# Kernel Methods

## Motivating Example: Binary Classification: Support Vector Machines (Izenman Ch. 11)

**Setting:** Given data $\{(\vec{x}_i, y_i)\}_{i=1}^{N}$ $x_i \in \mathbb{R}^d$, $y_i \in \{-1, 1\}$

**Goal:** Learn a function $f : \mathbb{R}^d \to \mathbb{R}$ from the given data so that

$$C(\vec{x}) = \mathrm{sgn}(f(x)) = \begin{cases} +1 & f(x) \geq 0 \\ -1 & f(x) < 0 \end{cases}$$

is a classifier, i.e. given new $\vec{x}_{N+1}$ we predict $y_{N+1} = C(\vec{x}_{N+1})$.

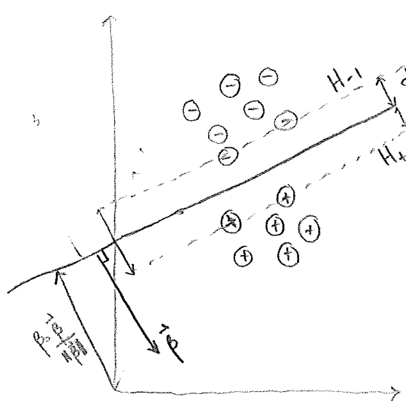
## Linearly Separable Case (Simplest Case)

Assume positive $(y_i = 1)$ and negative $(y_i = -1)$ points

can be perfectly separated by a hyperplane

$$\{\vec{x} : \beta_0 + \vec{\beta}^T \vec{x} = 0\}$$

called a separating hyperplane (decision boundary)




* The existence of a separating hyperplane implies existence of $\beta_0, \vec{\beta}$ s.t.

$$\beta_0 + \beta^T \vec{x}_i \geq +1 \quad (y_i = +1) \; ; \; \beta_0 + \beta^T \vec{x}_i \leq -1 \quad (y_i = -1)$$

* If $d_-$ $(d_+)$ is the shortest distance from hyperplane to nearest negative (positive) point

$$\text{margin} = d_- + d_+ = \frac{2}{\|\beta\|}$$

← maximum margin gives optimal separating hyperplane

Minimize $\frac{1}{2}\|\beta\|^2$

subject to $y_i(\beta_0 + \beta^T x_i) \geq 1 \quad i = 1, 2, \ldots, N$

## PRIMAL-DUAL OPTIMIZATION (KARUSH-KUHN-TUCKER CONDITIONS)

IT CAN BE SHOWN THAT THE OPTIMAL $\hat{\beta}_0, \hat{\beta}$ ARE GIVEN BY

$$\hat{\beta} = \sum_{i=1}^{N} \hat{\alpha}_i y_i \vec{x}_i$$

$$\hat{\beta}_0 = -\frac{1}{2}\left( \vec{x}_{-1}^T \hat{\beta} + x_{+1}^T \hat{\beta} \right) \qquad x_{-1} \text{ ON } H_{-1}, x_{+1} \text{ ON } H_{+1}$$

WHERE $\vec{\alpha} = (\alpha_1, \ldots \alpha_N)$ MAXIMIZES

$$1_N^T \vec{\alpha} - \frac{1}{2} \vec{\alpha}^T H \alpha$$

SUBJECT TO $\vec{\alpha}_1, \ldots \alpha_N \geq 0$, $\sum \alpha_i y_i = 0$, $H_{ij} = y_i y_j x_i^T x_j$,

AND THE OPTIMAL HYPERPLANE IS

$$f(\vec{x}) = \hat{\beta}_0 + \sum_{i=1}^{N} \hat{\alpha}_i y_i \left( \vec{x}_i^T \vec{x} \right)$$

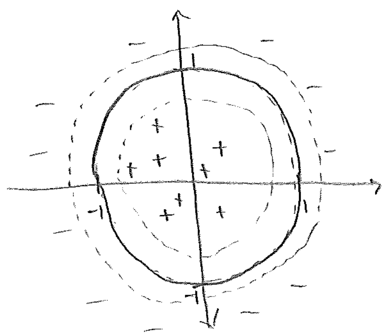
<u>NOTE</u>:
- TYPICALLY, $\vec{\alpha}$ IS SPARSE (MANY ZEROS) SO $\sum_{i=1}^{N} \ldots$ CAN BE REDUCED TO SUPPORT VECTORS FOR WHICH $\hat{\alpha}_i > 0$.
- CLASSIFICATION OF $\vec{x}_{N+1}$ REQUIRES INNER PRODUCTS $\vec{x}_i^T \vec{x}_N$

