Kernel Properties - Convexity

Leila Wehbe

October 1st 2013

Kernel Properties

- data is not linearly separable $\rightarrow$ use feature vector of the data $\Phi(x)$ in another space
- we can even use infinite feature vectors
- because of the Kernel trick you will not have to explicitly compute the feature vectors $\Phi(x)$. (you will Kernelize an algorithms in HW2).

Kernels

- dot product in feature space $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$
- we can write the kernel in matrix form over the data sample: $K_{ij} = \langle \Phi(x), \Phi(x') \rangle = k(x, x')$. This is called a Gram matrix.
- K is positive semi-definite, i.e. $\alpha K \alpha \geq 0$ for all $\alpha \in \mathbb{R}^m$ and all kernel matrices $K \in \mathbb{R}^{m \times m}$. Proof (from class):

$$\sum_{i,j}^{m} \alpha_i \alpha_j K_{ij} = \sum_{i,j}^{m} \alpha_i \alpha_j \langle \Phi(x_i), \Phi(x_j) \rangle$$

$$= \langle \sum_{i}^{m} \alpha_i \Phi(x_i), \sum_{j}^{m} \alpha_j \Phi(x_j) \rangle = || \sum_{i}^{m} \alpha_i \Phi(x_i) ||^2 \geq 0$$

Kernels

- by mercer's theorem, any symmetric, square integrable function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that satisfies

$$\int_{\mathcal{X}\times\mathcal{X}} k(x,x')f(x)f(x')dxdx' \geq 0$$

  there exist a feature space $\Phi(x)$ and a $\lambda \geq 0$
  $k(x,x') = \sum_i \lambda_i \phi_i(x)\phi_i(x')$ (we have $k(x,x') = \langle \Phi'(x), \Phi'(x')\rangle$)
- in discrete space: $\sum_i \sum_j K(x_i,x_j)c_i c_j$

any Gram matrix derived of a kernel k is positive semi definite $\leftrightarrow$ k is a valid kernel (dot product)

Exercices

$k(x, x')$ is a valid kernel

- show that $f(x)f(x')k(x, x')$ is a kernel

Exercices

Answer:

$$f(x)f(y)k(x,y) = f(x)f(y) < \phi(x), \phi(y) > = < f(x)\phi(x), f(y)\phi(y) >$$
$$= < \phi'(x), \phi'(y) >$$

Exercices

$k_1(x, x'), k_2(x, x')$ are valid kernels

- show that $c_1 * k_1(x, x') + c_2 * k_2(x, x')$, where $c_1, c_2 \geq 0$ is a valid Kernel (multiple ways to show it)

Exercices

Answer 1:
For any function $f(.)$:

$$\int_{x,x'} f(x)f(x')[c_1 k_1(x,x') + c_2 k_2(x,x')]\,dx\,dx'$$
$$= c_1 \int_{x,x'} f(x)f(x')k_1(x,x')\,dx\,dx' + c_2 \int_{x,x'} f(x)f(x')k_2(x,x')\,dx\,dx' \;\geq\; 0$$

since $\int_{x,x'} f(x)f(x')k_1(x,x')\,dx\,dx' \geq 0$ and $\int_{x,x'} f(x)f(x')k_2(x,x')\,dx\,dx' \geq 0$ since k_1 and k_2 are valid kernels.

Exercices

Answer 2:
Here is another way to prove it:

- Given any final set of instances $\{x_1, \ldots, x_n\}$, let K_1 (resp., K_2) be the $n \times n$ Gram matrix associated with k_1 (resp., k_2). The Gram matrix associated with $c_1k_1 + c_2k_2$ is just $K = c_1K_1 + c_2K_2$.
- K is PSD because any $v \in \mathbb{R}^n$, $v^T(c_1K_1 + c_2K_2)v \;=\; c_1(v^TK_1v) \;+\; c_2(v^TK_2v) \geq 0$ as $v^TK_1v \geq 0$ and $v^TK_2v \geq 0$ follows from K_1 and K_2 being positive semi definite.
- k is a valid kernel.

Exercices

Answer 3:
let Φ^1 and Φ^2 be the feature vectors associated with k_1 and k_2 respectively.
Take vector Φ which is the concatenation of $\sqrt{c_1}\Phi^1$ and $\sqrt{c_2}\Phi^2$.
i.e. $\Phi(x) =$
$[\sqrt{c_1}\phi_1^1(x), \sqrt{c_1}\phi_2^1(x), ...\sqrt{c_1}\phi_m^1(x), \sqrt{c_2}\phi_1^2(x), \sqrt{c_2}\phi_2^2(x), ...\sqrt{c_2}\phi_m^2(x)]$.
It's easy to check that

$$\begin{aligned}\langle\Phi(x), \Phi(x')\rangle = \sum_{i=1}^{N} \phi_i(x) \times \phi_i(x') &= c_1 \sum_{i=1}^{m} \phi_i^1(x) \times \phi_i^1(x') \\ &= c_1\langle\Phi^1(x), \Phi^1(x')\rangle + c_2\langle\Phi^2(x), \Phi^2(x')\rangle \\ &= c_1 k_1(x, x') + c_2 k_2(x, x') = k(x, x')\end{aligned}$$

therefore k is a valid kernel.

