Kernel Methods Continued

Xintian Han, David S. Rosenberg

NYU CDS

February 27, 2019

Contents

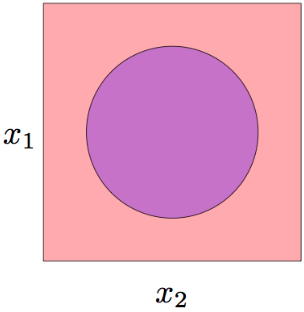
Recap

Linear Models with Explicit Feature Map

- Input space: $\mathcal{X}$ (no assumptions)
- Introduce **feature map** $\psi : \mathcal{X} \to \mathbf{R}^d$
- The feature map maps into the **feature space** $\mathbf{R}^d$.
- Hypothesis space of affine functions on feature space:

$$\mathcal{F} = \left\{ x \mapsto w^T \psi(x) + b \mid w \in \mathbf{R}^d, b \in \mathbf{R} \right\}.$$

Geometric Example: Two class problem, nonlinear boundary



- With identity feature map $\psi(x) = (x_1, x_2)$ and linear models, can't separate regions
- With appropriate featurization $\psi(x) = \left(x_1, x_2, x_1^2 + x_2^2\right)$, becomes linearly separable .
- Video: http://youtu.be/3liCbRZPrZA

From Percy Liang's "Lecture 3" slides from Stanford's CS221, Autumn 2014.

The Kernel Function

- **Input space**: $\mathcal{X}$
- **Feature space**: $\mathcal{H}$ (a Hilbert space, i.e. an inner product space with projections, e.g. $\mathbf{R}^d$)
- **Feature map**: $\psi : \mathcal{X} \to \mathcal{H}$
- The **kernel function** corresponding to ψ is

$$k(x, x') = \langle \psi(x), \psi(x') \rangle,$$

where $\langle \cdot, \cdot \rangle$ is the inner product associated with $\mathcal{H}$.

The Kernel Function: Why do we need this?

- **Feature map**: $\psi : \mathcal{X} \to \mathcal{H}$
- The **kernel function** corresponding to ψ is

$$k(x, x') = \left\langle \psi(x), \psi(x') \right\rangle.$$

- Why introduce this new notation $k(x, x')$?
- We can often evaluate $k(x, x')$ without explicitly computing $\psi(x)$ and $\psi(x')$.
- For large feature spaces, can be much faster.

What are the Benefits of Kernelization?

1. Computational (when optimizing over $\mathbf{R}^n$ is better than over $\mathbf{R}^d$)).
2. Can sometimes avoid any $O(d)$ operations
   - allows access to **infinite-dimensional feature spaces**.
3. Allows thinking in terms of "similarity" rather than features.

The Representer Theorem to Kernelize

The Representer Theorem

Theorem (Representer Theorem)

Let

$$J(w) = R(\|w\|) + L(\langle w, x_1\rangle, \ldots, \langle w, x_n\rangle),$$

where

- $w, x_1, \ldots, x_n \in \mathcal{H}$ *for some Hilbert space* $\mathcal{H}$. *(We typically have* $\mathcal{H} = \mathbf{R}^d$.*)*
- $\|\cdot\|$ *is the norm corresponding to the inner product of* $\mathcal{H}$. *(i.e.* $\|w\| = \sqrt{\langle w, w\rangle}$*)*
- $R : [0, \infty) \to \mathbf{R}$ *is nondecreasing (**Regularization term**), and*
- $L : \mathbf{R}^n \to \mathbf{R}$ *is arbitrary (**Loss term**).*

If $J(w)$ *has a minimizer, then it **has a minimizer of the form*** $w^* = \sum_{i=1}^n \alpha_i x_i$.
[If R *is strictly increasing, then all minimizers have this form. (Proof in homework.)]*

Questions on Representer Theorem

- If $J(w)$ is the objective function of the following problems, do all the minimizers have the form $w^* = \sum_{i=1}^{n} \alpha_i x_i$?
  - Lasso regression?
  - Ridge regression?

Questions on Representer Theorem

- If $J(w)$ is the objective function of the following problems, do all the minimizers have the form $w^* = \sum_{i=1}^n \alpha_i x_i$?
  - Lasso regression? **Not Always**
  - Ridge regression? **All the minimizers have the form.**
- (Copy from Representer Theorem)
  - $R : [0, \infty) \to \mathbf{R}$ is nondecreasing of $\|w\|$. If $J(w)$ has a minimizer, then it **has a minimizer of the form** $w^* = \sum_{i=1}^n \alpha_i x_i$.
  - If R is strictly increasing, then all minimizers have this form.

