Lecture 3
Theory of Kernel Functions

Pavel Laskov[1] Blaine Nelson[1]

[1]Cognitive Systems Group
Wilhelm Schickard Institute for Computer Science
Universität Tübingen, Germany

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Advanced Topics in Machine Learning, 2012

Part I

Introduction: Kernel Functions

Overview

- In this lecture, we will formally define kernel functions
- *Recall*: advantages of kernel-based learning:
    1. Kernels allow for learning in high-dimensional feature spaces without explicit mapping into feature space
    2. Kernels make learning in high-dimensional feature spaces computationally feasible
    3. Kernel methods learn non-linear function with the machinery of algorithms for learning linear functions
    4. Kernels provide an abstraction that separates data representation & learning
- Questions to be addressed:
    1. What properties do kernels have & what properties does a function need to be a kernel?
    2. How can we verify that a kernel function is valid?
    3. How does one construct a kernel function?

- Consider a (slightly modified) feature space for 2-dimensional polynomials of degree 3:

$$\Phi(\mathbf{x}) = [x_1^3, x_2^3, \sqrt{3}x_1^2 x_2, \sqrt{3}x_1 x_2^2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{6}x_1 x_2, \sqrt{3}x_1, \sqrt{3}x_2, 1]$$

Recall: Kernel Magic

Example 2: 2-dimensional Polynomials of Degree 3

- Consider a (slightly modified) feature space for 2-dimensional polynomials of degree 3:

$$\Phi(\mathbf{x}) = [x_1^3, x_2^3, \sqrt{3}x_1^2 x_2, \sqrt{3}x_1 x_2^2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{6}x_1 x_2, \sqrt{3}x_1, \sqrt{3}x_2, 1]$$

- Let us compute the inner product between two points in the feature space:

$$\begin{aligned}\Phi(\mathbf{x})^\top \Phi(\mathbf{y}) &= x_1^3 y_1^3 + x_2^3 y_2^3 + 3x_1^2 x_2 y_1^2 y_2 + 3x_1 x_2^2 y_1 y_2^2 + 3x_1^2 y_1^2 + 3x_2^2 y_2^2 \\ &+ 6x_1 x_2 y_1 y_2 + 3x_1 y_1 + 3x_2 y_2 + 1 \\ &= (x_1 y_1 + x_2 y_2 + 1)^3\end{aligned}$$

- Consider a (slightly modified) feature space for 2-dimensional polynomials of degree 3:

$$\Phi(\mathbf{x}) = [x_1^3, x_2^3, \sqrt{3}x_1^2 x_2, \sqrt{3}x_1 x_2^2, \sqrt{3}x_1^2, \sqrt{3}x_2^2, \sqrt{6}x_1 x_2, \sqrt{3}x_1, \sqrt{3}x_2, 1]$$

- Let us compute the inner product between two points in the feature space:

$$\begin{aligned}\Phi(\mathbf{x})^\top \Phi(\mathbf{y}) &= x_1^3 y_1^3 + x_2^3 y_2^3 + 3x_1^2 x_2 y_1^2 y_2 + 3x_1 x_2^2 y_1 y_2^2 + 3x_1^2 y_1^2 + 3x_2^2 y_2^2 \\ &+ 6x_1 x_2 y_1 y_2 + 3x_1 y_1 + 3x_2 y_2 + 1 \\ &= (x_1 y_1 + x_2 y_2 + 1)^3\end{aligned}$$

- Complexity: 3 multiplications instead of 10.

Kernel Questions

- Which of the following functions are kernels?

$$\kappa_1(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{D} (x_i + z_i) \qquad \kappa_2(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{D} h\left(\frac{x_i - c}{a}\right) h\left(\frac{z_i - c}{a}\right)$$

$$\kappa_3(\mathbf{x}, \mathbf{z}) = -\frac{\langle \mathbf{x}, \mathbf{z} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{z}\|_2} \qquad \kappa_4(\mathbf{x}, \mathbf{z}) = \sqrt{\|\mathbf{x} - \mathbf{z}\|_2^2 + 1}$$

where $h(x) = \cos(1.75x)\exp(-x^2/2)$

Part II

Linear Algebra Review

Vector Space

Definition 1

A set $\mathcal{X}$ is a vector space (over the reals) if it is *closed* under an addition operator '$+$' (*i.e.*, $\forall\, \mathbf{x}, \mathbf{z} \in \mathcal{X} \quad \mathbf{x} + \mathbf{z} \in \mathcal{X}$) & a scalar multiplication operator '$\cdot$' (*i.e.*, $\forall\, \mathbf{x} \in \mathcal{X}, a \in \Re \quad a \cdot \mathbf{x} \in \mathcal{X}$) and these operators satisfy;

1. (Additive Associativity) $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$
2. (Additive Commutativity) $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$
3. (Additive Identity) $\exists\, \mathbf{0} \in \mathcal{X} \text{ s.t. } \forall\, \mathbf{u} \in \mathcal{X} \quad \mathbf{u} + \mathbf{0} = \mathbf{u}$
4. (Additive Inverse) $\forall\, \mathbf{u} \in \mathcal{X} \quad \exists -\mathbf{u} \in \mathcal{X} \text{ s.t. } \mathbf{u} + (-\mathbf{u}) = \mathbf{0}$
5. (Distibutivity) $a \cdot (\mathbf{u} + \mathbf{v}) = a \cdot \mathbf{u} + a \cdot \mathbf{v} \quad \& \quad (a + b) \cdot \mathbf{u} = a \cdot \mathbf{u} + b \cdot \mathbf{u}$
6. (Multiplicative Associativity) $a \cdot (b \cdot \mathbf{u}) = (a \cdot b) \cdot \mathbf{u}$
7. (Multiplicative Identity) $1 \cdot \mathbf{u} = \mathbf{u}$

Example: For any $D \in \aleph$, $\quad \Re^D$ is a vector space.

