CSCE 990 Lecture 2: Kernels*

Stephen D. Scott

January 12, 2006

Outline

- Dot products as similarity measures
- Example: Product features
- Definitions
- All kernels are dot products
  - The “kernel trick”
- Examples of kernels
- Sections 1.1, 2.1, 2.2.1–2.2.2, 2.2.6–2.2.7, 2.3 (also read Sections 2.2.3–2.2.4, 2.5)

Introduction

- Remember that a kernel is simply a dot product under some mapping
  - We'll go into this more formally later

- Dot product $\Rightarrow$ similarity measure
  - E.g.: $\mathbf{x}_1 = (1/\sqrt{2}, 1/\sqrt{2})$, $\mathbf{x}_2 = (1/1.3, 1/1.565)$, $\mathbf{x}_3 = (1, 0)$ ($\|\mathbf{x}_i\| = 1 \ \forall i$)

$$\langle \mathbf{x}_1, \mathbf{x}_2 \rangle \approx 1/1.838 + 1/2.213 \approx 0.9958$$

$$\langle \mathbf{x}_1, \mathbf{x}_3 \rangle = 1/\sqrt{2} + 0 \approx 0.707$$

  - If $\|\mathbf{x}\| = 1$ and $\|\mathbf{x}'\| = 1$, then $\langle \mathbf{x}, \mathbf{x}' \rangle =$ cosine of angle between them

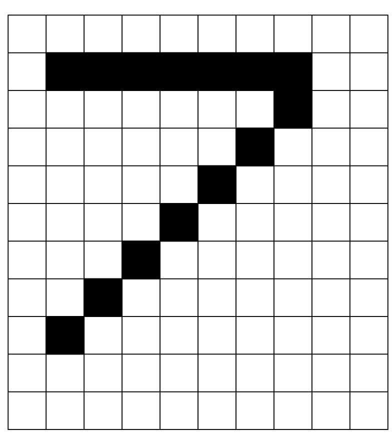
- So kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ gives measures of similarity under its corresponding remapping $\Phi : \mathcal{X} \to \mathcal{H}$
  - $\mathcal{X}$ is the original input space, where the labeled training examples x_i come from
  - $\mathcal{H}$ is the feature space, which is where we'll search for a separating hyperplane

Product Features

- Let $\mathcal{X} \subseteq \mathbb{R}^N$. We will consider the dth order products of the entries $[x]_j$ of $x \in \mathcal{X}$:
  $[x]_{j_1} \cdot [x]_{j_2} \cdots [x]_{j_d}$ for $j_1, \ldots, j_d \in \{1, \ldots, N\}$

- These are called product features, and $\mathcal{H}$ is the set of all products of d entries

- Popular in image processing:
  - Let each x correspond to a vector of the pixel intensities of an entire image (smoothed to remove noise)
  - Each product feature in $\Phi(x)$ is related to a logical "and" of a subset (up to size d) of pixels from the image x



Product Features
(cont'd)

- E.g. $\Phi(([x]_1, [x]_2)) = ([x]_1^2, [x]_2^2, [x]_1[x]_2)$

- Problem: If x has N dimensions, then for order-d products, the dimension of $\Phi(x)$ is

$$N_{\mathcal{H}} = \binom{d + N - 1}{d} \geq \left(\frac{d + N - 1}{d}\right)^d$$

  - E.g. images that are 16×16 ($N = 256$) and $d = 5$ yield $N_{\mathcal{H}} \approx 10^{10}$

- But if we're only concerned about the dot products, then we can define $\Phi_d(x)$ such that

$$\left\langle \Phi_d(x), \Phi_d(x') \right\rangle = \left\langle x, x' \right\rangle^d = k(x, x') \ ,$$

