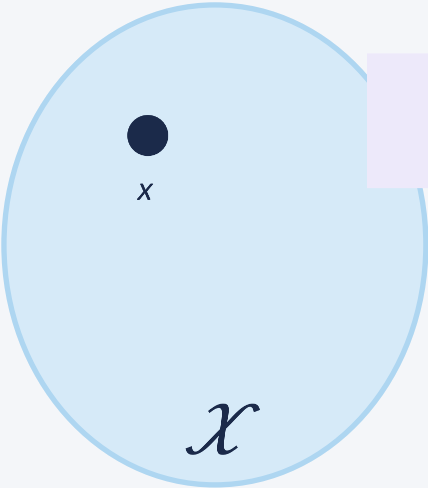
DNA Sequences

Alphabet {A,C,G,T} of arbitrary length

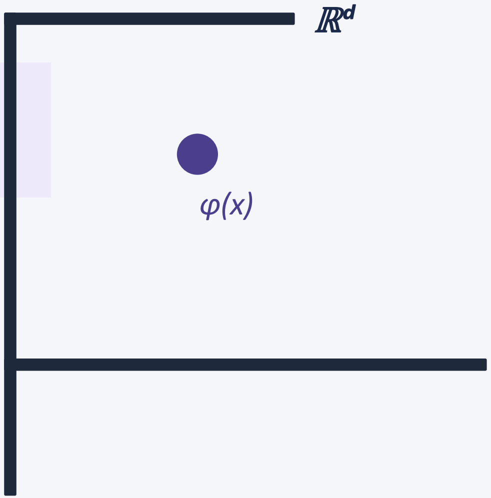
Problem: how do we feed these into a model that expects $x \in \mathbb{R}^d$?

Feature Extraction (Featurization)

Definition. Mapping an input from $\mathcal{X}$ to a vector in $\mathbb{R}^d$ is called **feature extraction** or **featurization.**



$$\phi : \mathcal{X} \longrightarrow \mathbb{R}^d$$



New hypothesis space:

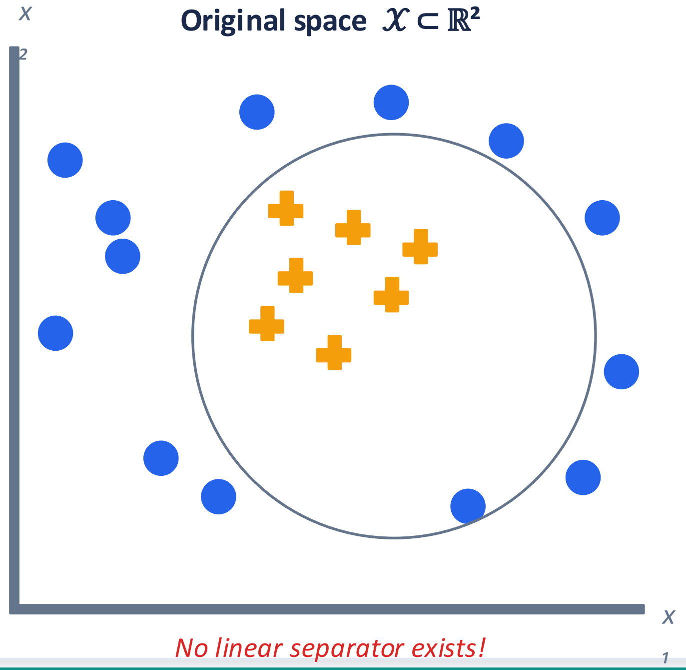
$$\mathcal{H} = \{\, x \mapsto w^\top \phi(x) + b \;:\; w \in \mathbb{R}^d,\; b \in \mathbb{R} \,\}$$

Geometric Example: Not Linearly Separable in $\mathbb{R}^2$

Binary classification in $\mathbb{R}^2$. Goal: find f_{w,b} such that f > 0 for y=+1 and f < 0 for y=−1.

$$f_{w,b}(x) = w^\top x + b$$

Original space $\mathcal{X} \subset \mathbb{R}^2$



No linear separator exists!