Vectors I

- A D-dimensional vector $\mathbf{x}$ is a list of D-reals in vector space $\Re^D$
- Vectors $\{\mathbf{x}_i\}_{i=1}^N$ are linearly dependent if there exists $c_1, c_2, \ldots, c_D$ (at least one not 0) such that

$$\sum_{i=1}^N c_i \cdot \mathbf{x}_i = \mathbf{0} \;\; ;$$

  otherwise, they are linearly independent.
- The inner product of $\mathbf{x}$ and $\mathbf{z}$ is defined as: $\mathbf{x}^\top \mathbf{z} = \sum_{i=1}^D x_i \cdot z_i$
  - Non-trivial vectors $\{\mathbf{x}_i\}_{i=1}^N$ are orthogonal if for $i \neq j$, $\mathbf{x}_i^\top \mathbf{x}_j = 0$. They are normal vectors if for all i, $\mathbf{x}_i^\top \mathbf{x}_i = 1$. They are orthonormal if both hold; *i.e.*, $\forall\, i, j \quad \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \delta_{i,j} = \mathrm{I}\left[i == j\right]$
  - A set of orthogonal vectors are *linearly independent*
  - Euclidean norm of a vector: $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$
  - Angle between vectors: $\theta\left(\mathbf{u}, \mathbf{v}\right) = \arccos \frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\| \, \|\mathbf{v}\|}$
  - Projection onto vector $\mathbf{z}$: $proj_{\mathbf{z}}(\mathbf{x}) = \frac{\langle \mathbf{z}, \mathbf{x} \rangle}{\langle \mathbf{z}, \mathbf{z} \rangle} \mathbf{z}$

Vectors II

- Vectors $\{\mathbf{x}_i\}_{i=1}^{N}$ spans $\mathcal{X}$ if for every $\mathbf{x} \in \mathcal{X}$, $\mathbf{x} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i$
  - $\{\mathbf{x}_i\}_{i=1}^{N}$ is a basis for $\mathcal{X}$ if it spans $\mathcal{X}$ & is *linearly independent*
  - The dimension of $\mathcal{X}$ is the number of elements in any basis of $\mathcal{X}$:
  - Every vector in $\mathcal{X}$ can be represented as its projection onto an orthonormal basis $\{\mathbf{x}_i\}_{i=1}^{N}$ of $\mathcal{X}$:

$$\mathbf{z} = \sum_{i=1}^{D} proj_{\mathbf{x}_i}(\mathbf{z}) = \sum_{i=1}^{D} \langle \mathbf{x}_i, \mathbf{z} \rangle \cdot \mathbf{x}_i$$

This is a Fourier decomposition of $\mathbf{z}$ in that basis.

Matrices

Matrices are a $N \times D$ grid of reals.

$$\mathbf{A} = \begin{bmatrix} A_{1,1} & A_{1,2} & \dots & A_{1,D} \\ A_{2,1} & A_{2,2} & \dots & A_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ A_{N,1} & A_{N,2} & \dots & A_{N,D} \end{bmatrix} = \begin{bmatrix} -\mathbf{A}_{1,\bullet}^\top- \\ -\mathbf{A}_{2,\bullet}^\top- \\ \vdots \\ -\mathbf{A}_{N,\bullet}^\top- \end{bmatrix} = \begin{bmatrix} | & | & & | \\ \mathbf{A}_{\bullet,1} & \mathbf{A}_{\bullet,2} & \dots & \mathbf{A}_{\bullet,D} \\ | & | & & | \end{bmatrix}$$

Matrix-Vector Multiplication ($\mathbf{x}$ is a D-vector & $\mathbf{z}$ is a N-vector):

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} -\mathbf{A}_{1,\bullet}^\top- \\ -\mathbf{A}_{2,\bullet}^\top- \\ \vdots \\ -\mathbf{A}_{N,\bullet}^\top- \end{bmatrix} \begin{bmatrix} | \\ \mathbf{x} \\ | \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1,\bullet}^\top \mathbf{x} \\ \mathbf{A}_{2,\bullet}^\top \mathbf{x} \\ \vdots \\ \mathbf{A}_{N,\bullet}^\top \mathbf{x} \end{bmatrix} \qquad \mathbf{z}^\top \mathbf{A} = \begin{bmatrix} \mathbf{z}^\top \mathbf{A}_{\bullet,1} & \mathbf{z}^\top \mathbf{A}_{\bullet,2} & \dots & \mathbf{z}^\top \mathbf{A}_{\bullet,1} \end{bmatrix}$$

Matrix-Matrix Multiplication:

$$\mathbf{AB} = \begin{bmatrix} -\mathbf{A}_{1,\bullet}^\top- \\ -\mathbf{A}_{2,\bullet}^\top- \\ \vdots \\ -\mathbf{A}_{N,\bullet}^\top- \end{bmatrix} \begin{bmatrix} | & | & & | \\ \mathbf{B}_{\bullet,1} & \mathbf{B}_{\bullet,2} & \dots & \mathbf{B}_{\bullet,D} \\ | & | & & | \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1,\bullet}^\top \mathbf{B}_{\bullet,1} & \mathbf{A}_{1,\bullet}^\top \mathbf{B}_{\bullet,2} & \dots & \mathbf{A}_{1,\bullet}^\top \mathbf{B}_{\bullet,D} \\ \mathbf{A}_{2,\bullet}^\top \mathbf{B}_{\bullet,1} & \mathbf{A}_{2,\bullet}^\top \mathbf{B}_{\bullet,2} & \dots & \mathbf{A}_{2,\bullet}^\top \mathbf{B}_{\bullet,D} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{A}_{N,\bullet}^\top \mathbf{B}_{\bullet,1} & \mathbf{A}_{N,\bullet}^\top \mathbf{B}_{\bullet,2} & \dots & \mathbf{A}_{N,\bullet}^\top \mathbf{B}_{\bullet,D} \end{bmatrix}$$

Matrices

- Matrix Multiplication as summations:

$$[\mathbf{Ax}]_i = \sum_\ell A_{i,\ell} x_\ell \qquad \left[\mathbf{z}^\top \mathbf{A}\right]_j = \sum_k z_k A_{k,j}$$

$$\mathbf{z}^\top \mathbf{Ax} = \sum_{k,\ell} z_k A_{k,\ell} x_\ell \qquad [\mathbf{AB}]_{i,k} = \sum_j A_{i,j} B_{j,k}$$

- Special forms of matrices:

Lower Triangular

$$\begin{bmatrix} \bullet & 0 & 0 & \dots & 0 \\ \bullet & \bullet & 0 & \dots & 0 \\ \bullet & \bullet & \bullet & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \bullet & \bullet & \bullet & \dots & \bullet \end{bmatrix}$$

Diagonal Matrix

$$\begin{bmatrix} \bullet & 0 & 0 & \dots & 0 \\ 0 & \bullet & 0 & \dots & 0 \\ 0 & 0 & \bullet & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \bullet \end{bmatrix}$$

Upper Triangular

$$\begin{bmatrix} \bullet & \bullet & \bullet & \dots & \bullet \\ 0 & \bullet & \bullet & \dots & \bullet \\ 0 & 0 & \bullet & \dots & \bullet \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \bullet \end{bmatrix}$$