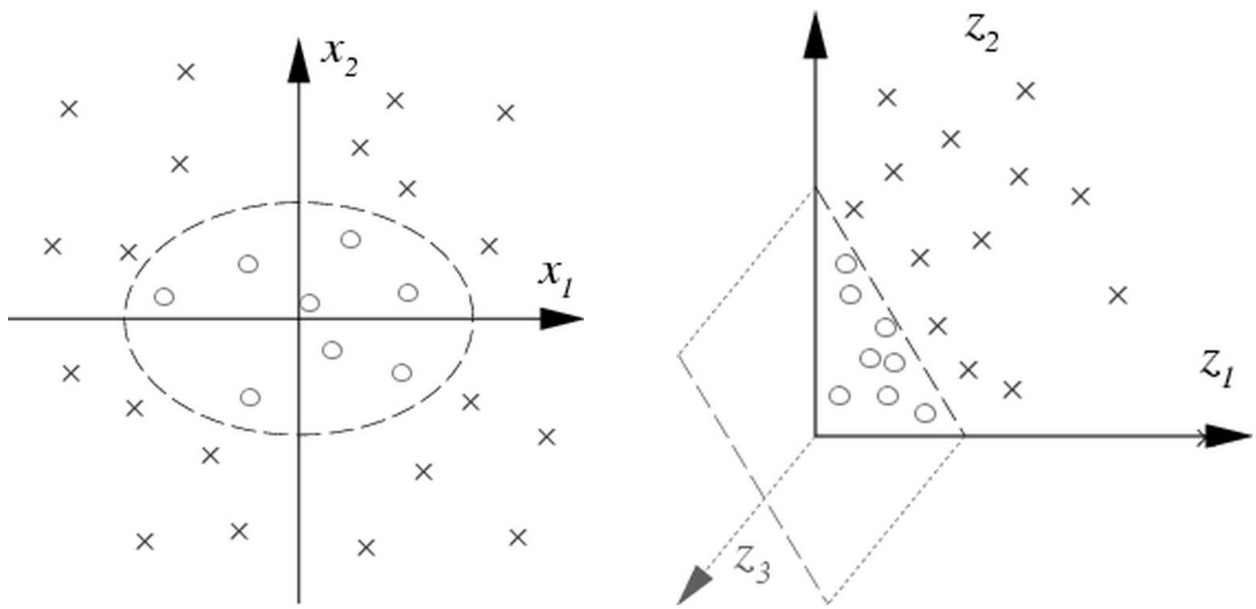
  which is easy to compute

  - E.g. $\Phi_2(x) = \left([x]_1^2, [x]_2^2, \sqrt{2}[x]_1[x]_2\right)$

- Can also use $k(x, x') = (\langle x, x' \rangle + 1)^d$ to get terms of degree $\leq d$

Product Features

An Example



Definitions

- Up to now, we assumed $\mathcal{X} \subseteq \mathbb{R}^N$. For the rest of this course, $\mathcal{X}$ can be arbitrary, e.g. sequences of letters from some alphabet (such as protein sequences)

- We will require the range of kernels to be $\mathbb{R}$, even though the book allows it to be complex

D2.3 Given a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and patterns $x_1, \ldots, x_m \in \mathcal{X}$, the $m \times m$ matrix K (where $K_{ij} = k(x_i, x_j)$) is called the Gram matrix or kernel matrix of k wrt $x_1, \ldots, x_m$

D2.4 A real, symmetric $m \times m$ matrix K that satisfies $\langle \mathbf{x}, K\mathbf{x} \rangle \geq 0$ for all $\mathbf{x} \in \mathcal{H}$ is positive definite. Equivalently, K is PD if it is symmetric and satisfies $\sum_{i,j} c_i c_j K_{ij} \geq 0$ for all $c_i, c_j \in \mathbb{R}$

- PD $\Leftrightarrow$ all eigenvalues ≥ 0

Definitions
(cont'd)

D2.5 A function k on $\mathcal{X} \times \mathcal{X}$ which for all positive integers m and all $x_1, \ldots, x_m \in \mathcal{X}$ yields a PD Gram matrix is called a positive definite kernel, aka kernel, reproducing kernel, Mercer kernel, admissible kernel, support vector kernel, non-negative definite kernel, positive semidefinite kernel, covariance function

- Properties of PD kernels:

  1. If Φ maps $\mathcal{X}$ to $\mathcal{H}$, then $\langle \Phi(x), \Phi(x') \rangle$ is a PD kernel on $\mathcal{X} \times \mathcal{X}$

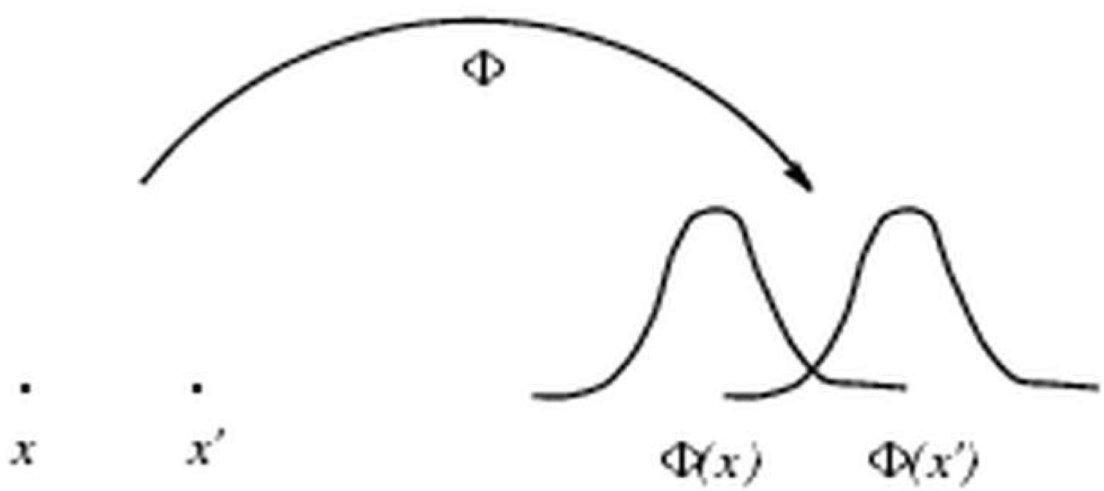
  2. $k(x,x) \geq 0$ for all $x \in \mathcal{X}$