Class y = +1

Outer ring of points

Class y = −1

Inner cluster of points

Key Question:

Can a feature map ϕ transform the data so it becomes linearly separable?

$$\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$$

Geometric Example: Mapping to $\mathbb{R}^3$

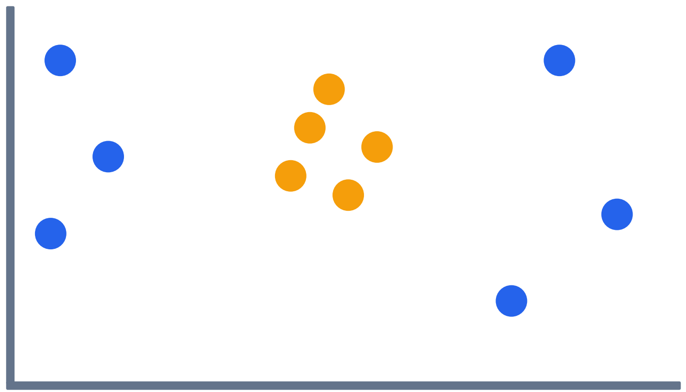
Data not linearly separable in lower-dim space might be separable in higher dimensions!

The Feature Map:

$$\phi : \mathbb{R}^2 \longrightarrow \mathbb{R}^3$$

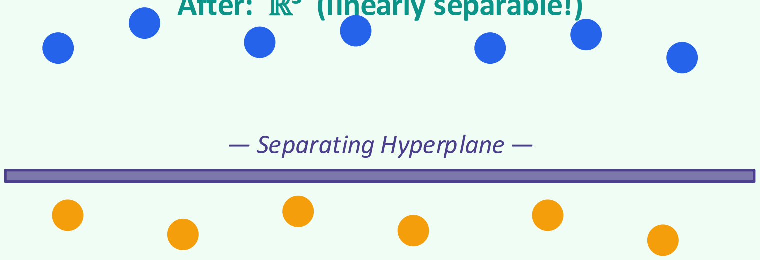
$$(x_1, x_2) \mapsto \left(x_1^2,\ \sqrt{2}\, x_1 x_2,\ x_2^2\right)$$

Before: $\mathbb{R}^2$ (not separable)



→ Apply ϕ

After: $\mathbb{R}^3$ (linearly separable!)



+1 above / −1 below

Goal: find ϕ s.t. data become linearly separable $\Longrightarrow$ fit SVM / Logistic Regression on $\phi(x)$

SECTION 02

Deep Dive:
PSD Matrices

The mathematical foundation of valid kernel functions

Positive Semidefinite (PSD) Matrix — Definition

Definition. A real, symmetric matrix $\mathbf{M} \in \mathbb{R}^{\mathbf{n \times n}}$ is **positive semidefinite (PSD)** if for any vector $x \in \mathbb{R}^n$:

$$x^\top M x \geq 0 \qquad \forall\, x \in \mathbb{R}^n$$

Two required properties for M:

- M is real-valued (entries in $\mathbb{R}$)
- M is symmetric: $M = M^T$

Intuition: The quadratic form $x^T M x$ generalises 'squared length.' PSD means M acts like a valid inner product — which can never be negative.

Why does PSD matter for kernels? A kernel matrix $K = (k(x^i, x^j))$ must be PSD for the kernel to correspond to a valid inner product in some feature space (Mercer's Theorem).

PSD Matrices — Two Equivalent Conditions

A symmetric matrix M is PSD if and only if EITHER condition holds (they are equivalent):

1 Factorization

$$M = R^\top R$$

M can be written as $R^T R$ for some matrix $R \in \mathbb{R}^{k\times n}$.

