CS3780/5780
Kernel Method 2

Recap: $x \longrightarrow \phi(x)$

x: d-dim, $\phi(x)$: D-dim, $D \gg d$ (e.g. $D = 2^d$ even)

Linear in $\phi(x)$ is non-linear in x

Never explicitly enumerate in feature space

Kernel function: $k(x, y) = \phi(x)^T \phi(y)$

while $\phi(x)$ might be infinite dimensional, $k(x,y)$ can be computed efficiently

Eg. $k(x, y) = (1 + x^\top y)^p \qquad D = O(d^p)$

$k(x, y) = \prod_{\alpha=1}^{d} (1 + x_\alpha y_\alpha) \qquad D = O(2^d)$

How do we use this kernel trick?

SVM:

Minimize $\sum_{i=1}^{n} \max(0, 1 - y_i w^\top \phi(x_i)) + \frac{1}{C} \|w\|_2^2$

Logistic Regression:

Minimize $\sum_{i=1}^{n} \log(1 + \exp(-y_i w^\top \phi(x_i))) + \frac{\lambda}{2} \|w\|_2^2$

Linear Regression:

Minimize $\sum_{i=1}^{n} (w^\top \phi(x_i) - y_i)^2 + \frac{\lambda}{2} \|w\|_2^2$

If we can write Algo only in terms of inner products (ip) between data points, we can replace ip by kernel functions.

More generally : $$L(w) = \sum_{i=1}^{n} \ell\left(w^\top \phi(x_i), y_i\right) + \frac{\lambda}{2}\|w\|_2^2$$

Claim: w that minimizes L(w) admits form

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

For some $\alpha_1 \cdots \alpha_n \in \mathbb{R}$

(ie. w is in the span of $\phi(x_1) \ldots, \phi(x_n)$)

Proof:

Say $w = w_D + w_\perp$ where $w_D = \sum_{i=1}^{n} \alpha_i \phi(x_i)$, $w_\perp \perp w_D$

w_D in span of data

$w_\perp \perp$ to subspace containing data (ie $w_D \perp w_\perp$)

$\forall_i,\ w^\top \phi(x_i) = w_D^\top \phi(x_i)$

$\|w\|_2^2 = \|w_D\|_2^2 + \|w_\perp\|_2^2 + 2\, w_D^\top w_\perp$ (where $w_D^\top w_\perp = 0$)

Hence

$$L_D(w) = L_D(w_D) + \frac{\lambda}{2}\|w_\perp\|_2^2 \geq L_D(w_D)$$

Hence minimizer of L(w) will be in span of Data ($w_\perp = 0$)

What does this buy us? w is still very high dim (even ∞)

For a new point x, $$w^\top \phi(x) = \sum_{i=1}^{n} \alpha_i \phi(x_i)^T \phi(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$$

Hence is we had the α_i's, then we can compute prediction for any new x using only kernel function

Two Questions:

1. Can any function $k(x,y)$ be a kernel function for some feature space?
2. How do we compute α's given data set D?

1. A function k is a kernel function if and only if

$\forall\ x_1, \ldots, x_n$ and K the nxn kernel matrix given by $K_{ij} = k(x_i, x_j)$

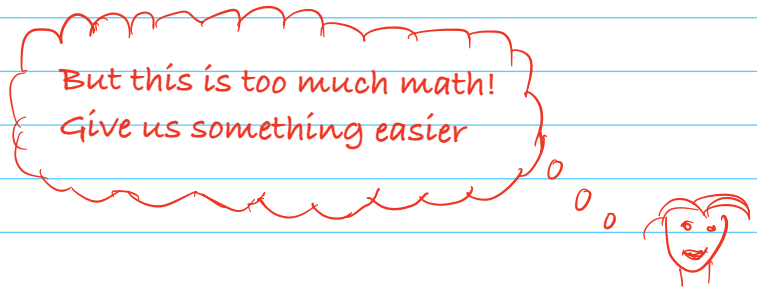
a. All eigen values of K are non-negative

$\Updownarrow$

b. $\exists$ matrix P s.t. $K = P^T P$

$\Updownarrow$

c. $\forall\ x \in \mathbb{R}^n$, $x^T K x \geq 0$



We can construct new kernels by recursively combining one or more rules from the following list:

1. $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{z}$
2. $k(\mathbf{x}, \mathbf{z}) = c k_1(\mathbf{x}, \mathbf{z})$
3. $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) + k_2(\mathbf{x}, \mathbf{z})$
4. $k(\mathbf{x}, \mathbf{z}) = g(k(\mathbf{x}, \mathbf{z}))$
5. $k(\mathbf{x}, \mathbf{z}) = k_1(\mathbf{x}, \mathbf{z}) k_2(\mathbf{x}, \mathbf{z})$
6. $k(\mathbf{x}, \mathbf{z}) = f(\mathbf{x}) k_1(\mathbf{x}, \mathbf{z}) f(\mathbf{z})$
7. $k(\mathbf{x}, \mathbf{z}) = e^{k_1(\mathbf{x}, \mathbf{z})}$
8. $k(\mathbf{x}, \mathbf{z}) = \mathbf{x}^\top \mathbf{A} \mathbf{z}$

where $c \geq 0$ and $g()$ is a polynomial with positive coefficients.