  3. Cauchy-Schwarz: $k(x,x')^2 \leq k(x,x)k(x',x')$

  4. $k(x,x) = 0$ for all $x \in \mathcal{X}$ implies $k(x,x') = 0$ for all $x, x' \in \mathcal{X}$

All Kernels are Dot Products

- All kernels are dot products in some feature space $\mathcal{H}$
- Consider a kernel k and some $x \in \mathcal{X}$
- Then $\Phi(x)(\cdot) = k(\cdot, x)$ is a function that measures similarity of all $x' \in \mathcal{X}$ to x
  - I.e. $\Phi(x)(x') = k(x', x)$
  - One such function for each $x \in \mathcal{X}$



- Can now think of each $x \in \mathcal{X}$ as a function over $\mathcal{X}$

All Kernels are Dot Products
(cont'd)

- We can turn the set of functions $\Phi(\mathcal{X})$ into a linear space

- Let m, m' be positive ints, $\alpha_i, \beta_j \in \mathbb{R}$, and $x_1, \ldots, x_m, x'_1, \ldots, x'_{m'} \in \mathcal{X}$ be arbitrary

- Let

  $$f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \qquad g(\cdot) = \sum_{j=1}^{m'} \beta_j k(\cdot, x'_j)$$

  and define the dot product as

  $$\langle f, g \rangle := \sum_{i=1}^{m} \sum_{j=1}^{m'} \alpha_i \beta_j k(x_i, x'_j) \tag{1}$$

- Can show that (1) is a valid dot product and that

  $$\Big\langle k(\cdot, x), k(\cdot, x') \Big\rangle = k(x, x') \ ,$$

  which implies

  $$\Big\langle \Phi(x), \Phi(x') \Big\rangle = k(x, x')$$

The "Kernel Trick"

- Thus we see that any algorithm formulated in terms of a PD kernel k can be changed by replacing k with another PD kernel k'

- Holds for <u>any</u> algorithm, not just SVMs

Examples of Kernels

- Polynomial: $k(x, x') = (\langle x, x' \rangle + c)^d$ for $c \geq 0$
  - When $c = 0$, then k is invariant under all rotations and mirroring operations of $\mathcal{X}$

- Gaussian radial basis function (Gaussian RBF):

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

  with $\sigma > 0$

  - Invariant under rotations and translations
  - Its remapping has $\|\Phi(x)\| = 1$ for all $x \in \mathcal{X}$

- Sigmoid: $k(x, x') = \tanh(\kappa \langle x, x' \rangle + \vartheta)$ with $\kappa > 0$ and $\vartheta < 0$
  - Invariant under rotations
  - Not PD, but still used in practice

Examples of Kernels
(cont'd)

- Can make new kernels from other kernels: if k_1 and k_2 are PD kernels, then so are

  $\Rightarrow$ αk_1 for all $\alpha \geq 0$

  $\Rightarrow$ $k_1 + k_2$

  $\Rightarrow$ $k_1 \, k_2$

  $\Rightarrow$ $k(A, B) := \sum_{x \in A, x' \in B} k_1(x, x')$, where $A, B \subseteq \mathcal{X}$

  – More on this later

Empirical Kernel Map

- Given a kernel k and a data set $Z = \{z_1, \ldots, z_n\}$, can define an empirical kernel map $\Phi_m(x) = (k(z_1, x), \ldots, k(z_n, x))^\top$

- I.e. remap x to a new representation based on its similarities to the patterns in Z

- Can then use each $\Phi_m(x)$ as training patterns in an SVM, etc.
  - Can feed pairs of $\Phi_m(x)$ into a different kernel k'
  - If k' is a straight dot product, then this is the same as squaring K, k's Gram matrix

- This remapping is valid even if k is not PD!

Topic summary due in 1 week!