Intuition:

Think of R as encoding coordinates:

$$x^\top M x = x^\top R^\top R x = \|Rx\|^2 \geq 0$$

Known as the Cholesky decomposition.

2 Eigenvalues

$$\lambda_i(M) \geq 0 \quad \forall\, i$$

All eigenvalues of M are non-negative.

Intuition:

By spectral theorem $M = Q\Lambda Q^T$, so:

$$M = Q\Lambda Q^\top \Rightarrow x^\top M x = \|\Lambda^{1/2} Q^\top x\|^2 \geq 0 \iff \lambda_i \geq 0$$

Useful: check eigenvalues numerically.

These conditions are EQUIVALENT (iff) — checking either one is sufficient to establish PSD.

Proof: Factorization $\Longrightarrow$ PSD

Claim: If $M = R^T R$ for some matrix $R \in \mathbb{R}^{k\times n}$, then M is PSD.

Symmetry check: $(R^TR)^T = R^T(R^T)^T = R^TR = M$ ✓

Proof.

For any vector $x \in \mathbb{R}^n$, compute the quadratic form x^TMx step by step:

1. $x^T M x = x^T (R^TR)\, x$ ← *Substitute $M = R^TR$*
2. $= (Rx)^T (Rx)$ ← *$(AB)^T = B^TA^T$, so $x^TR^T \cdot Rx = (Rx)^T(Rx)$*
3. $= \|Rx\|^2$ ← *Definition of squared Euclidean norm*
4. ≥ 0 ← *Squared norm always non-negative* □

Key Insight: The proof converts x^TMx into $\|Rx\|^2$ — always ≥ 0. The matrix R encodes feature coordinates; the factorization structure guarantees PSD.
Cholesky factorization is the constructive characterization: given $M \sqsupseteq 0$, factor it as R^TR.

Proof: PSD $\Longrightarrow$ Factorization (Reverse Direction)

Claim: If M is PSD (real, symmetric, $x^T M x \geq 0$ for all x), then $M = R^T R$ for some R.

Key tool: Spectral Theorem — any real symmetric M admits $M = Q\Lambda Q^T$ (Q orthogonal, Λ diagonal)

Proof.

Apply the Spectral Theorem and construct R explicitly:

1 $M = Q\Lambda Q^T$ ← *Spectral theorem: Q orthogonal, $\Lambda = \mathrm{diag}(\lambda_1,\ldots,\lambda_n)$*

2 Define $R = \Lambda^{1/2} Q^T$ ← *where $\Lambda^{1/2} \stackrel{def}{=} \mathrm{diag}(\sqrt{\lambda_1},\ldots,\sqrt{\lambda_n})$; valid because:*

↳ *take $x=q_i$ (eigenvec.) in $x^T M x \geq 0$: $\lambda_i \|q_i\|^2 \geq 0 \Rightarrow \lambda_i \geq 0$ ✓*

3 $R^T R = Q\Lambda^{1/2} \cdot \Lambda^{1/2} Q^T = Q\Lambda Q^T = M$ □ ← *$Q^T Q = I$ (Q orthogonal), $\Lambda^{1/2} \cdot \Lambda^{1/2} = \Lambda$*

Key Insight: The non-negativity of eigenvalues ($\lambda_i \geq 0$) is not an extra assumption — it is derived from the PSD condition by substituting $x = q_i$. This is what makes $\Lambda^{1/2}$ real, and the construction valid.

Together: M is PSD $\Longleftrightarrow$ $M = R^T R$ (fully proved in both directions)

SECTION 03

Representer Theorem

Ridge Regression & SVM — then the general statement

SVM Dual — Solution in the Span of the Data

Given an optimal α*, the primal solution w* is:

SVM Dual:

$$\max_{\alpha} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^\top x^{(j)}$$

$$\text{s.t.} \quad \sum_{i=1}^{n} \alpha_i y^{(i)} = 0 \quad \text{and} \quad \alpha_i \in \left[0, \frac{C}{n}\right]$$

$$w^* = \sum_{i=1}^{n} \alpha_i^* y^{(i)} x^{(i)}$$

What this means:

w* is a linear combination of the training inputs $x^{(1)}$, ..., $x^{(n)}$.