$$C(\vec{x}_{N+1}) = \operatorname{SGN}\left( f(\vec{x}_{N+1}) \right)$$

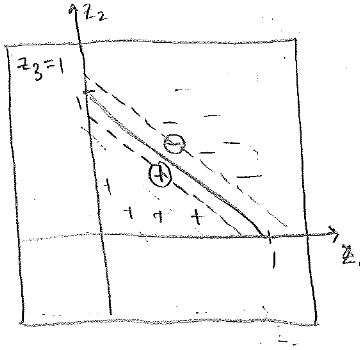
  SO WE CAN DERIVE A CLASSIFICATION RULE IF WE HAVE ONLY THE INNER PRODUCTS OF ANY PAIR OF POINTS BUT NOT THE POINTS THEMSELVES (SIMILAR TO INNER PRODUCT MATRIX ; CLASSICAL SCALING MDS)!
- LINEARLY SEPARABLE (HARD MARGIN) CASE CAN BE RELAXED TO SOFT-MARGIN (STRICT SEPARATION NOT REQUIRED) W/ CONNECTIONS TO LOGISTIC REGRESSION.

## Nonlinear SVM

Ex: Non-planar Decision Boundary




$$\varphi\left((x_1, x_2)\right) = \begin{pmatrix} x_1^2 \\ x_2^2 \\ 1 \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix}$$




* A curved decision boundary can be represented implicitly as a function (polynomial in some cases) of components of $\vec{x}$, i.e. $x_1^2 + x_2^2 - 1 = 0$ in case above

* Can find a nonlinear mapping from original space ($\mathbb{R}^2$ here) to a higher dimensional "feature" space (here $\mathbb{R}^3$) where data are linearly separable!

$$\varphi : \mathbb{R}^d \longrightarrow \mathcal{H} \quad (\text{w/ } \mathbb{R}^p)$$

$\uparrow$ where our data live $\qquad$ $\uparrow$ feature space

* Linear SVM in $\mathcal{H}$

  - Decision boundary is $f(\vec{x}) = \beta_0 + \beta^T \varphi(x)$

$$= \beta_0 + \sum_{i=1}^{N} \alpha_i y_i \underbrace{\varphi(x_i)^T \varphi(\vec{x})}$$

(requires an inner product on $\mathcal{H}$ !!)

  - Finding $\vec{\alpha}$ requires maximizing

$$1^T \vec{\alpha} - \vec{\alpha}^T H \vec{\alpha}, \quad H_{ij} = y_i y_j \overbrace{\varphi(x_i)^T \varphi(x_j)}$$

Ex: $2^{\text{ND}}$ DEGREE POLYNOMIAL IN $\mathbb{R}^d$

- GIVEN $\vec{x} = (x_1 \dots x_d)^T$ WE WANT TO FIND THE DECISION BOUNDARY OF THE FORM

$$f(\vec{x}) = \beta_0 + \beta^T \varphi(x) \quad , \quad \varphi(x) = \begin{pmatrix} x_1^2 \\ \vdots \\ x_d^2 \\ \sqrt{2}\, x_d x_{d-1} \\ \vdots \\ \sqrt{2}\, x_d x_1 \\ \sqrt{2}\, x_{d-1} x_{d-2} \\ \vdots \\ \sqrt{2}\, x_2 x_1 \\ \sqrt{2c}\, x_d \\ \vdots \\ \sqrt{2c}\, x_1 \\ c \end{pmatrix} \in \mathbb{R}^{\frac{(d+1)(d+2)}{2}}$$

- DIMENSION OF $\varphi(x)$ IS QUADRATIC IN $d$!
    - EVEN WORSE IF WE INCLUDE CUBIC, QUARTIC, ... TERMS
    - EXPONENTIALLY BAD AS WE INCREASE COMPLEXITY, FROM A COMPUTATIONAL PERSPECTIVE

- OBSERVE:

$$\begin{aligned}
\varphi(x)^T \varphi(y) &= \sum_{i=1}^{d} (x_i)^2 (y_i)^2 + \sum_{i=2}^{d} \sum_{j=1}^{i-1} 2 x_i x_j y_i y_j + \sum_{i=1}^{d} 2c\, x_i y_i + c \\
&= \sum_{i=1}^{d} (x_i y_i)^2 + 2 \sum_{i=1}^{d} \sum_{j=1}^{i-1} (x_i y_i)(x_j y_j) + 2c \sum_{i=1}^{d} x_i y_i + c^2 \\
&= \left(\vec{x}^T \vec{y} + c\right)^2 = k(\vec{x}, \vec{y})
\end{aligned}$$

* NO NEED TO COMPUTE $\varphi(x)$ OR $\varphi(y)$ IF WE ONLY NEED $\varphi(x)^T \varphi(y)$!

