CS480/680: Intro to ML

Lecture 07: Kernels
Announcement

- Final exam: Wednesday December 18, 4:00-6:30PM
  - Place TBD
Outline

• Feature map
• Kernels
• The Kernel Trick
• Advanced
XOR

\[ \text{XOR} \]

\[ \begin{array}{c}
\begin{array}{c}
\text{x}_1 \\
\text{x}_2 \\
\end{array}
\end{array} \]

\[ \begin{array}{c}
\begin{array}{c}
-1 \\
0 \\
1 \\
\end{array}
\end{array} \]

\[ \begin{array}{c}
\begin{array}{c}
-1 \\
0 \\
1 \\
\end{array}
\end{array} \]
Quadratic classifier

\[ \hat{y} = f(x) = 1 \]

Weights (to be learned)

\[ x^T Q x + \sqrt{2} x^T p + \gamma \geq 0 \]
The power of lifting

\[ x^\top Q x + \sqrt{2}x^\top p + \gamma \geq 0 \]

\[ w^\top \phi(x) \geq 0 \]

\[ \phi(x) = \begin{bmatrix} x^\top \hat{Q} x \\ \sqrt{2}x^\top \\ 1 \end{bmatrix} \]

\[ w = \begin{bmatrix} \hat{Q} \\ p \\ \gamma \end{bmatrix} \]
Example

\[ \phi(x) = [x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1] \]

\[ \phi(x) = [x_1^2, x_1x_2, x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1] \]
Does it work?
The curse of dimensionality?

\[ \phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d^2 + d + 1} \]

Computation in this space now

\[ \phi(x) = \begin{bmatrix} \overrightarrow{x}x \\ \sqrt{2}x \\ 1 \end{bmatrix} \]

But, all we need is the dot product!!!

\[ \phi(x)^T \phi(x') = (x^T x')^2 + 2x^T x' + 1 \]
\[ = (x^T x' + 1)^2 \]

- This is still computable in \( O(d) \)!
Feature transform

\[ \phi : \mathbb{R}^d \rightarrow \mathbb{R}^h \]

- NN: learn \( \phi \) simultaneously with \( w \)

- Here: choose a nonlinear \( \phi \) so that for some \( f : \mathbb{R} \rightarrow \mathbb{R} \)

\[ \phi(x) \mathbf{T} \phi(x') = f(x \mathbf{T} x') \]

save computation
Outline

- Feature map
- Kernels
- The Kernel Trick
- Advanced
Reverse engineering

• Start with some function $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$, s.t.
  exists feature transform $\phi$ with

  \[ \phi(x)^\top \phi(x') = k(x, x') \]

• As long as $k$ is efficiently computable, don’t care the dim of $\phi$ (could be infinite!)

• Such $k$ is called a (reproducing) kernel.
Examples

- Polynomial kernel
  \[ k(x, x') = (x^\top x')^p \]
  \[ k(x, x') = (x^\top x' + 1)^p \]

- Gaussian Kernel
  \[ k(x, x') = \exp\left(-\|x - x'|^2 / \sigma\right) \]

- Laplace Kernel
  \[ k(x, x') = \exp\left(-\|x - x'|_2 / \sigma\right) \]

- Matérn Kernel
  \[ \frac{1}{2^{\nu-1} \Gamma(\nu)} \left( \frac{2\sqrt{\nu}\|x - x'|_2}{\theta} \right)^\nu H_\nu \left( \frac{2\sqrt{\nu}\|x - x'|_2}{\theta} \right) \]
Verifying a kernel

For any $n$, for any $x_1, x_2, \ldots, x_n$, the kernel matrix $K$ with

$$K_{ij} = k(x_i, x_j)$$

is symmetric and positive semidefinite ($K \in S_+^d$)

- Symmetric: $K_{ij} = K_{ji}$
- Positive semidefinite (PSD): for all $\alpha \in \mathbb{R}^n$

$$\alpha^\top K \alpha = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j K_{ij} \geq 0$$
Kernel calculus

- If $k$ is a kernel, so is $\lambda k$ for any $\lambda \geq 0$
- If $k_1$ and $k_2$ are kernels, so is $k_1 + k_2$
  - $k_1$ with $\varphi_1$, $k_2$ with $\varphi_2$
  - $k_1 + k_2$ with ??
- If $k_1$ and $k_2$ are kernels, so is $k_1 k_2$
Outline

- Feature map
- Kernels
- The Kernel Trick
- Advanced
Kernel SVM (dual)

\[
\begin{aligned}
\min_{C \geq \alpha \geq 0} & \quad \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K_{ij} - \sum_{i=1}^{n} \alpha_i \\
\text{s.t.} & \quad \sum_{i} \alpha_i y_i = 0
\end{aligned}
\]

With \( \alpha \), \( w = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \) but \( \phi \) is implicit...
Does it work?