A Simple Example

- Suppose we only have one data point $x_1 = 1, x_2 = 1, y = 1$.
- Lasso regression: $J(w) = (y - w_1 x_1 - w_2 x_2)^2 + |w_1| + |w_2|$.
- Lasso regression is equivalent to (Homework 4):

$$\begin{aligned} \min_{w} \quad & J(w) = (y - w_1 x_1 - w_2 x_2)^2 \\ s.t. \quad & |w_1| + |w_2| \leqslant r \end{aligned}$$

- There is no closed form solution of r. But we can still analyze using r. All solutions (w_1, w_2) are on the line segment $w_1 + w_2 = r, \quad 0 \leqslant w_1, w_2 \leqslant r$. Only the one $(w_1 = r/2, w_2 = r/2)$ is a linear combination of (x_1, x_2).
- For ridge regression: $J(w) = (y - w_1 x_1 - w_2 x_2)^2 + w_1^2 + w_2^2$
- Solution is $(w_1 = 1/3, w_2 = 1/3)$, which is a linear combination of (x_1, x_2).

Representer Theorem (Baby Version)

Theorem ((Baby) Representer Theorem)

Suppose you have a loss function of the form

$$J(w) = L(w^T\phi(x_1), \dots, w^T\phi(x_n)) + R(\|w\|_2)$$

where

- $w, \phi(x_i) \in \mathbf{R}^D$.
- $L : \mathbf{R}^n \to \mathbf{R}$ *is an arbitrary function (loss term).*
- $R : \mathbf{R}_{\geqslant 0} \to \mathbf{R}$ *is increasing (regularization term).*

Assume J has at least one minimizer. Then J has a minimizer w^* of the form $w^* = \sum_{i=1}^n \alpha_i \phi(x_i)$ for some $\alpha \in \mathbf{R}^n$. If R is strictly increasing, then all minimizers have this form.

Kernels

Linear Kernel

- Input space: $\mathcal{X} = \mathbf{R}^d$
- Feature space: $\mathcal{H} = \mathbf{R}^d$, with standard inner product
- Feature map

$$\psi(x) = x$$

- Kernel:

$$k(x, x') = x^T x'$$

Quadratic Kernel in $\mathbf{R}^d$

- Input space $\mathcal{X} = \mathbf{R}^d$
- Feature space: $\mathcal{H} = \mathbf{R}^D$, where $D = d + \binom{d}{2} \approx d^2/2$.
- Feature map:

$$\psi(x) = (x_1, \ldots, x_d, x_1^2, \ldots, x_d^2, \sqrt{2}x_1x_2, \ldots, \sqrt{2}x_ix_j, \ldots \sqrt{2}x_{d-1}x_d)^T$$

- Then for $\forall x, x' \in \mathbf{R}^d$

$$\begin{aligned} k(x,x') &= \langle \psi(x), \psi(x') \rangle \\ &= \langle x, x' \rangle + \langle x, x' \rangle^2 \end{aligned}$$

- Computation for inner product with explicit mapping: $O(d^2)$
- Computation for implicit kernel calculation: $O(d)$.

Based on Guillaume Obozinski's Statistical Machine Learning course at Louvain, Feb 2014.

Polynomial Kernel in $\mathbf{R}^d$

- Input space $\mathcal{X} = \mathbf{R}^d$
- Kernel function:

$$k(x, x') = \left(1 + \langle x, x' \rangle\right)^M$$

- Corresponds to a feature map with all monomials up to degree M.
- For any M, computing the kernel has same computational cost
- Cost of explicit inner product computation grows rapidly in M.

The RBF Kernel

Radial Basis Function (RBF) / Gaussian Kernel

- Input space $\mathcal{X} = \mathbf{R}^d$

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

  where σ^2 is known as the bandwidth parameter.
- Does it act like a similarity score?
- Why "radial"?
- Have we departed from our "inner product of feature vector" recipe?
  - Yes and no: corresponds to an infinite dimensional feature vector
- Probably the most common nonlinear kernel.

The Infinite Dimensional Feature Vector for RBF

- Consider RBF kernel (1-dim): $k(x,x') = \exp\left(-(x-x')^2/2\right)$
- We claim that $\psi : \mathbf{R} \to \ell_2$, defined by

$$[\psi(x)]_j = \frac{1}{\sqrt{j!}} e^{-x^2/2} x^j$$

  gives the **"infinite-dimensional feature vector" corresponding to RBF kernel**.
- Is this mapping even well-defined? Is $\psi(x)$ even an element of ℓ_2?
- Yes:

$$\sum_{j=0}^{\infty} \frac{1}{j!} e^{-x^2} x^{2j} = e^{-x^2} \sum_{j=0}^{\infty} \frac{\left(x^2\right)^j}{j!} = 1 < \infty$$

.

When is $k(x, x')$ a kernel function? (Mercer's Theorem)

How to Get Kernels?