Exercices

k_1, k_2 are valid kernels

- show that $k_1(x,x') - k_2(x,x')$ is not necessarily a kernel

Exercices

Proof by counter example:
Consider the kernel k_1 being the identity ($k_1(x, x') = 1$ iff $x = x'$ and $= 0$ otherwise), and k_2 being twice the identity ($k_1(x, x') = 2$ iff $x = x'$ and $= 0$ otherwise).
Let $K_1 = I_p$ be the $p \times p$ identity matrix and $K_p = 2I_p$ be 2 times that identity matrix. K_1 and K_2 are the Gram matrices associated with k_1 and k_2 respectively. Clearly both K_1 and K_2 are positive semi definite, however $K_1 - K_2 = -I$ is not, as its eigenvalues are -1.
Therefore k is not a valid kernel.

Exercices

PSD matrices A and B

- show that AB is not necessarily PSD

Exercices

for PSD matrices A and B, it suffices to show that AB is not symmetric – so just use $A = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}$ and $B = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$; here $AB = \begin{pmatrix} 2 & 1 \\ 2 & 4 \end{pmatrix}$ which is not symmetric.

Exercices

k_1, k_2 are valid kernels

- show that the element wise product $k(x_i, x_j) = k_1(x_i, x_j) \times k_2(x_i, x_j)$ is a valid kernel.
- start by showing that if matrices A and B are PSD, then $C_{ij} = A_{ij} \times B_{ij}$ is PSD

Exercices

Answer: First show that C s.t. $C_{ij} = A_{ij} \times B_{ij}$ is PSD:
One way to show it:

1. Any PSD matrix Q is a covariance matrix.
   To see this, think of a p-dimensional random variable $\mathbf{x}$ with a covariance matrix $\mathbf{I}_p$, the identity matrix. (Q is $p \times p$)
   Because Q is PSD it admits a non-negative symmetric square root $Q^{\frac{1}{2}}$.
   Then:
   $$cov(Q^{\frac{1}{2}}\mathbf{x}) = Q^{\frac{1}{2}}cov(\mathbf{x}))Q^{\frac{1}{2}} = Q^{\frac{1}{2}}\mathbf{I}Q^{\frac{1}{2}} = Q$$
   And therefore Q is a covariance matrix.
2. We also know that any covariance matrix is PSD. So given A and B PSD, we know that they are covariance matrices. We want to show that C is also a covariance matrix and therefore PSD.

Exercices

3. Let $u = (u_1, \dots, u_n)^T \sim N(0_p, A)$ and $v = (v_1, \dots, v_n)^T \sim N(0_p, B)$ where $0 + p$ is a p-dimensional vector of zeros
   Define the vector $w = (u_1 v_1, \dots, u_n v_n)^T$
4.

$$cov(w) = E[(w - \mu^w)(w - \mu^w)^T] = E[ww^T]$$

This is because $\mu_i^w = 0$ for all i. This is because u and v are independent so $\mu^w = \mu^u \times \mu^v = 0_p$

$$\begin{aligned} cov(w)_{i,j} = E[w_i w_j^T] = E[(u_i v_i)(u_j v_j)] = E[(u_i u_j)(v_i v_j)] \\ = E[u_i u_j]\, E[v_i v_j] \end{aligned}$$

This is again because u and v are independent.

$$cov(w)_{i,j} = E[u_i u_j]\, E[v_i v_j] = A_{i,j} \times B_{i,j} = C_{i,j}$$

Exercices

5. Therefore C is a covariance matrix and therefore PSD
6. Since any kernel matrix created from $k(x_i, x_j) = k_1(x_i, x_j) \times k_2(x_i, x_j)$ is PSD, then k is PSD.

Exercices

A is PSD

- show that A^m is PSD

Exercices

Answer:
Recall $A = UDU^T$
First we show that $A^m = UD^mU^T$.
Proof by induction:

- trivially true for $m = 1$.
- $A^{m+1} = AA^m = UDU^T(UD^mU^T) = UD(U^TU)D^mU^T = UDD^mU^T = UD^{m+1}U^T$

Hence, the eigenvalues of A^m are the diagonal elements of D^m, which are λ_i^m (where $\{\lambda_i\}$ are the diagonal elements of D). Since $\lambda_i \geq 0$, these eigenvalues λ_i^m are also ≥ 0. This means A^m is PSD.

