Kernel Methods: Wrapup and Review

David Rosenberg

New York University

February 28, 2017

Kernelization

Linear Models

- So far we've discussed
  - Linear regression
  - Ridge regression
  - Lasso regression
  - Support Vector Machines
  - Perceptrons
- Each of these methods assumes
  - Input space $\mathcal{X}$.
  - Feature map $\psi : \mathcal{X} \to \mathbf{R}^d$.
  - Linear (or affine) hypothesis space:

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

What is a Kernelized Method?

Definition

A method is **kernelized** if every reference to an element of the input space $x_1 \in \mathcal{X}$ occurs in an inner product with another element of the input space, such as $\langle \psi(x_1), \psi(x_2) \rangle$ for some $x_2 \in \mathcal{X}$.

- The **kernel function** corresponding to ψ is

$$k(x_1, x_2) = \langle \psi(x_1), \psi(x_2) \rangle .$$

Is it Kernelized?

- What if $\mathcal{X} = \mathbf{R}^d$ and we see x's always show up as $x_i^T x_j$. Is that kernelized?
- Yes! Consider the identity feature map $\psi(x) = x$ with the standard inner product.

- What if $x's$ only show up in XX^T?
- Yes! Every matrix entry is an inner product: $\left(XX^T\right)_{ij} = x_i^T x_j$. .

- What if $x's$ only show up in $X^T X$?
- No! Every matrix entry is inner product between single features:

$$\left(X^T X\right)_{ij} = f_i^T f_j,$$

where f_i is the ith coordinate for all x's.

A Generalized Linear Objective Function

Generalize from SVM Objective

- Featurized SVM objective:

$$\min_{w \in \mathbf{R}^d} \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\left(1 - y_i\left[\langle w, \psi(x_i)\rangle\right]\right)_+ .$$

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, \psi(x_1)\rangle, \ldots, \langle w, \psi(x_n)\rangle\right),$$

where

  - $R : \mathbf{R}^{\geqslant 0} \to \mathbf{R}$ is nondecreasing (**Regularization term**)
  - and $L : \mathbf{R}^n \to \mathbf{R}$ is arbitrary. (**Loss term**)

Generalized Linear Objective Function (Details)

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, \psi(x_1)\rangle, \ldots, \langle w, \psi(x_n)\rangle\right),$$

where

- $w, \psi(x_1), \ldots, \psi(x_n) \in \mathcal{H}$ 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. $\|w\| = \sqrt{\langle w, w\rangle}$)
- $R : \mathbf{R}^{\geqslant 0} \to \mathbf{R}$ is nondecreasing (**Regularization term**), and
- $L : \mathbf{R}^n \to \mathbf{R}$ is arbitrary (**Loss term**).

Generalized Linear Objective Function

- **Generalized objective**:

$$\min_{w \in \mathcal{H}} R\left(\|w\|\right) + L\left(\langle w, \psi(x_1)\rangle, \ldots, \langle w, \psi(x_n)\rangle\right),$$

- Why "linear"? $\langle w, \psi(x_i)\rangle$ is a generalization of predictions $w^T \psi(x_i)$
  - a linear function of $\psi(x_i) \in \mathbf{R}^d$.
- Ridge regression and SVM are of this form.
- What if we penalize with $\lambda\|w\|_2$ instead of $\lambda\|w\|_2^2$? Yes!.
- What if we use lasso regression? No! ℓ_1 norm does not correspond to an inner product.

The Representer Theorem

Theorem (Representer Theorem)

Let

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

where

- $w, \psi(x_1), \ldots, \psi(x_n) \in \mathcal{H}$ *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.* $\|w\| = \sqrt{\langle w, w\rangle}$*)*
- $R : \mathbf{R}^{\geqslant 0} \to \mathbf{R}$ *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* $w^* = \sum_{i=1}^n \alpha_i \psi(x_i)$.
[If R is strictly increasing, then all minimizers have this form. (Proof in homework.)]

The Representer Theorem (Proof)

1. Let w^* be a minimizer.
2. Let $M = \text{span}\left(\psi(x_1), \ldots, \psi(x_n)\right)$. [the "**span of the data**"]
3. Let $w = \text{Proj}_M w^*$. So $\exists \alpha$ s.t. $w = \sum_{i=1}^n \alpha_i \psi(x_i)$.
4. Then $w^\perp := w^* - w$ is orthogonal to M.
5. Projections decrease norms: $\|w\| \leqslant \|w^*\|$.
6. Since R is nondecreasing, $R(\|w\|) \leqslant R(\|w^*\|)$.
7. By (4), $\langle w^*, \psi(x_i) \rangle = \left\langle w + w^\perp, \psi(x_i) \right\rangle = \langle w, \psi(x_i) \rangle$.
8. $L\left(\langle w^*, \psi(x_1) \rangle, \ldots, \langle w^*, \psi(x_n) \rangle\right) = L\left(\langle w, \psi(x_1) \rangle, \ldots, \langle w, \psi(x_n) \rangle\right)$
9. $J(w) \leqslant J(w^*)$.
10. Therefore $w = \sum_{i=1}^n \alpha_i \psi(x_i)$ is also a minimizer.