1. Explicitly construct $\psi(x) : \mathcal{X} \to \mathbf{R}^d$ and define $k(x, x') = \psi(x)^T \psi(x')$.
2. Directly define the kernel function $k(x, x')$, and verify it corresponds to $\langle \psi(x), \psi(x') \rangle$ for some ψ.

There are many theorems to help us with the second approach

Positive Semidefinite Matrices

Definition

A real, symmetric matrix $M \in \mathbf{R}^{n \times n}$ is **positive semidefinite (psd)** if for any $x \in \mathbf{R}^n$,

$$x^T M x \geqslant 0.$$

Theorem

The following conditions are each necessary and sufficient for a symmetric matrix M to be positive semidefinite:

- *M has can be factorized as $M = R^T R$, for some matrix R.*
- *All eigenvalues of M are greater than or equal to 0.*

Positive Semidefinite Function

Definition

A symmetric kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ is **positive semidefinite (psd)** if for any finite set $\{x_1, \ldots, x_n\} \in \mathcal{X}$, the kernel matrix on this set

$$K = \left(k(x_i, x_j)\right)_{i,j} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \cdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix}$$

is a positive semidefinite matrix.

Mercer's Theorem

Theorem

A symmetric function $k(x,x')$ can be expressed as an inner product

$$k(x,x') = \langle \psi(x), \psi(x') \rangle$$

for some ψ if and only if $k(x,x')$ is ***positive semidefinite.***

Generating New Kernels from Old

- Suppose $k, k_1, k_2 : \mathcal{X} \times \mathcal{X} \to \mathbf{R}$ are psd kernels. Then so are the following:

$$
\begin{aligned}
k_{\text{new}}(x, x') &= k_1(x, x') + k_2(x, x') \\
k_{\text{new}}(x, x') &= \alpha k(x, x') \\
k_{\text{new}}(x, x') &= f(x) f(x') \text{ for any function } f(\cdot) \\
k_{\text{new}}(x, x') &= k_1(x, x') k_2(x, x')
\end{aligned}
$$

- See Appendix for details.
- Lots more theorems to help you construct new kernels from old...

Details on New Kernels from Old [Optional]

Additive Closure

- Suppose k_1 and k_2 are psd kernels with feature maps ϕ_1 and ϕ_2, respectively.
- Then
$$k_1(x,x') + k_2(x,x')$$
is a psd kernel.
- Proof: Concatenate the feature vectors to get
$$\phi(x) = (\phi_1(x), \phi_2(x)).$$
Then ϕ is a feature map for $k_1 + k_2$.

Closure under Positive Scaling

- Suppose k is a psd kernel with feature maps ϕ.
- Then for any $\alpha > 0$,

$$\alpha k$$

  is a psd kernel.
- Proof: Note that

$$\phi(x) = \sqrt{\alpha}\phi(x)$$

  is a feature map for αk.

Scalar Function Gives a Kernel

- For any function $f(x)$,

$$k(x,x') = f(x)f(x')$$

  is a kernel.
- Proof: Let $f(x)$ be the feature mapping. (It maps into a 1-dimensional feature space.)

$$\left\langle f(x), f(x')\right\rangle = f(x)f(x') = k(x,x').$$

Closure under Hadamard Products

- Suppose k_1 and k_2 are psd kernels with feature maps ϕ_1 and ϕ_2, respectively.
- Then
$$k_1(x,x')k_2(x,x')$$
is a psd kernel.
- Proof: Take the outer product of the feature vectors:
$$\phi(x) = \phi_1(x)\left[\phi_2(x)\right]^T.$$
Note that $\phi(x)$ is a matrix.
- Continued...

Closure under Hadamard Products

- Then

$$
\begin{aligned}
\langle \phi(x), \phi(x') \rangle &= \sum_{i,j} \phi(x)\phi(x') \\
&= \sum_{i,j} \left[\phi_1(x) \left[\phi_2(x)\right]^T \right]_{ij} \left[\phi_1(x') \left[\phi_2(x')\right]^T \right]_{ij} \\
&= \sum_{i,j} \left[\phi_1(x)\right]_i \left[\phi_2(x)\right]_j \left[\phi_1(x')\right]_i \left[\phi_2(x')\right]_j \\
&= \left(\sum_i \left[\phi_1(x)\right]_i \left[\phi_1(x')\right]_i \right) \left(\sum_j \left[\phi_2(x)\right]_j \left[\phi_2(x')\right]_j \right) \\
&= k_1(x, x') k_2(x, x')
\end{aligned}
$$

Questions on Kernel Methods

1. Fix $n > 0$. For $x, y \in \{1, 2, \ldots, n\}$ define $k(x, y) = \min(x, y)$. Give an explicit feature map $\phi : \{1, 2, \ldots, n\}$ to $\mathbf{R}^D$ (for some D) such that $k(x, y) = \phi(x)^T \phi(y)$.
2. Show that $k(x, y) = (x^T y)^4$ is a positive semidefinite kernel on $\mathbf{R}^d \times \mathbf{R}^d$.
3. Let $A \in \mathbf{R}^{d \times d}$ be a positive semidefinite matrix. Prove that $k(x, y) = x^T A y$ is a positive semidefinite kernel.