Basic Linear Algebra

Suppose matrix $\mathbf{A}$ is an $N \times D$ matrix

- A square matrix has same # of rows & columns; *i.e.*, $N = D$
- The identity matrix $\mathbf{I}_N$ is a $N \times N$ diagonal matrix of 1's
  - For any $\mathbf{A}$, $\mathbf{A}\mathbf{I}_D = \mathbf{A}$ and $\mathbf{I}_N\mathbf{A} = \mathbf{A}$
- The transpose of $\mathbf{A}$ is denoted by $\mathbf{A}^\top$ (it is $D \times N$
- A symmetric matrix is its own transpose: $\mathbf{A} = \mathbf{A}^\top$
- The inverse of $\mathbf{A}$ is denoted by $\mathbf{A}^{-1}$: $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ & $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
- An orthonormal or unitary matrix is its own inverse: $\mathbf{A}\mathbf{A}^\top = \mathbf{A}^\top\mathbf{A} = \mathbf{I}$... both the columns & rows of $\mathbf{A}$ form a basis
- The rank of matrix $\mathbf{A}$ is the maximum number of columns of $\mathbf{A}$ that are *linearly independent* (*i.e.*, the dimension of its column space). $\mathbf{A}$ is full-rank if $rank(\mathbf{A}) = \min(M, N)$

Singularity

Definition 2

A matrix $\mathbf{A}$ is singular if there exists some $\mathbf{x} \neq \mathbf{0}$ such that $\mathbf{Ax} = \mathbf{0}$; otherwise, $\mathbf{A}$ is nonsingular.

Theorem 3

The following are equivalent:

- *Matrix* $\mathbf{A}$ *is invertible*
- *Matrix* $\mathbf{A}$ *is nonsingular*
- *Matrix* $\mathbf{A}$ *is full-rank*
- *The spectrum of* $\mathbf{A}$ *does not contain* 0*; i.e.,* $0 \notin eig(\mathbf{A})$
- *If* $\mathbf{A}$ *is square, the (linear) function* $f(\mathbf{x}) = \mathbf{Ax}$ *is one-to-one & onto,* $f(\mathbf{x}) = \mathbf{b}$ *has at least* 1 *solution, and* $f(\mathbf{x}) = \mathbf{0}$ *only has solution* $\mathbf{x} = \mathbf{0}$

Eigenvalues & Eigenvectors

- Given an $N \times N$ matrix $\mathbf{A}$, an eigenvector of $\mathbf{A}$ is a *non-trivial* vector $\mathbf{v}$ that satisfies

$$\mathbf{A}\mathbf{v} = \lambda \mathbf{v} \;\; ;$$

  the corresponding value λ is an eigenvalue

- The Rayleigh quotient is defined by

$$\lambda = \frac{\mathbf{v}^\top \mathbf{A}\mathbf{v}}{\mathbf{v}^\top \mathbf{v}}$$

- In fact, the maximum eigen-value/vector pair of $\mathbf{A}$ is a solution to

$$\max_{\|\mathbf{x}\|=1} \frac{\mathbf{x}^\top \mathbf{A}\mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

  with $\mathbf{x}$ restricted to have norm 1

Eigen-Decomposition & Deflation

- Deflation: for any eigen-value/vector pair $(\lambda,\mathbf{v})$ of $\mathbf{A}$, the transform

$$\tilde{\mathbf{A}} \leftarrow \mathbf{A} - \lambda \mathbf{v}\mathbf{v}^\top$$

deflates the matrix; *i.e.*, $\mathbf{v}$ is an eigenvector of $\tilde{\mathbf{A}}$ but has eigenvalue 0

- A *symmetric* matrix has N *orthonormal* eigenvectors $\{\mathbf{v}_i\}$ corresponding to N eigenvalues—its spectrum; *eig* $(\mathbf{A})$

$$\lambda_1(\mathbf{A}) \geq \lambda_2(\mathbf{A}) \geq \ldots \geq \lambda_N(\mathbf{A})$$

- Eigen-vectors/values form orthonormal matrix $\mathbf{V}$ & diagonal matrix $\mathbf{\Lambda}$

$$\mathbf{V} = \begin{bmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_N \\ | & | & & | \end{bmatrix} \qquad \mathbf{\Lambda} = \begin{bmatrix} \lambda_1(\mathbf{A}) & 0 & \ldots & 0 \\ 0 & \lambda_2(\mathbf{A}) & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & & \lambda_N(\mathbf{A}) \end{bmatrix}$$

which form the eigen-decomposition of $\mathbf{A}$: $\quad \mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$

The Spectral Theorem

Theorem 4

If $\mathbf{A}$ *is a symmetric* $N \times N$ *real-valued matrix, it can be written as*

$$\mathbf{A} = \sum_{i=1}^{N} \lambda_N(\mathbf{A})\, \mathbf{v}_i \mathbf{v}_i^\top$$

where $(\lambda_i, \mathbf{v}_i)$ *are eigen-value/vector pairs of* $\mathbf{A}$*. This is called the spectral decomposition of* $\mathbf{A}$

- For a matrix with rank $K < N$, the spectral decomposition of $\mathbf{A}$ only has K summands

Matrix Functions

- Properties of *diagonal* matrix $\mathbf{D}$ with entries $D_{i,i}$:
  - For $k = 0, 1, 2, \ldots$: $\mathbf{D}^k$ is diagonal with entries $[\mathbf{D}^k]_{i,i} = (D_{i,i})^k$
  - If $\nexists\, i$ s.t. $D_{i,i} = 0$ then $\mathbf{D}^{-1}$ exists, is diagonal, & $[\mathbf{D}^{-1}]_{i,i} = (D_{i,i})^{-1}$
  - $\sqrt{\mathbf{D}}$ is diagonal with entries $[\sqrt{\mathbf{D}}]_{i,i} = \sqrt{D_{i,i}}$
- Functions of $\mathbf{A}$ are defined by its eigen-decomposition $\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$ & the fact that $\mathbf{V}^\top\mathbf{V} = \mathbf{I}$
  - For $k = 0, 1, 2, \ldots$ $\mathbf{A}^k = \mathbf{V}\boldsymbol{\Lambda}^k\mathbf{V}^\top$
  - If $\mathbf{A}$ is *non-singular*, then $\mathbf{A}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^\top$
  - $\sqrt{\mathbf{A}} = \mathbf{V}\sqrt{\boldsymbol{\Lambda}}\mathbf{V}^\top$ (Note: this satisfies $\sqrt{\mathbf{A}}\sqrt{\mathbf{A}} = \mathbf{A}$)
  - $\exp(\mathbf{A}) = \mathbf{V}\exp(\boldsymbol{\Lambda})\,\mathbf{V}^\top$
  - $\log(\mathbf{A}) = \mathbf{V}\log(\boldsymbol{\Lambda})\,\mathbf{V}^\top$