Exercices

$k(x, x')$ is a valid kernel

- show that $k(x, y)^2 \leq k(x, x)k(y, y)$

Exercices

Answer:

$$k(x,y)^2 =< \phi(x), \phi(y) >^2 = ||\phi(x)||^2 ||\phi(y)||^2 (cos(\theta_{\phi(x),\phi(y)}))^2$$
$$\leq ||\phi(x)||^2 ||\phi(y)||^2 = k(x,x)k(y,y)$$

Introduction to Convex Optimization

Xuezhi Wang

Computer Science Department
Carnegie Mellon University

10701-recitation, Jan 29

Outline

Outline

Convex Sets

- Definition
  For $x, x' \in X$ it follows that $\lambda x + (1-\lambda)x' \in X$ for $\lambda \in [0,1]$
- Examples
  - Empty set $\emptyset$, single point $\{x_0\}$, the whole space $\mathbb{R}^n$
  - Hyperplane: $\{x \mid a^\top x = b\}$, halfspaces $\{x \mid a^\top x \le b\}$
  - Euclidean balls: $\{x \mid ||x - x_c||_2 \le r\}$
  - Positive semidefinite matrices: $\mathbf{S}^n_+ = \{A \in \mathbf{S}^n | A \succeq 0\}$ ($\mathbf{S}^n$ is the set of symmetric $n \times n$ matrices)

Convexity Preserving Set Operations

Convex Set C, D

- Translation $\{x + b \mid x \in C\}$
- Scaling $\{\lambda x \mid x \in C\}$
- Affine function $\{Ax + b \mid x \in C\}$
- Intersection $C \cap D$
- Set sum $C + D = \{x + y \mid x \in C, y \in D\}$

Outline

1. Convexity
   - Convex Sets
   - Convex Functions
2. Unconstrained Convex Optimization
   - First-order Methods
   - Newton's Method

Convex Functions

dom f is convex, $\lambda \in [0, 1]$
$\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y)$

- **First-order condition**: if f is differentiable,

$$f(y) \geq f(x) + \nabla f(x)^\top (y - x)$$

- **Second-order condition**: if f is twice differentiable,

$$\nabla^2 f(x) \succeq 0$$

- **Strictly convex**: $\nabla^2 f(x) \succ 0$
  Strongly convex: $\nabla^2 f(x) \succeq dI$ with $d > 0$

Convex Functions

A quick matrix calculus reference: http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html

Convex Functions

- **Below-set of a convex function** is convex:
  $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y)$
  hence $\lambda x + (1-\lambda)y \in X$ for $x, y \in X$
- **Convex functions don't have local minima**:
  Proof by contradiction:
  linear interpolation breaks local minimum condition
- **Convex Hull**:
  $Conv(X) = \{\bar{x} \mid \bar{x} = \sum \alpha_i x_i \textit{ where } \alpha_i \geq 0 \textit{ and } \sum \alpha_i = 1\}$
  Convex hull of a set is always a convex set

Convex Functions examples

- Exponential. e^{ax} convex on $\mathbb{R}$, any $a \in \mathbb{R}$
- Powers. x^a convex on $\mathbb{R}_{++}$ when $a \geq 1$ or $a \leq 0$, and concave for $0 \leq a \leq 1$.
- Powers of absolute value. $|x|^p$ for $p \geq 1$, convex on $\mathbb{R}$.
- Logarithm. $\log x$ concave on $\mathbb{R}_{++}$.
- Norms. Every norm on $\mathbb{R}^n$ is convex.
- $f(x) = \max\{x_1, ..., x_n\}$ convex on $\mathbb{R}^n$
- Log-sum-exp. $f(x) = \log(e^{x_1} + ... + e^{x_n})$ convex on $\mathbb{R}^n$.

Convexity Preserving Function Operations

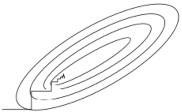
Convex function $f(x), g(x)$

- Nonnegative weighted sum: $af(x) + bg(x)$
- Pointwise Maximum: $f(x) = \max\{f_1(x), ..., f_m(x)\}$
- Composition with affine function: $f(Ax + b)$
- Composition with nondecreasing convex g: $g(f(x))$

Outline

1. Convexity
   - Convex Sets
   - Convex Functions
2. Unconstrained Convex Optimization
   - First-order Methods
   - Newton's Method

Gradient Descent



given a starting point $x \in \mathbf{dom} f$.
repeat
1. $\Delta x := -\nabla f(x)$
2. Choose step size t via exact or backtracking line search.
3. update. $x := x + t\Delta x$.

Until stopping criterion is satisfied.

- Key idea
  - Gradient points into descent direction
  - Locally gradient is good approximation of objective function
- Gradient Descent with line search
  - Get descent direction
  - Unconstrained line search
  - Exponential convergence for strongly convex objective

Outline

1. Convexity
   - Convex Sets
   - Convex Functions
2. Unconstrained Convex Optimization
   - First-order Methods
   - Newton's Method

Newton's method

- Convex objective function f
- Nonnegative second derivative

$$\partial_x^2 f(x) \succeq 0$$

- Taylor expansion

$$f(x+\delta) = f(x) + \delta^\top \partial_x f(x) + \frac{1}{2}\delta^\top \partial_x^2 f(x)\delta + O(\delta^3)$$

- Minimize approximation & iterate til converged

$$x \leftarrow x - [\partial_x^2 f(x)]^{-1} \partial_x f(x)$$