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 $\psi : \mathcal{X} \rightarrow \mathbb{R}^d$.
  - Linear (or affine) hypothesis space:

$$\mathcal{H} = \left\{ x \mapsto w^T \psi(x) \mid w \in \mathbb{R}^d \right\}.$$
What is a Kernelized Method?

Definition

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

- The **kernel function** corresponding to $\psi$ is

$$k(x_1, x_2) = \langle \psi(x_1), \psi(x_2) \rangle.$$
### Is it Kernelized?

- What if $\mathcal{X} = \mathbb{R}^d$ 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: $(XX^T)_{ij} = x_i^T x_j$.

- What if $x$’s only show up in $X^T X$?
  - No! Every matrix entry is inner product between single features:
    
    $$(X^T X)_{ij} = f_i^T f_j,$$

    where $f_i$ is the $i$th coordinate for all $x$’s.
A Generalized Linear Objective Function
Generalize from SVM Objective

- Featurized SVM objective:
  \[
  \min_{w \in \mathbb{R}^d} \frac{1}{2} \|w\|^2 + c \sum_{i=1}^{n} (1 - y_i [\langle w, \psi(x_i) \rangle])_+.
  \]

- Generalized objective:
  \[
  \min_{w \in \mathcal{H}} R(\|w\|) + L(\langle w, \psi(x_1) \rangle, \ldots, \langle w, \psi(x_n) \rangle),
  \]
  where
  - \( R : \mathbb{R}^{\geq 0} \to \mathbb{R} \) is nondecreasing (Regularization term)
  - and \( L : \mathbb{R}^n \to \mathbb{R} \) is arbitrary. (Loss term)
Generalized Linear Objective Function (Details)

- **Generalized objective:**

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

where

- \(w, \psi(x_1), \ldots, \psi(x_n) \in \mathcal{H}\) for some Hilbert space \(\mathcal{H}\). (We typically have \(\mathcal{H} = \mathbb{R}^d\).)
- \(\| \cdot \|\) is the norm corresponding to the inner product of \(\mathcal{H}\). (i.e. \(\|w\| = \sqrt{\langle w, w \rangle}\))
- \(R : \mathbb{R}^+ \to \mathbb{R}\) is nondecreasing (Regularization term), and
- \(L : \mathbb{R}^n \to \mathbb{R}\) is arbitrary (Loss term).
Generalized Linear Objective Function

- **Generalized objective:**
  \[
  \min_{w \in \mathcal{H}} R(\|w\|) + L(\langle w, \psi(x_1) \rangle, \ldots, \langle w, \psi(x_n) \rangle),
  \]

- Why “linear”? \( \langle w, \psi(x_i) \rangle \) is a generalization of predictions \( w^T \psi(x_i) \)
  - a linear function of \( \psi(x_i) \in \mathbb{R}^d \).

- 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! \( \ell_1 \) norm does not correspond to an inner product.
The Representer Theorem

Theorem (Representer Theorem)

Let

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

where

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

Let \( w^* \) be a minimizer.

Let \( M = \text{span} (\psi(x_1), \ldots, \psi(x_n)) \). [the “span of the data”]

Let \( w = \text{Proj}_M w^* \). So \( \exists \alpha \text{ s.t. } w = \sum_{i=1}^n \alpha_i \psi(x_i) \).

Then \( w^\perp := w^* - w \) is orthogonal to \( M \).

Projections decrease norms: \( \|w\| \leq \|w^*\| \).

Since \( R \) is nondecreasing, \( R(\|w\|) \leq R(\|w^*\|) \).

By (4), \( \langle w^*, \psi(x_i) \rangle = \langle w + w^\perp, \psi(x_i) \rangle = \langle w, \psi(x_i) \rangle \).

\[ L (\langle w^*, \psi(x_1) \rangle, \ldots, \langle w^*, \psi(x_n) \rangle) = L (\langle w, \psi(x_1) \rangle, \ldots, \langle w, \psi(x_n) \rangle) \]

\[ J(w) \leq J(w^*). \]

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 $x \in X$?

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

$$= \sum_{i=1}^{n} \alpha_i \langle \psi(x_i), \psi(x) \rangle$$

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

**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?

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

(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

$$K = (k(x_i, x_j))_{i,j} = \begin{pmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ \vdots & \ddots & \vdots \\ k(x_n, x_1) & \cdots & k(x_n, x_n) \end{pmatrix} \in \mathbb{R}^{n \times n}.$$
Consider $w = \sum_{i=1}^{n} \alpha_i \psi(x_i)$.

What does $R(\|w\|)$ look like?

$$\|w\|^2 = \sum_{i,j=1}^{n} \alpha_i \alpha_j k(x_i, x_j)$$

$$= \alpha^T K \alpha$$

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

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

- Substituting

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

into generalized objective, we get

\[ \min_{\alpha \in \mathbb{R}^n} R \left( \sqrt{\alpha^T K \alpha} \right) + L(K \alpha). \]

- 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:

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

- Kernelizing yields

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

- Ridge Regression:
  \[
  \min_{w \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (w^T x_i - y_i)^2 + \lambda \|w\|^2
  \]

- Featurized Ridge Regression
  \[
  \min_{w \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (\langle w, \psi(x_i) \rangle - y_i)^2 + \lambda \|w\|^2
  \]

- Kernelized Ridge Regression
  \[
  \min_{\alpha \in \mathbb{R}^n} \frac{1}{n} \|K\alpha - y\|^2 + \lambda \alpha^T K\alpha,
  \]
  where \( y = (y_1, \ldots, y_n)^T. \)
Prediction Functions with RBF Kernel
Radial Basis Function (RBF) / Gaussian Kernel

- Input space \( \mathcal{X} = \mathbb{R}^d \)

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

where \( \sigma^2 \) 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 $\mathcal{X} = \mathbb{R}$
- Output space: $\mathcal{Y} = \mathbb{R}$
- 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

![Basis functions graph](image)

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

![Prediction function graph](image)

- 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 $\ell_2$**

- 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: $x = x_\perp + x_\parallel$ where $x_\parallel \in S = \text{span}(w_1,\ldots,w_n)$ and $\langle x_\perp, s \rangle = 0 \quad \forall s \in S$.
- Basically, we need a Hilbert space.

**Definition**

$\ell_2$ 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 inner product $\langle x, x' \rangle = \sum_{i=0}^{\infty} x_i x'_i$, $\ell_2$ 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 \( \psi : \mathbb{R} \rightarrow \ell_2 \) 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 \( \ell_2 \)?
- Yes:

\[
\sum_{n=0}^{\infty} \frac{1}{n!} e^{-x^2} x^{2n} = e^{-x^2} \sum_{n=0}^{\infty} \frac{(x^2)^n}{n!} = 1 < \infty
\]
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:

$$\langle \psi(w), \psi(x) \rangle = \sum_{n=0}^{\infty} \frac{1}{n!} e^{-\frac{(x^2+w^2)}{2}} x^n w^n$$

$$= e^{-\frac{(x^2+w^2)}{2}} \sum_{n=0}^{\infty} \frac{(xw)^n}{n!}$$

$$= \exp\left(-\frac{x^2 + w^2}{2}\right) \exp(xw)$$

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

QED