Part III

Positive (Semi-)Definiteness

Positive (Semi-)Definite Matrices

Definition 5 (Positive Semi-Definite Matrix)

Matrix $\mathbf{A}$ is positive semi-definite (PSD) if all its eigenvalues are non-negative ($\forall\, i \quad \lambda_i(\mathbf{A}) \geq 0$); *i.e.*, for all $\mathbf{x} \in \mathcal{X}$:

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} \geq 0$$

from the Rayleigh quotient. We use $\mathbf{A} \succeq 0$ to denote that $\mathbf{A}$ *is PSD*

Definition 6 (Positive Definite Matrix)

Matrix $\mathbf{A}$ is positive definite if all its eigenvalues are positive ($\forall\, i \quad \lambda_i(\mathbf{A}) > 0$); *i.e.*, $\mathbf{A}$ is PSD &

$$\mathbf{x}^\top \mathbf{A}\mathbf{x} = 0 \;\Leftrightarrow\; \mathbf{x} = \mathbf{0}$$

We denote this as $\mathbf{A} \succ 0$.

PSD Matrices

Proposition 7

Matrix $\mathbf{A}$ *is PSD iff there exists a real matrix* $\mathbf{B}$ *such that* $\mathbf{A} = \mathbf{B}^\top \mathbf{B}$

Proof

Case $\Leftarrow$: Suppose $\mathbf{A} = \mathbf{B}^\top \mathbf{B}$, then for any $\mathbf{x}$

$$\mathbf{x}^\top \mathbf{A} \mathbf{x} = \mathbf{x}^\top \mathbf{B}^\top \mathbf{B} \mathbf{x} = \|\mathbf{B}\mathbf{x}\|^2 \geq 0$$

Case $\Rightarrow$: If $\mathbf{A} \succeq 0$ then its eigen-decomposition ($\mathbf{A} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top$) has only non-negative eigenvalues and thus, $\sqrt{\boldsymbol{\Lambda}}$ is a real-valued matrix. Thus, let $\mathbf{B} = \sqrt{\boldsymbol{\Lambda}}\mathbf{V}^\top$ and we have

$$\mathbf{B}^\top \mathbf{B} = \mathbf{V}\sqrt{\boldsymbol{\Lambda}}\sqrt{\boldsymbol{\Lambda}}\mathbf{V}^\top = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\top = \mathbf{A}$$

$\square$

Part IV

Reproducing Kernel Hilbert Spaces

Inner Product Space

Definition 8

An inner product space $\mathcal{X}$ is a vector space with an associated inner product $\langle \cdot, \cdot \rangle : \mathcal{X} \times \mathcal{X} \to \Re$ that satisfies:

1. (Symmetry) $\langle \mathbf{x}, \mathbf{z} \rangle = \langle \mathbf{z}, \mathbf{x} \rangle$
2. (Linearity) $\langle a \cdot \mathbf{x}, \mathbf{z} \rangle = a \cdot \langle \mathbf{x}, \mathbf{z} \rangle \quad \& \quad \langle \mathbf{w} + \mathbf{x}, \mathbf{z} \rangle = \langle \mathbf{w}, \mathbf{z} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$
3. (Positive Semi-Definiteness) $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$

The inner product space is strict if $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \quad \Leftrightarrow \quad \mathbf{x} = \mathbf{0}$

- A strict inner product space $\mathcal{X}$ has a natural norm given by $\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$. The associated metric is $d(\mathbf{x}, \mathbf{z}) = \|\mathbf{x} - \mathbf{z}\|_2$
- The space $\Re^D$ has the inner product $\langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x}^\top \mathbf{z}$ which yields the Euclidean norm:

$$\|\mathbf{x}\|_2^2 = \sum_{i=1}^{D} x_i^2$$

Hilbert Space

Definition 9

A *strict inner product space* $\mathcal{F}$ is a Hilbert space if it is

1. **Complete**: Every (Cauchy) sequence $\{h_i \in \mathcal{F}\}_{i=1}^{\infty}$ such that

$$\lim_{n\to\infty} \sup_{m>n} \|h_n - h_m\| = 0$$

   converges to an element $h \in \mathcal{F}$; *i.e.*, $h_i \to h$

2. **Separable**: There is a *countable* subset $\hat{\mathcal{F}} = \{h_i \in \mathcal{F}\}_{i=1}^{\infty}$ such that for all $h \in \mathcal{F}$ and $\epsilon > 0$, there exists $h_i \in \hat{\mathcal{F}}$ such that

$$\|h_i - h\| < \epsilon$$

Hilbert Space Examples: the interval $[0, 1]$, the reals $\Re$, the complex numbers $\mathcal{C}$, & Euclidean spaces $\Re^D$ for $D \in \aleph$.

Hilbert Space

Definition 9

A *strict inner product space* $\mathcal{F}$ is a Hilbert space if it is

1. **Complete**: Every (Cauchy) sequence $\{h_i \in \mathcal{F}\}_{i=1}^{\infty}$ such that

   Technical Condition required for potentially infinite-dimensional sets

   $$\lim_{n\to\infty} \sup_{m>n} \|h_n - h_m\| = 0$$

   converges to an element $h \in \mathcal{F}$; *i.e.*, $h_i \to h$

2. **Separable**: There is a *countable* subset $\hat{\mathcal{F}} = \{h_i \in \mathcal{F}\}_{i=1}^{\infty}$ such that for all $h \in \mathcal{F}$ and $\epsilon > 0$, there exists $h_i \in \hat{\mathcal{F}}$ such that

   $$\|h_i - h\| < \epsilon$$

Hilbert Space Examples: the interval $[0, 1]$, the reals $\Re$, the complex numbers $\mathcal{C}$, & Euclidean spaces $\Re^D$ for $D \in \aleph$.