- COMPUTE INNER PRODUCTS IN FEATURE SPACE USING ONLY ORIGINAL VECTORS $\vec{x}, \vec{y}$!

"KERNEL TRICK" $\quad k(\vec{x}, \vec{y}) = \varphi(\vec{x})^T \varphi(\vec{y})$

* HIGHER ORDER POLYNOMIALS USING $k(\vec{x}, \vec{y}) = (\vec{x}^T \vec{y} + c)^d$

POLYNOMIAL KERNEL

- LINEAR METHODS REQUIRING INNER PRODUCTS CAN BE "KERNELIZED" TO GIVE A NONLINEAR METHOD W/O EVER COMPUTING $\varphi(x_i)$ !

<u>Ex:</u> RADIAL BASIS FUNCTION (RBF) KERNEL

$$k(\vec{x},\vec{y}) = e^{-\frac{\|\vec{x}-\vec{y}\|^2}{2\sigma^2}} = \exp\left\{ -\frac{1}{2\sigma^2}\left(\|\vec{x}\|^2 + \|\vec{y}\|^2\right) + \frac{\vec{x}^T\vec{y}}{\sigma^2} \right\}$$

$$= \sum_{j=0}^{\infty} \frac{(\vec{x}^T\vec{y}/\sigma^2)^j}{j!} e^{-\frac{\|x\|^2}{2\sigma^2}} e^{-\frac{\|y\|^2}{2\sigma^2}} = \underbrace{\sum_{j=0}^{\infty} \sum_{\sum n_i = j}} e^{-\frac{\|x\|^2}{2\sigma^2}} \frac{x_1^{n_1}\cdots x_d^{n_d}}{\sigma^j\sqrt{n_1!\cdots n_d!}} e^{-\frac{\|y\|^2}{2\sigma^2}} \frac{y_1^{n_1}\cdots y_d^{n_d}}{\sigma^j\sqrt{n_1!\cdots n_d!}}$$

- INFINITE DIMENSIONAL FEATURE SPACE !!
- GIVEN ANY LABELED DATA, USING THIS KERNEL GIVES A LINEARLY SEPARABLE SET IN FEATURE SPACE
  - BASICALLY UNLIMITED FLEXIBILITY (DANGER OF OVERFITTING)

TRANSLATION INVARIANT (DEPEND ONLY ON $\|\vec{x}-\vec{y}\|$)

<u>OTHER EXAMPLES</u>

- LAPLACIAN $\quad \exp\left(-\|x-y\|/\sigma\right)$
- THIN-PLATE SPLINE $\quad \left(\frac{\|x-y\|}{\sigma}\right)^2 \log\left(\frac{\|x-y\|}{\sigma}\right)$
- SIGMOID $\quad \tanh\left(a(x^Ty)+b\right)$ $\quad$ * ONLY THEORETICAL SUPPORT FOR CERTAIN VALUES OF $a, b$

<u>NOTE:</u>

- KERNELS CAN BE COMBINED TOGETHER TO CAPTURE MORE COMPLEX RELATIONSHIPS

$$k(x,y) = k_1(x,y) + k_2(x,y)$$
$$k(x,y) = \alpha\, k(x,y) \qquad \alpha > 0$$
$$k(x,y) = k_1(x,y)\, k_2(x,y)$$
$$k(x,y) = f(x) f(y) \qquad f: \mathbb{R}^d \to \mathbb{R}$$

- THE CHOICE OF KERNEL IMPLICITLY DEFINES AN INNER PRODUCT (HILBERT) SPACE OF FEATURES AND MAP $\varphi: \mathbb{R}^d \to \mathcal{H}$
  - BEST CHOICE OF KERNEL IS DEPENDENT ON THE PROBLEM
    - <u>BAD NEWS</u>: CONTENT KNOWLEDGE OR SHAPE OF SOLUTION IS USEFUL
    - <u>GOOD NEWS</u>: CAN EXTRACT SOME DETAILS (OPTIMAL PARAMETER TUNING) FOR A CHOSEN MODEL

## Theoretical Details:

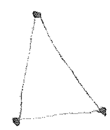
- **Motivation:** linear separability easier in higher dimensions

  **Thm. (Cover 1965)**

  "A complex pattern-classification problem, cast in a high dimensional space nonlinearly, is more likely to be linearly separable than in a low-dimensional space, provided that the space is not densely populated."