Quiz: Prove that the following functions are valid kernels

1. $k(x,y) = e^{-\frac{(x-y)^2}{\sigma^2}}$

2. For any sets $S_1, S_2 \subseteq \{1,\dots,m\}$:

$$k(S_1,S_2) = e^{|S_1 \cap S_2|}$$

How to find α's :

Lets kernelize (Ride) Regression: $w = \sum_{j=1}^{n} \alpha_j \phi(x_j)$

$$\arg\min_w \frac{1}{2}\sum_{i=1}^{n}\left(w^T\phi(x_i) - y_i\right)^2 + \frac{\lambda}{2}\|w\|_2^2$$

$$= \arg\min_\alpha \frac{1}{2}\sum_{i=1}^{n}\left(\left(\sum_{j=1}^{n}\alpha_j\phi(x_j)\right)^\top \phi(x_i) - y_i\right)^2 + \frac{\lambda}{2}\left(\sum_{j=1}^{n}\alpha_j\phi(x_j)\right)^\top\left(\sum_{i=1}^{n}\alpha_i\phi(x_i)\right)$$

$$= \arg\min_\alpha \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n}\alpha_j \underbrace{\phi(x_j)^\top\phi(x_i)}_{k(x_i,x_j)} - y_i\right)^2 + \frac{\lambda}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j \underbrace{\phi(x_j)^\top\phi(x_i)}_{k(x_i,x_j)}$$

$$= \arg\min_{\alpha} \frac{1}{2} \sum_{i=1}^{n} \left(\sum_{n}^{j=1} \alpha_j K_{i,j} - y_i \right)^2 + \frac{\lambda}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_j \alpha_i K_{i,j}$$

$$= \arg\min_{\alpha} \frac{1}{2} \|K\alpha - \mathbf{y}\|_2^2 + \frac{\lambda}{2} \alpha^\top K \alpha$$

Take gradient equate to 0

$$0 = K\left(K\alpha - \mathbf{y}\right) + \lambda K \alpha$$
$$0 = K\left(K\alpha - \mathbf{y} + \lambda I \alpha\right)$$
$$\left(K + \lambda I\right)\alpha = \mathbf{y}$$
$$\alpha = \left(K + \lambda I\right)^{-1} \mathbf{y}$$

Let us kernelize SVM: $w = \sum_{i=1}^{n} y_i \alpha_i \phi(x_i)$

original form:

$$\min_{\xi_1,\ldots,\xi_n,w,b} w^\top w + C \sum_{i=1}^{n} \xi_i$$

$$\text{s. t. } \forall i,\ y_i \left(w^\top \phi\left(x_i\right) + b\right) \geq \xi_i, \quad \xi_i \geq 0$$

$1 - \xi_i$

Dual form:

$$\min_{\alpha_1,\ldots,\alpha_n} \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K_{i,j} - \sum_{i=} \alpha_i$$

$$\text{s. t. } \forall i, 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^{n} y_i \alpha_i = 0$$

Decision function: $h_{\mathrm{SVM}}(x) = \operatorname{sign}\left(\sum_{i=1}^{n} y_i \alpha_i k(x_i, x) + b\right)$

How to compute b?

pick i s.t. $\alpha_j > 0$ (support vector)

$\left(w^\top \phi(x_i) + b\right) y_i = 1$, hence:

$$\left(\sum_{j=1}^{n} y_j \alpha_j \phi(x_j)^\top \phi(x_i) + b\right) y_i = 1$$

$$\left(\sum_{j=1}^{n} y_j \alpha_j K_{i,j} + b\right) = y_i$$

$$b = y_i - \sum_{j=1}^{n} y_j \alpha_j K_{i,j}$$

K-NN vs SVM:

$$h_{\mathrm{k-NN}}(x) = \operatorname{sign}\left(\sum_{i=1}^{n} y_i \delta_{\mathrm{k-NN}}(x, x_i)\right)$$

$\delta_{\mathrm{k-NN}}(x, x_i)$: 1 if x_i is amongst k-NN of x

$$h_{\mathrm{SVM}}(x) = \operatorname{sign}\left(\sum_{i=1}^{n} y_i \alpha_i k(x, x_i) + b\right)$$

$k(x, x_i)$: similarity; α_i: weight for each training point; b: bias

1. Often many α_i are 0, only few support vectors
2. SVM as a soft (and smarter) NN approach