Hilbert Space

Definition 9

A *strict inner product space* $\mathcal{F}$ is a Hilbert space if it is

1. **Complete**: Every (Cauchy) sequence $\{h_i \in \mathcal{F}\}_{i=1}^{\infty}$ such that

   Technical Condition required for potentially infinite-dimensional sets

   $$\lim_{n\to\infty} \sup_{m>n} \|h_n - h_m\| = 0$$

   converges to an element $h \in \mathcal{F}$; *i.e.*, $h_i \to h$

2. **Separable**: There is a *countable* subset $\hat{\mathcal{F}} = \{h_i \in \mathcal{F}\}_{i=1}^{\infty}$ such that for all $h \in \mathcal{F}$ and $\epsilon > 0$, there exists $h_i \in \hat{\mathcal{F}}$ such that

   Condition required to make Hilbert space isomorphisms

   $$\|h_i - h\| < \epsilon$$

Hilbert Space Examples: the interval $[0, 1]$, the reals $\Re$, the complex numbers $\mathcal{C}$, & Euclidean spaces $\Re^D$ for $D \in \aleph$.

Function Spaces

- *What is a vector?* An ordered list of D elements from $\Re$ indexed by the index set $\mathbb{I}_D = \{1, 2, \ldots, D\}$. The set of all such lists is $\Re^D$
- We can extend this notion to countable sequences $\mathbf{x} = (x_1, x_2, \ldots)$ by using the index set $\mathbb{I} = \aleph$
  - Inner-product generalizes naturally as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i \in \aleph} x_i \cdot y_i$$

  - However, we need additional restrictions to make such spaces well-behaved
  - The subspace ℓ^2 for which $\forall\, \mathbf{x}\ \langle \mathbf{x}, \mathbf{x} \rangle < \infty$ is a Hilbert space
- Further, a function $f : \mathcal{X} \to \Re$ maps each $\mathbf{x} \in \mathcal{X}$ to exactly one $y \in \Re$; *i.e.*, it is also *a vector with an uncountable index set* (*e.g.*, $\mathbb{I} = \Re^D$)
  - Inner-product again generalizes naturally as

$$\langle f, g \rangle = \int_{\mathcal{X}} f(\mathbf{x})\, g(\mathbf{x})\, \mathrm{d}\mathbf{x}$$

  - Subspace $\mathcal{L}_2(\mathcal{X})$ defined on $\mathcal{X}$, a *compact* subspace of $\Re^D$, for which $\forall\, f \in \mathcal{L}_2(\mathcal{X}),\ \langle f, f \rangle < \infty$ is a Hilbert space

Properties of (Separable) Hilbert Spaces

- Hilbert space $\mathcal{F}$ is isomorphic to $\mathcal{H}$ if there is a one-to-one linear mapping $T : \mathcal{F} \to \mathcal{H}$ such that for all $\mathbf{x}, \mathbf{z} \in \mathcal{F}$

$$\langle T(\mathbf{x}), T(\mathbf{z}) \rangle_{\mathcal{H}} = \langle \mathbf{x}, \mathbf{z} \rangle_{\mathcal{F}}$$

- Every *separable* Hilbert space **(A)** of dimension D is isomorphic to $\Re^D$ and **(B)** of infinite dimension is isomorphic to ℓ^2
- Since Hilbert space $\mathcal{F}$ is isomorphic to $\Re^D$ or ℓ^2, $\mathcal{F}$ has an orthonormal basis $\{\phi_i\}$ & element in $\mathbf{x} \in \mathcal{F}$ have a Fourier decomposition:

$$\mathbf{x} = \sum_i \langle \phi_i, \mathbf{x} \rangle_{\mathcal{F}} \cdot \phi_i$$

Part V

Characterizing Kernel Functions

Kernel Terminology

Definition 10

A kernel is a two-argument real-valued function over $\mathcal{X} \times \mathcal{X}$ ($\kappa : \mathcal{X} \times \mathcal{X} \to \Re$) such that for any $\mathbf{x}, \mathbf{z} \in \mathcal{X}$

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle_{\mathcal{F}} \tag{1}$$

for some inner-product space $\mathcal{F}$ such that $\forall\, \mathbf{x} \in \mathcal{X} \quad \phi(\mathbf{x}) \in \mathcal{F}$

- Kernel functions must be symmetric since inner products are symmetric
- *To show that κ is a valid kernel, it is sufficient to show that a mapping ϕ exists that yields Eq. 1. However, this is generally difficult to construct.*
- In this rest of this lecture, we will demonstrate additional ways to construct & validate kernels

Kernel Matrices

Definition 11

A kernel matrix (or Gram matrix) $\mathbf{K}$ is the matrix that results from applying κ to all pairs of datapoints in set $\{\mathbf{x}_i\}_{i=1}^{N}$

$$\mathbf{K} = \begin{bmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \kappa(\mathbf{x}_1, \mathbf{x}_2) & \dots & \kappa(\mathbf{x}_1, \mathbf{x}_N) \\ \kappa(\mathbf{x}_2, \mathbf{x}_1) & \kappa(\mathbf{x}_2, \mathbf{x}_2) & \dots & \kappa(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_N, \mathbf{x}_1) & \kappa(\mathbf{x}_N, \mathbf{x}_2) & \dots & \kappa(\mathbf{x}_N, \mathbf{x}_N) \end{bmatrix}$$

that is, $K_{i,j} = \kappa(\mathbf{x}_i, \mathbf{x}_j)$

- Kernel matrices are square & symmetric

Kernel Matrices are PSD

Proposition 12

Kernel matrices, which are constructed from a kernel corresponding to a strict inner product space $\mathcal{F}$, are PSD.

Proof

By definition of a kernel matrix, for all $i, j \in 1, \ldots, N$

$$K_{i,j} = \kappa\left(\mathbf{x}_i, \mathbf{x}_j\right) = \left\langle \phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{x}_j\right)\right\rangle_{\mathcal{F}}$$

Thus, for any $\mathbf{v} \in \Re^N$:

$$\begin{aligned}
\mathbf{v}^\top \mathbf{K} \mathbf{v} &= \sum_{i,j} v_i K_{i,j} v_j = \sum_{i,j} v_i \left\langle \phi\left(\mathbf{x}_i\right), \phi\left(\mathbf{x}_j\right)\right\rangle_{\mathcal{F}} v_j \\
&= \left\langle \sum_i v_i \phi\left(\mathbf{x}_i\right), \sum_j v_j \phi\left(\mathbf{x}_j\right)\right\rangle_{\mathcal{F}} \\
&= \left\| \sum_i v_i \phi\left(\mathbf{x}_i\right)\right\|_{\mathcal{F}}^2 \geq 0
\end{aligned}$$

$\square$

Reproducing Kernel Function

Definition 13 (Reproducing Kernel Function (Aronszajn, 1950) [1])