1. Fix $n > 0$. For $x, y \in \{1, 2, \ldots, n\}$ define $k(x, y) = \min(x, y)$. Give an explicit feature map $\phi : \{1, 2, \ldots, n\}$ to $\mathbf{R}^D$ (for some D) such that $k(x, y) = \phi(x)^T \phi(y)$.

   Solution:

   Define $\phi(x) = (\mathbb{1}(x \geqslant 1), \mathbb{1}(x \geqslant 2), \ldots, \mathbb{1}(x \geqslant n))$. Then $\phi(x)^T \phi(y) = \min(x, y)$.

2. Show that $k(x, y) = (x^T y)^4$ is a positive semidefinite kernel on $\mathbf{R}^d \times \mathbf{R}^d$.

   Solution:

   $k_1(x, y) = x^T y$ is a psd kernel, since $x^T y$ is an inner product on $\mathbf{R}^d$. Using the product rule for psd kernels, we see that

$$k(x, y) = k_1(x, y) k_1(x, y) k_1(x, y) k_1(x, y) = k_1(x, y)^4$$

   is psd as well.

1. Let $A \in \mathbf{R}^{d \times d}$ be a positive semidefinite matrix. Prove that $k(x,y) = x^T Ay$ is a positive semidefinite kernel.

   Solution:
   Fix $x_1, \ldots, x_n \in \mathbf{R}^d$ and let X denote the matrix that has x_i^T as its ith row. Then note that $(XAX^T)_{ij} = x_i^T Ax_j = k(x_i, x_j)$. Thus we are done if we can show XAX^T is positive semidefinite. But note that, for any $\alpha \in \mathbf{R}^n$,

   $$\alpha^T XAX^T \alpha = (X^T \alpha)^T A (X^T \alpha) \geqslant 0,$$

   since A is positive semidefinite.

1. Suppose you are given an training set of distinct points $x_1, x_2, \dots, x_n \in \mathbf{R}^n$ and labels $y_1, \dots, y_n \in \{-1, +1\}$. Show that by properly selecting σ you can achieve perfect $0-1$ loss on the training data using a linear decision function and the RBF kernel.
2. Consider the standard (unregularized) linear regression problem where we minimize $L(w) = \|Xw - y\|_2^2$ for some $X \in \mathbf{R}^{n \times m}$ and $y \in \mathbf{R}^n$. Assume $m > n$.
    1. Let w^* be one minimizer of the loss function L above. Give an infinite set of minimizers of the loss function.
    2. What property defines the minimizer given by the representer theorem (in terms of X)?

1. Suppose you are given an training set of distinct points $x_1, x_2, \ldots, x_n \in \mathbf{R}^n$ and labels $y_1, \ldots, y_n \in \{-1, +1\}$. Show that by properly selecting σ you can achieve perfect $0-1$ loss on the training data using a linear decision function and the RBF kernel.

Solution:

By selecting σ sufficiently small (say, much smaller than $\min_{i \neq j} \|x_i - x_j\|_2$) we can use $\alpha_i = y_i$ and get very pointy spikes at each data point. Kernelized prediction function will be:

$$f(x) = \sum_{i=1}^{n} y_i \exp(-\|x - x_i\|_2^2 / \sigma^2),$$

$$f(x_j) = y_j + \sum_{i \neq j} y_i \exp(-\|x_j - x_i\|_2^2 / \sigma^2),$$

where $|y_j| >> |\sum_{i \neq j} y_i \exp(-\|x_j - x_i\|_2^2 / \sigma^2)|$.

[Note: This is not possible if any repeated points have different labels, which is not unusual in real data.]

1. Consider the standard (unregularized) linear regression problem where we minimize $L(w) = \|Xw - y\|_2^2$ for some $X \in \mathbf{R}^{n \times m}$ and $y \in \mathbf{R}^n$. Assume $m > n$.
    1. Let w^* be one minimizer of the loss function L above. Give an infinite set of minimizers of the loss function.
    2. What property defines the minimizer given by the representer theorem (in terms of X)?

Solution:

1. $\{w^* + v \mid v \in \text{Null}(X)\}$. Using the standard inner product on $\mathbf{R}^n$, we can also write $\text{Null}(X)$ as the set of all vectors orthogonal to the row space of X.x
2. w^* lies in the row space of X.