  **Proof:**

  $$x_i \longmapsto \vec{e}_i \in \mathbb{R}^N$$

  - map data to vertices of simplex
    - any partition into two sets is separable by a hyperplane.

- **Requirements & Properties of Kernels**

  - Can we choose any function $k: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ as a kernel? Equivalently, does a given $k$ have a corresponding Hilbert space $\mathcal{H}$ and map $\varphi: \mathbb{R}^d \to \mathcal{H}$ s.t. $k(\vec{x},\vec{y}) = \underbrace{\langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}}_{\text{inner product of } \mathcal{H} \text{ (usually the standard dot product)}}$ ?

  - **Mercer's Condition**

    There exists a Hilbert space and map $\varphi: \mathbb{R}^d \to \mathcal{H}$ s.t. $k(x,y) = \langle \varphi(x), \varphi(y) \rangle_{\mathcal{H}}$

    IFF

    For any $g(x)$ s.t. $\int_{\mathbb{R}^d} [g(\vec{x})]^2 \, d\vec{x} < \infty$, then $\underbrace{\int g(x)\, k(\vec{x},\vec{y})\, g(y)\, d\vec{x}\, d\vec{y} \geq 0}_{\text{Note similarity to P.S.D. matrices!}}$.

    * $k$ is semi-positive definite

    Equivalently, given any finite set of points $x_1, \dots x_N \in \mathbb{R}^d$ and $c_1, \dots c_N \in \mathbb{R}$

    $$\sum_{i=1}^{N} \sum_{j=1}^{N} c_i c_j \, k(\vec{x}_i, x_j) = \vec{c}^{\,T} \mathbb{K} \vec{c}$$

    $\mathbb{K}_{ij} = k(x_i, x_j)$ is called the Gram matrix

## Viewing Kernels as Operators on Function Spaces

- Let $\mathbb{L}^2(\mathbb{R}^d) = \left\{ g : \mathbb{R}^d \to \mathbb{R} \mid \int [g(x)]^2 \, dx < \infty \right\}$

- Associated w/ a kernel $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a linear operator

$$\left(T_K \; f\right)(x) = \int k(\vec{x}, \vec{y}) \, f(\vec{y}) \, dy \qquad \left(\text{assuming } \iint |k(x,y)|^2 \, dx dy < \infty\right)$$

$$T_K : \mathbb{L}^2(\mathbb{R}^d) \longrightarrow \mathbb{L}^2(\mathbb{R}^d)$$

- <u>Mercer's Thm.</u> (Spectral Thm.)

$k(x,y) = k(y,x)$

Suppose $k$ is a continuous, symmetric kernel satisfying Mercer's condition. Then there exists an orthonormal basis $\{\phi_j\}_{j=1}^{\infty} \subset \mathbb{L}^2(\mathbb{R}^d)$ of continuous eigenfunctions of $T_K$ w/ a corresponding sequence of nonnegative eigenvalues $\{\lambda_j\}_{j=1}^{\infty}$, and

$$k(\vec{x}, \vec{y}) = \sum_{j=1}^{\infty} \lambda_j \, \phi_j(\vec{x}) \, \phi_j(\vec{y}) \qquad (\text{Diagonalization})$$

- <u>Reproducing Property</u>

  - A kernel $k$ (with corresponding Hilbert space $\mathcal{H}$) has the reproducing property if for any $f \in \mathcal{H}$

$$\langle f(\cdot), k(x, \cdot) \rangle_{\mathcal{H}} = \langle f, \phi(x) \rangle_{\mathcal{H}} = f(x)$$

  and we say $k$ is a reproducing kernel.

  - For any positive semidefinite (Mercer) kernel, we can construct a <u>unique</u> Hilbert space $\mathcal{H}_K$ for which $k$ is its reproducing kernel. $\mathcal{H}_K$ is called a Reproducing Kernel Hilbert Space (RKHS)

    - RKHSs have many applications in statistics (kernel PCA for example)

    - Moore-Aronszajn Theorem proves this claim in a constructive manner.