Suppose $\mathcal{F}$ is a Hilbert space of functions over $\mathcal{X}$; the function $\kappa : \mathcal{X} \times \mathcal{X} \to \Re$ is a reproducing kernel of $\mathcal{F}$ if

1. For every $\mathbf{x} \in \mathcal{X}$, the function $f_{\mathbf{x}}(\cdot) = \kappa(\cdot, \mathbf{x})$ is in $\mathcal{F}$.
2. Reproducing Property: for every $\mathbf{z} \in \mathcal{X}$ and every $f \in \mathcal{F}$

$$f(\mathbf{z}) = \langle f, \kappa(\cdot, \mathbf{z}) \rangle_{\mathcal{F}}$$

Further, the space is called a Reproducing Kernel Hilbert Space (RKHS)

- By 1^{st} property & closure of $\mathcal{F}$, for any $\alpha_i \in \Re$ and $\mathbf{x}_i \in \mathcal{X}$, we have

$$\textstyle\sum_{i=1}^{N} \alpha_i \cdot \kappa(\cdot, \mathbf{x}_i) \in \hat{\mathcal{X}}$$

- Applying $f_{\mathbf{x}}$ from 1^{st} property to 2^{nd} property, for any $\mathbf{x}, \mathbf{z} \in \mathcal{X}$, we have

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \kappa(\cdot, \mathbf{x}), \kappa(\cdot, \mathbf{z}) \rangle_{\mathcal{F}}$$

Kernel Functions

Definition 14 (Finitely Positive Semi-definite)

A function $\kappa : \mathcal{X} \times \mathcal{X} \to \Re$ is finitely positive semi-definite (FPSD) if

- It is *symmetric*; *i.e.*, $\forall\, \mathbf{x}, \mathbf{z} \in \mathcal{X} \quad \kappa(\mathbf{x}, \mathbf{z}) = \kappa(\mathbf{z}, \mathbf{x})$
- The matrix $\mathbf{K}$ formed by applying κ to *any* finite subset of $\mathcal{X}$ is positive semi-definite: $\mathbf{K} \succeq 0$

Kernel Functions I

Theorem 15

$\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \Re$ (either continuous or with a countable domain) is FPSD iff $\exists$ Hilbert space $\mathcal{F}$ with feature map $\phi : \mathcal{X} \rightarrow \mathcal{F}$ s.t.

$$\kappa(\mathbf{x}, \mathbf{z}) = \langle \phi(\mathbf{x}), \phi(\mathbf{z}) \rangle$$

Proof

Case $\Leftarrow$: Follows from Proposition 12.

Case $\Rightarrow$: Suppose κ if FPSD & we construct Hilbert Space $\mathcal{F}_\kappa$ with κ as its reproducing kernel; *i.e.*, $\mathcal{F}_\kappa$ is the closure of functions: $f_{\mathbf{x}}(\cdot) = \kappa(\cdot, \mathbf{x})$. Thus, for any α_i, $\mathbf{x}_i$, $g(\cdot) = \sum_i \alpha_i \kappa(\cdot, \mathbf{x}_i)$ is in $\mathcal{F}_\kappa$ &, by the reproducing property,

$$\langle g, g \rangle = \sum_{i,j} \alpha_i \alpha_j \kappa(\mathbf{x}_i, \mathbf{x}_j) = \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}$$

where $\mathbf{K}$ is the kernel matrix $\{\mathbf{x}_i\}$, & thus $\boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} \geq 0$ since $\mathbf{K} \succeq 0$.

Kernel Functions II

(**Completeness**) Follows from the Cauchy-Schwarz inequality, but beyond the scope of this course.

(**Separability**) Separability follows from κ being continuous or having a countable domain, but is not shown here.

Finally, the mapping ϕ is specified by κ and $\phi(\mathbf{x}) = \kappa(\cdot, \mathbf{x}) \in \mathcal{F}_\kappa$. □

Note, the inner product defined above is *strict* since if $\|f\| = 0$, then for all $\mathbf{x}$, $|f(\mathbf{x})| \leq \|f\| \, \|\phi(\mathbf{x})\| = 0$

Part VI

Kernel Constructions

Simple Kernels

- Clearly, the linear kernel defined by

$$\kappa_{\mathrm{lin}}(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x}, \mathbf{z} \rangle = \mathbf{x}^\top \mathbf{z}$$

  is a valid kernel function since it is an inner product in $\mathcal{X}$

- For any $N \times N$ matrix $\mathbf{B} \succeq 0$,

$$\kappa_{\mathbf{B}}(\mathbf{x}, \mathbf{z}) = \langle \mathbf{x} | \mathbf{B} | \mathbf{z} \rangle = \mathbf{x}^\top \mathbf{B} \mathbf{z}$$

  is a valid kernel function

Closure Properties of Kernels I

Proposition 16

Suppose κ_1 & κ_2 are kernels on $\mathcal{X}$, $a > 0$, $f : \mathcal{X} \to \Re$, $\phi : \mathcal{X} \to \Re^M$, & κ_3 is a kernel on $\Re^M$. Then these are all kernel functions on $\mathcal{X}$:

16.1 $\kappa(\mathbf{x}, \mathbf{z}) = \kappa_1(\mathbf{x}, \mathbf{z}) + \kappa_2(\mathbf{x}, \mathbf{z})$

16.2 $\kappa(\mathbf{x}, \mathbf{z}) = a \cdot \kappa_1(\mathbf{x}, \mathbf{z})$

16.3 $\kappa(\mathbf{x}, \mathbf{z}) = \kappa_1(\mathbf{x}, \mathbf{z}) \cdot \kappa_2(\mathbf{x}, \mathbf{z})$

16.4 $\kappa(\mathbf{x}, \mathbf{z}) = f(\mathbf{x})\, f(\mathbf{z})$

16.5 $\kappa(\mathbf{x}, \mathbf{z}) = \kappa_3(\phi(\mathbf{x}), \phi(\mathbf{z}))$

Closure Properties of Kernels II

Proof

Let $\mathbf{K}_1$ & $\mathbf{K}_2$ be the kernel matrices of κ_1 & κ_2 applied to any set $\{\mathbf{x}_i\}_{i=1}^N$—both these matrices are PSD. Also let $\boldsymbol{\alpha}$ be any N-vector:

(**Part 1**): $\mathbf{K} = \mathbf{K}_1 + \mathbf{K}_2 \quad \Rightarrow \quad \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} = \boldsymbol{\alpha}^\top \mathbf{K}_1 \boldsymbol{\alpha} + \boldsymbol{\alpha}^\top \mathbf{K}_2 \boldsymbol{\alpha} \geq 0$
(**Part 2**): $\mathbf{K} = a\mathbf{K}_1 \quad \Rightarrow \quad \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha} = a \cdot \boldsymbol{\alpha}^\top \mathbf{K}_1 \boldsymbol{\alpha} \geq 0$
(**Part 3**): Take the spectral decomposition of $\mathbf{K}_1 = \sum_{i=1}^N \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$ and $\mathbf{K}_2 = \sum_{i=1}^N \gamma_i \mathbf{w}_i \mathbf{w}_i^\top$. The spectral decomposition of their element-wise product, $\mathbf{K} = \mathbf{K}_1 \odot \mathbf{K}_2$, is then $\mathbf{K} = \sum_{i,j=1}^N \sqrt{\lambda_i \gamma_j} (\mathbf{v}_i \odot \mathbf{w}_j)(\mathbf{v}_i \odot \mathbf{w}_j)^\top$; *i.e.*, a summation of rank-1 matrices with positive coefficients $\Rightarrow$ PSD.
(**Part 4**): $\kappa(\mathbf{x}, \mathbf{z}) = \langle \psi(\mathbf{x}), \psi(\mathbf{z}) \rangle$ where $\psi : \mathbf{x} \mapsto f(\mathbf{x})$; thus, κ is PSD.
(**Part 5**): Since κ_3 is a kernel, applying it to any set of vectors $\{\phi(\mathbf{x}_i)\}_{i=1}^N$ yields a PSD matrix. $\square$

Closure Properties of Kernels III

The feature spaces for these kernels are as follows:

- For kernel $\kappa_1(\mathbf{x}, \mathbf{z}) + \kappa_2(\mathbf{x}, \mathbf{z})$, the new feature map is equivalent to stacking the feature maps of κ_1 & κ_2:

$$\phi(\mathbf{x}) = \begin{bmatrix} \phi_1(\mathbf{x}) \\ \phi_2(\mathbf{x}) \end{bmatrix}$$

- For kernel $a \cdot \kappa_1(\mathbf{x}, \mathbf{z})$, its feature space is scaled by $\sqrt{a}$
- For kernel $\kappa_1(\mathbf{x}, \mathbf{z}) \cdot \kappa_2(\mathbf{x}, \mathbf{z})$, if ϕ_1 has dimension N_1 and ϕ_2 has dimension N_2, ϕ has $N_1 N_2$ features given by

$$[\phi(\mathbf{x})]_{ij} = [\phi_1(\mathbf{x})]_i \cdot [\phi_2(\mathbf{x})]_j$$

- It follows that the features of $\kappa_1(\mathbf{x}, \mathbf{z})^d$ are all monomials of the form

$$[\phi_1(\mathbf{x})]_1^{d_1} [\phi_1(\mathbf{x})]_2^{d_2} \ldots [\phi_1(\mathbf{x})]_N^{d_N} \qquad \sum_i d_i = d$$

Additional Kernel Functions

Proposition 17

Suppose κ_1 is a kernel on $\mathcal{X}$ & $p : \Re \to \Re$ is a polynomial with non-negative coefficients. Then, the following are kernels:

1. $\kappa(\mathbf{x}, \mathbf{z}) = p(\kappa_1(\mathbf{x}, \mathbf{z}))$
2. $\kappa(\mathbf{x}, \mathbf{z}) = \exp(\kappa_1(\mathbf{x}, \mathbf{z}))$
3. *Gaussian or RBF kernel:* $\kappa(\mathbf{x}, \mathbf{z}) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{z}\|_2^2}{2\sigma^2}\right)$

Proof

(**Part 1**) Constructing a polynomial kernel from base kernel κ_1 proceeds directly from Proposition 16.1, 16.2, & 16.3
(**Part 2**) Consider that $\exp(x) = 1 + x + \frac{1}{2}x^2 + \ldots \frac{1}{i!}x^i + \ldots$. Thus, it is a limit of polynomials & the PSD property is closed under pointwise limits.
(**Part 3**) Left as an exercise. □

Common Kernel Functions

- Linear Kernel: $\kappa_{\mathrm{lin}}(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{z}$
- Polynomial Kernel: $\kappa_{\mathrm{poly}}(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z} + R)^d$
- RBF Kernel: $\kappa_{\mathrm{rbf}}(\mathbf{x}, \mathbf{z}) = \exp\left(\frac{-\|\mathbf{x}-\mathbf{z}\|_2^2}{2\sigma^2}\right)$

- Which of the following functions are kernels?

$$\kappa_1(\mathbf{x}, \mathbf{z}) = \sum_{i=1}^{D} (x_i + z_i) \qquad \kappa_2(\mathbf{x}, \mathbf{z}) = \prod_{i=1}^{D} h(\tfrac{x_i - c}{a}) h(\tfrac{z_i - c}{a})$$

$$\kappa_3(\mathbf{x}, \mathbf{z}) = -\frac{\langle \mathbf{x}, \mathbf{z} \rangle}{\|\mathbf{x}\|_2 \|\mathbf{z}\|_2} \qquad \kappa_4(\mathbf{x}, \mathbf{z}) = \sqrt{\|\mathbf{x} - \mathbf{z}\|_2^2 + 1}$$

  where $h(x) = \cos(1.75x)\exp(-x^2/2)$
- κ_1 is *not* a kernel. Consider $\mathbf{x}_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$ & $\mathbf{x}_2 = \begin{bmatrix} 0 & 2 \end{bmatrix}^\top$. Their kernel matrix has eigenvalues -1 and 5.
- κ_2 is a kernel because it can be written as the product $f(\mathbf{x})f(\mathbf{z})$ where $f(x) = \prod_{i=1}^{D} h(\frac{x_i - c}{a})$
- κ_3 is *not* a kernel because it is the negation of a valid non-trivial kernel & thus will have negative eigenvalues
- κ_4 is *not* a kernel. Consider $\mathbf{x}_1 = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$ & $\mathbf{x}_2 = \begin{bmatrix} 0 & 1 \end{bmatrix}^\top$. Again, their kernel matrix has a negative eigenvalue