Key terminology:

$$w^* \in \operatorname{span}(x^{(1)}, \ldots, x^{(n)})$$

Ridge Regression — Closed-Form Solution

Objective and closed-form solution:

$$\min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (w^\top x^{(i)} - y^{(i)})^2 + \lambda \|w\|^2$$

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

where $X \in \mathbb{R}^{nxd}$ is the design matrix, $y \in \mathbb{R}^n$ is the label vector.

Rearranging to show $w^* \in$ span of the data:

1. *Start:*

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

2. *Push-through identity:*

$$w^* = X^\top (X X^\top + \lambda I)^{-1} y \qquad \leftarrow n \times n \text{ inverse!}$$

3. *Let $\alpha^* = (XX^T + \lambda I)^{-1} y$:*

$$w^* = X^\top \alpha^* = \sum_{i=1}^{n} \alpha_i^* x^{(i)}$$

Same structure as SVM! w^* is again a linear combination of training inputs → $w^* \in \text{span}(x^{(1)}, \ldots, x^{(n)})$

Large Feature Spaces — Why Reparameterization Matters

Both SVM dual and reparameterized Ridge Regression solve for $\alpha^* \in \mathbb{R}^n$. When is this useful?

When $d \gg n$!

Concrete Example:

Training Set

$n = 300{,}000$

examples (fairly large)

vs.

Feature Space

$d = 300{,}000{,}000$

dimensions
e.g. high-degree monomials

Approach	Parameters to optimize	Cost
Solve for w directly	$w \in \mathbb{R}^d \rightarrow$ 300,000,000 params	Very expensive
Solve for α (dual)	**$\alpha \in \mathbb{R}^n \rightarrow$ 300,000 params**	**300× cheaper!**

The Representer Theorem

***Theorem** (Representer Theorem).* Suppose:

$$J(w) = R(\|w\|) + L(\langle w, x^{(1)}\rangle, \ldots, \langle w, x^{(n)}\rangle)$$

- $w, x^{(1)}, \ldots, x^{(n)} \in H$ for some Hilbert space H,
- $\|\cdot\|$ is the norm of H (i.e. $\|w\| = \sqrt{\langle w, w\rangle}$),
- $R : [0,\infty) \to \mathbb{R}$ is nondecreasing (the regularizer),
- $L : \mathbb{R}^n \to \mathbb{R}$ is arbitrary (the loss function).

Then, if J has a minimizer, there exists one of the form:

$$w^* = \sum_{i=1}^{n} \alpha_i x^{(i)}$$

1. We can always restrict the search to span($x^{(1)}, \ldots, x^{(n)}$) — no matter the loss function.
2. All norm-regularized linear models can be kernelized — we only need inner products $\langle x^{(i)}, x^{(j)}\rangle$.

Summary

01

Feature Maps $\phi : \mathcal{X} \rightarrow \mathbb{R}^d$

Turn arbitrary inputs (text, images, DNA) into fixed-length vectors. A good ϕ can make non-linearly-separable data separable in higher dimensions.

02

PSD Matrices

Symmetric M with $x^T M x \geq 0$. Equivalent to: (1) $M = R^T R$ or (2) all eigenvalues ≥ 0. Necessary for a valid kernel (Mercer's theorem).

03

Representer Theorem

For any norm-regularized objective, the minimizer lies in $\mathrm{span}(x^{(1)}, \ldots, x^{(n)})$. We can always reparameterize with $\alpha \in \mathbb{R}^n$ and kernelize.

Feature Maps + PSD Kernels + Representer Theorem → Kernel Methods