Q.E.D.

Using Representer Theorem to Kernelize

Kernelized Predictions

- Consider $w = \sum_{i=1}^{n} \alpha_i \psi(x_i)$. (As representer theorem implies.)
- How do we make predictions for a given $x \in \mathcal{X}$?

$$
\begin{aligned}
f(x) = \langle w, \psi(x) \rangle &= \left\langle \sum_{i=1}^{n} \alpha_i \psi(x_i), \psi(x) \right\rangle \\
&= \sum_{i=1}^{n} \alpha_i \langle \psi(x_i), \psi(x) \rangle \\
&= \sum_{i=1}^{n} \alpha_i k(x_i, x)
\end{aligned}
$$

Note: $f(x)$ is a linear combination of $k(x_1, x), \ldots, k(x_n, x)$, all considered as functions of x.

Kernelized Regularization

- Consider $w=\sum_{i=1}^{n}\alpha_i\psi(x_i)$.
- What does $R(\|w\|)$ look like?

$$\begin{aligned}
\|w\|^2 &= \langle w,w\rangle \\
&= \left\langle \sum_{i=1}^{n}\alpha_i\psi(x_i), \sum_{j=1}^{n}\alpha_j\psi(x_j)\right\rangle \\
&= \sum_{i,j=1}^{n}\alpha_i\alpha_j\langle\psi(x_i),\psi(x_j)\rangle \\
&= \sum_{i,j=1}^{n}\alpha_i\alpha_j k(x_i,x_j)
\end{aligned}$$

(You should recognize the last expression as a quadratic form.)

The Kernel Matrix (a.k.a. Gram Matrix)

Definition

The **kernel matrix** or **Gram matrix** for a kernel k on a set $\{x_1, \ldots, x_n\}$ is

$$K = \left(k(x_i, x_j)\right)_{i,j} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \cdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \in \mathbf{R}^{n \times n}.$$

Kernelized Regularization: Matrix Form

- Consider $w = \sum_{i=1}^{n} \alpha_i \psi(x_i)$.
- What does $R(\|w\|)$ look like?

$$
\begin{aligned}
\|w\|^2 &= \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \\
&= \alpha^T K \alpha
\end{aligned}
$$

- So $R(\|w\|) = R\left(\sqrt{\alpha^T K \alpha}\right)$.

Kernelized Predictions

- Write $f_\alpha(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$. (Switched from $k(x_i, x)$ by symmetry of inner product.)
- Predictions on the training points have a particularly simple form:

$$
\begin{aligned}
\begin{pmatrix} f_\alpha(x_1) \\ \vdots \\ f_\alpha(x_n) \end{pmatrix} &= \begin{pmatrix} \alpha_1 k(x_1, x_1) + \cdots + \alpha_n k(x_1, x_n) \\ \vdots \\ \alpha_1 k(x_n, x_1) + \cdots + \alpha_n k(x_n, x_n) \end{pmatrix} \\
&= \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \cdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} \\
&= K\alpha
\end{aligned}
$$

Kernelized Objective

- Substituting

$$w = \sum_{i=1}^{n} \alpha_i \psi(x_i)$$

into generalized objective, we get

$$\min_{\alpha \in \mathbf{R}^n} R\left(\sqrt{\alpha^T K \alpha}\right) + L(K\alpha).$$

- No direct access to $\psi(x_i)$.
- All references are via kernel matrix K.
- (Assumes R and L do not hide any references to $\psi(x_i)$.)
- This is the **kernelized objective function.**

Kernelized SVM

- The SVM objective:

$$\min_{w \in \mathcal{H}} \frac{1}{2}\|w\|^2 + \frac{c}{n}\sum_{i=1}^{n}\left(1 - y_i\left[\langle w, \psi(x_i)\rangle\right]\right)_+ .$$

- Kernelizing yields

$$\min_{\alpha \in \mathbf{R}^n} \frac{1}{2}\alpha^T K\alpha + \frac{c}{n}\sum_{i=1}^{n}\left(1 - y_i\left(K\alpha\right)_i\right)_+$$

Kernelized Ridge Regression

- Ridge Regression:

$$\min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^{n} \left(w^T x_i - y_i\right)^2 + \lambda \|w\|^2$$

- Featurized Ridge Regression

$$\min_{w \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \left(\langle w, \psi(x_i) \rangle - y_i\right)^2 + \lambda \|w\|^2$$

- Kernelized Ridge Regression

$$\min_{\alpha \in \mathbf{R}^n} \frac{1}{n} \|K\alpha - y\|^2 + \lambda \alpha^T K \alpha,$$

where $y = (y_1, \ldots, y_n)^T$.