\[ 9\alpha_1 - \alpha_2 - \alpha_3 + \alpha_4 = 1 \]
\[-\alpha_1 + 9\alpha_2 + \alpha_3 - \alpha_4 = 1 \]
\[-\alpha_1 + 0.5\alpha_2 + 9\alpha_3 - \alpha_4 = 1 \]
\[\alpha_1 - \alpha_2 - \alpha_3 + 9\alpha_4 = 1 \]

\[ \alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 1/8 \]

\[ w = [0, +1/\sqrt{2}, 0, 0, 0, 0, 0] \]
\[ \phi(x) = [x_1^2, \sqrt{2}x_1x_2, x_2^2, \sqrt{2}x_1, \sqrt{2}x_2, 1] \]

\[ k(x, x') = (x^T x' + 1)^2 \]

\[ K = \begin{bmatrix} 9 & 1 & 1 & 1 & 1 \\ 1 & 9 & 1 & 1 & 0 \\ 1 & 1 & 9 & 1 & 0 \\ 1 & 1 & 1 & 9 & 0 \end{bmatrix} \]
Testing

• Given test sample $\mathbf{x}'$, how to perform testing?

$$\mathbf{w}^\top \phi(\mathbf{x}') = \sum_{i=1}^{n} \alpha_i y_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}')$$

$$= \sum_{i=1}^{n} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}')$$

No explicit access to $\phi$, again!

dual variables  \hspace{1cm}  \hspace{1cm}  training set  \hspace{1cm}  \hspace{1cm}  kernel
Tradeoff

• Previously: training $O(nd)$, test $O(d)$

• Kernel: training $O(n^2d)$, test $O(nd)$

• Nice to avoid explicit dependence on $h$ (could be inf)

• But if $n$ is also large… (maybe later)
Learning the kernel (Lanckriet et al.’04)

\[
\min_{C \geq \alpha \geq 0} \max_{\zeta \geq 0} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \left[ \sum_{s=1}^{t} \zeta_s K_{ij}^{(s)} \right] - \sum_{i=1}^{n} \alpha_i
\]

s.t. \( \sum_{i} \alpha_i y_i = 0 \)

- Nonnegative combination of \( t \) pre-selected kernels, with coefficients \( \zeta \) simultaneously learned
Logistic regression revisited

The optimal $w$ has the following form:

$$
\min_{w \in \mathbb{R}^d} \sum_i \log(1 + e^{-y_i w^\top x_i}) + \lambda \|w\|^2
$$

kernelize

$$
\min_{w \in \mathbb{R}^h} \sum_i \log(1 + e^{-y_i w^\top \phi(x_i)}) + \lambda \|w\|^2
$$

Representer Theorem (Wabha, Schölkopf, Herbrich, Smola, Dinuzzo, …).

The optimal $w$ has the following form:

$$
w = \sum_i \alpha_i \phi(x_i)$$
Outline

- Feature map
- Kernels
- The Kernel Trick
- Advanced
• What does it mean to use a kernel $k$?
  
  $w = \sum_{i=1}^{n} \alpha_i y_i \varphi(x_i) = \sum_{i=1}^{n} \beta_i \varphi(x_i)$
  
  testing: $f(x) = w^T \varphi(x) = \sum_{i=1}^{n} \beta_i k(x_i, x)$

• Take $k(x, x') = (x^T x')^2$

  $f(x) = \sum_{i=1}^{n} \beta_i (x^T z_i)^2 = \sum_{i=1}^{n} \beta_i \sum_{j=1}^{d} \sum_{k=1}^{d} x_j x_k z_{ji} z_{ki} = \sum_{j=1}^{d} \sum_{k=1}^{d} x_j x_k (\sum_{i=1}^{n} \beta_i z_{ji} z_{ki}) = \sum_{j=1}^{d} \sum_{k=1}^{d} \mu_{jk} x_j x_k$
Reproducing Kernel Hilbert Space

- Fix $x$, $k(., x) : X \to \mathbb{R}$, $z \mapsto k(z, x)$
- Vary $x$ in $X$: $\{ k(., x) : x \text{ in } X \}$
  - A set of functions from $X$ to $\mathbb{R}$
- Take linear combinations: $\{ \sum_{i=1}^{n} \beta_i k(., x_i) : x_i \text{ in } X \}$
- Define dot product: $\langle \sum_{i=1}^{n} \beta_i k(., x_i), \sum_{j=1}^{m} \gamma_j k(., z_j) \rangle = \sum_{i=1}^{n} \sum_{j=1}^{m} \beta_i \gamma_j k(x_i, z_j)$
- Complete

- Reproducing: $\langle f, k(., x) \rangle = f(x)$
Universal approximation (Micchelli, Xu, Zhang’06)

Universal kernel. For any compact set $Z$, for any continuous function $f: Z \to \mathbb{R}$, for any $\varepsilon > 0$, there exist $x_1, x_2, \ldots, x_n$ in $Z$ and $\alpha_1, \alpha_2, \ldots, \alpha_n$ in $\mathbb{R}$ such that

$$\max_{x \in Z} \left| f(x) - \sum_{i=1}^{n} \alpha_i k(x, x_i) \right| \leq \varepsilon$$

Example. The Gaussian kernel.

$$\exp\left(-\frac{\|x - x'\|^2}{2\sigma}\right)$$
Kernel mean embedding (Smola, Song, Gretton, Schölkopf, …)

\[ P \mapsto \mu_P := \mathbb{E}(\phi(X)), \text{ where } X \sim P \]

- **Characteristic kernel**: the above mapping is 1-1
- Completely preserve the information in the distribution \( P \)
- Lots of applications

\[ \mu_P = \mathbb{E}(\phi(X)) \]
Questions?