Part VII

Transforming Kernel Matrices

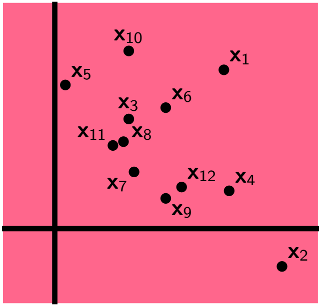
- *Adding a non-negative constant to the Kernel Matrix*: corresponds to adding a new constant feature to each training example; *i.e.*, given the matrix $\mathbf{\Phi}$ of features such that $\mathbf{K} = \mathbf{\Phi}\mathbf{\Phi}^\top$,

$$\begin{bmatrix} \mathbf{\Phi} & c\mathbf{1} \end{bmatrix} * \begin{bmatrix} \mathbf{\Phi} & c\mathbf{1} \end{bmatrix}^\top = \mathbf{K} + c^2 \mathbf{1}\mathbf{1}^\top$$

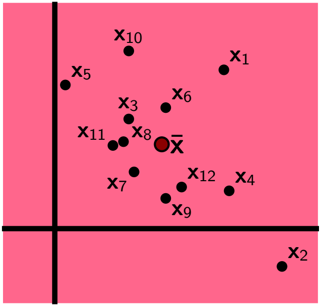
- *Adding a non-negative constant to its diagonal*: corresponds to adding an *indicator* feature for every data point

$$\begin{bmatrix} \phi(\mathbf{x}_1) & c & 0 & \dots & 0 \\ \phi(\mathbf{x}_2) & 0 & c & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_N) & 0 & 0 & \dots & c \end{bmatrix} \begin{bmatrix} \phi(\mathbf{x}_1) & c & 0 & \dots & 0 \\ \phi(\mathbf{x}_2) & 0 & c & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi(\mathbf{x}_N) & 0 & 0 & \dots & c \end{bmatrix}^\top = \mathbf{K} + c^2 \mathbf{I}$$

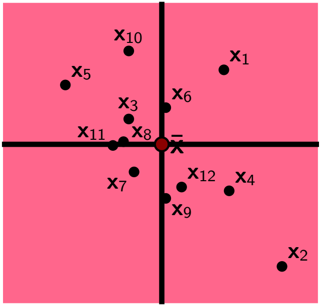
- Suppose we want to translate the origin to the data's center of mass...



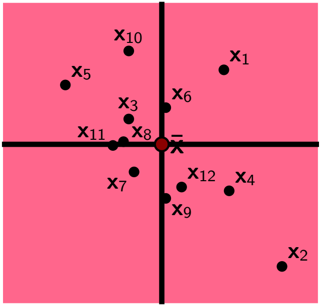
- Suppose we want to translate the origin to the data's center of mass...



- Suppose we want to translate the origin to the data's center of mass...



- Suppose we want to translate the origin to the data's center of mass...

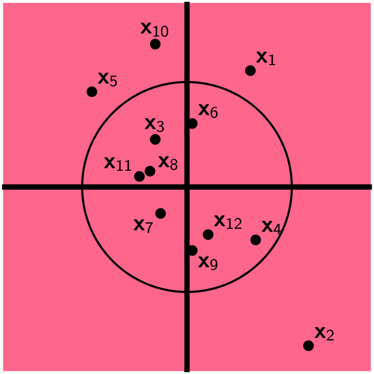


- As we will see next lecture, this transformation can be expressed as kernel transform

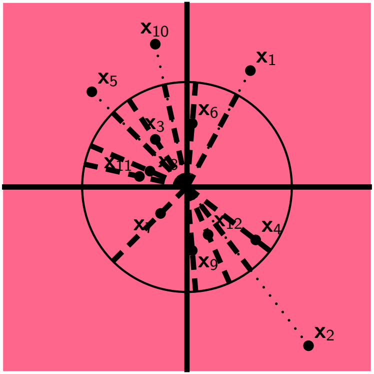
$$\hat{\mathbf{K}} \leftarrow \mathbf{K} - \frac{1}{N}\mathbf{1}\mathbf{1}^\top\mathbf{K} - \frac{1}{N}\mathbf{K}\mathbf{1}\mathbf{1}^\top + \frac{\mathbf{1}^\top\mathbf{K}\mathbf{1}}{N^2}\mathbf{1}\mathbf{1}^\top$$

Operations on Kernel Matrices
Normalizing Data

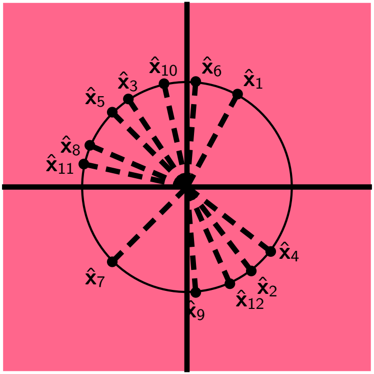
- Suppose we want to project all data to be norm 1; *i.e.*, $\|\hat{\mathbf{x}}\| = 1 \ldots$



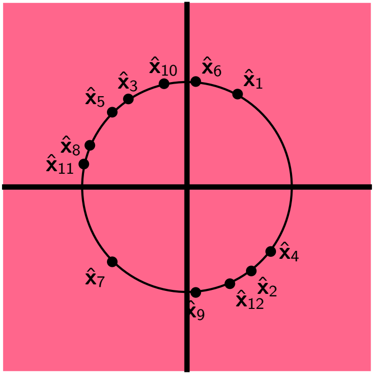
- Suppose we want to project all data to be norm 1; *i.e.*, $\|\hat{\mathbf{x}}\| = 1 \ldots$



- Suppose we want to project all data to be norm 1; *i.e.*, $\|\hat{\mathbf{x}}\| = 1\ldots$



- Suppose we want to project all data to be norm 1; *i.e.*, $\|\hat{\mathbf{x}}\| = 1\ldots$



- This transformation can be achieved using only the information from the kernel matrix:

$$\hat{\kappa}(\mathbf{x}, \mathbf{z}) = \frac{\kappa(\mathbf{x}, \mathbf{z})}{\sqrt{\kappa(\mathbf{x}, \mathbf{x})\,\kappa(\mathbf{z}, \mathbf{z})}}$$

Summary

- We explored a formal framework for kernels
- We saw a formal defintion for kernel functions & matrices
- We saw the properties that kernels must exhibit and how those properties can be used to validate kernel functions & construct new kernels from existing kernels
- We explored some operations that allow us to manipulate data in feature space
- *Next Lecture*: we will see basic kernel-based learning algorithms
  - We will explore how to take the mean of data in feature space & use that to construct a novelty detection algorithm
  - We will explore how to project data in feature space & use that for a basic subspace algorithm

Bibliography I

The Majority of the work from this talk can be found in the lecuture's accompanying book, "Kernel Methods for Pattern Analysis."

[1] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.