Prediction Functions with RBF Kernel

Radial Basis Function (RBF) / Gaussian Kernel

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

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

  where σ^2 is known as the bandwidth parameter.

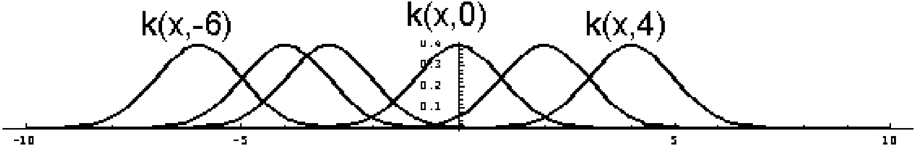
- Does it act like a similarity score?
- Why "radial"?
- Have we departed from our "inner product of feature vector" recipe?
  - Yes and no: corresponds to an infinite dimensional feature vector
- Probably the most common nonlinear kernel.

RBF Basis

- Input space $\mathcal{X} = \mathbf{R}$
- Output space: $\mathcal{Y} = \mathbf{R}$
- RBF kernel $k(w,x) = \exp\left(-(w-x)^2\right)$.
- Suppose we have 6 training examples: $x_i \in \{-6,-4,-3,0,2,4\}$.
- If representer theorem applies, then

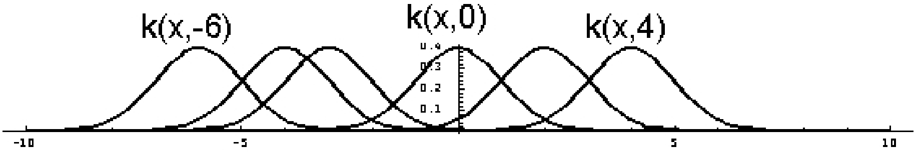
$$f(x) = \sum_{i=1}^{6} \alpha_i k(x_i, x).$$

- f is a linear combination of 6 basis functions of form $k(x_i,\cdot)$:

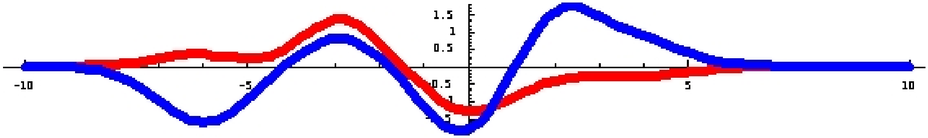


RBF Predictions

- Basis functions



- Predictions of the form $f(x) = \sum_{i=1}^{6} \alpha_i k(x_i, x)$:



- When kernelizing with RBF kernel, prediction functions always look this way.
- (Whether we get w from SVM, ridge regression, etc...)

RBF Feature Space: The Sequence Space ℓ_2

- To work with infinite dimensional feature vectors, we need a space with certain properties.
  - an inner product
  - a norm related to the inner product
  - projection theorem: $x = x_\perp + x_\|$ where $x_\| \in S = \text{span}(w_1, \ldots, w_n)$ and $\langle x_\perp, s \rangle = 0 \quad \forall s \in S$.
- Basically, we need a Hilbert space.

Definition

ℓ_2 is the space of all real-valued sequences: $(x_0, x_1, x_2, x_3, \ldots)$ with $\sum_{i=0}^{\infty} x_i^2 < \infty$.

Theorem

*With the the inner product $\langle x, x' \rangle = \sum_{i=0}^{\infty} x_i x_i'$, ℓ_2 is a **Hilbert space**.*

The Infinite Dimensional Feature Vector for RBF

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

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

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:

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

.

The Infinite Dimensional Feature Vector for RBF

- Does feature vector $[\psi(x)]_n = \frac{1}{\sqrt{n!}} e^{-x^2/2} x^n$ actually correspond to the RBF kernel?
- Yes! Proof:

$$
\begin{aligned}
\langle \psi(w), \psi(x) \rangle &= \sum_{n=0}^{\infty} \frac{1}{n!} e^{-\left(x^2+w^2\right)/2} x^n w^n \\
&= e^{-\left(x^2+w^2\right)/2} \sum_{n=0}^{\infty} \frac{(xw)^n}{n!} \\
&= \exp\left(-\left[x^2+w^2\right]/2\right) \exp(xw) \\
&= \exp\left(-\left[(x-w)^2/2\right]\right)
\end{aligned}
$$

QED