CS 419M Introduction to Machine Learning **Spring 2021-22**

Lecture 12: Non-linear Classification using Kernels

Lecturer: Abir De *Scribe: Group 2*

12.1 Introducing Kernel for non-linear Classification

The SVM optimization problem in Dual is to maximise J where

$$J = \sum_{i \in D} \alpha_i - \frac{1}{2} \sum_{i \in D} \sum_{j \in D} \alpha_i \alpha_j y_i y_j x_i^T x_j \tag{12.1}$$

subject to constraints (2) and (3):

$$0 \leq \alpha \leq c \; \forall \; i \in D \tag{12.2}$$

$$\sum_{i \in D} \alpha_i y_i = 0 \tag{12.3}$$

We note that $\beta_i = c - \alpha_i$.
The present formulation of SVM cannot separate data-points which are not linearly separable. We transform the data-point to a higher dimensional space using operator Φ, where the data-points are linearly separable.
We define a kernel function which describes the inner-product of data-points in the higher-dimensional space.

$$ker(x_i, x_j) = \Phi(x_i)^T \Phi(x_j) \tag{12.4}$$

An example of a kernel function is $ker(x_i, x_j) = (x_i^T x_j + 1)^2$. Next, we will find the expression for transformation Φ for this kernel. We have,

$$\begin{aligned} ker(x_i, x_j) &= (x_i^T x_j + 1)^2 \\ &= x_i^T x_j x_j^T x_i + 2x_i^T x_j + 1 \\ &= trace(x_i^T x_j x_j^T x_i) + 2x_i^T x_j + 1 \end{aligned} \tag{12.5}$$

Using the result that $trace(A_1 A_2 A_3 A_4) = trace(A_2 A_3 A_4 A_1)$, we get

$$ker(x_i, x_j) = trace(x_j x_j^T x_i x_i^T) + 2x_i^T x_j + 1 \tag{12.6}$$

Define matrices $A = x_j x_j^T$ and $B = x_i x_i^T$. We note that B is a symmetric matrix. Then we have

$$\begin{aligned}
ker(x_i, x_j) &= trace(AB) + 2x_i^T x_j + 1 \\
&= \sum_{t,k} A_{tk} B_{kt} + 2x_i^T x_j + 1 \\
&= \sum_{t,k} A_{tk} B_{tk} + 2x_i^T x_j + 1 \\
&= \begin{bmatrix} (x_j x_j^T)_{11} \\ (x_j x_j^T)_{12} \\ (x_j x_j^T)_{13} \\ (x_j x_j^T)_{14} \\ . \\ . \\ (x_j x_j^T)_{nn} \end{bmatrix}^T \begin{bmatrix} (x_i x_i^T)_{11} \\ (x_i x_i^T)_{12} \\ (x_i x_i^T)_{13} \\ (x_i x_i^T)_{14} \\ . \\ . \\ (x_i x_i^T)_{nn} \end{bmatrix} + 2x_i^T x_j + 1
\end{aligned} \tag{12.7}$$

Using mathematical induction techniques, we can also prove that $(x_i^T x_j + 1)^d = \Phi(x_i)^T \Phi(x_j)$ for some transformation Φ.

Q. **Can we write $e^{x_i^T x_j}$ as a kernel?**
Ans. To write the function $e^{x_i^T x_j}$ as a kernel, we need to find a feature map Φ, such that $e^{x_i^T x_j} = \Phi(x_i)\Phi(x_j)^T$. Therefore, using the Taylor series expansion, it will be an infinite polynomial, $1 + x_i^T x_j + \frac{(x_i^T x_j)^2}{2!} + \frac{(x_i^T x_j)^3}{3!} + \dots$
Thus, we can make a feature map which is infinitely long such that $e^{x_i^T x_j} = \Phi(x_i)\Phi(x_j)^T$.

Q. **Can we write any $F(x_i, x_j)$ as a kernel?**
Ans. A property of kernels is that they are linear, that is,
$\alpha_1 K_1(x_1, x_2) + \alpha_2 K_2(x_1, x_2) \rightarrow K_3(x_1, x_2)$ as long as $\alpha_1, \alpha_2 > 0$.
Since we can write any function $F(x_i, x_j)$ as a polynomial using Taylor Series expansion,

$$\begin{aligned}
F(X^T X') &= \sum_{n=0}^{\infty} a_n (X^T X')^n \\
&= \sum_{n=0}^{\infty} a_n \Phi_n(X)^T \Phi_n(X')
\end{aligned}$$

And $a_n \geq 0 \; \forall \; n$.
So, using the linear property of kernels we can write the function as a kernel as long as its Taylor series converges for $n \rightarrow \infty$ and its coefficients of the series expansion are all non-negative.

12.2 Properties of a Kernel

The kernel function K has following properties.

- $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ where Φ is a transformation.
- $K(x, x) \geq 0$

  Proof:
  We have,

$$K(x_1, x_2) = \Phi(x_1)^T \Phi(x_2)$$
$$K(x, x) = ||\Phi(x)||^2 \geq 0$$

- A kernel satisfies following inequality

$$\sum_{i=0,j=0}^{n,n} c_i c_j K(x_i, x_j) \geq 0 \ \ \forall \ \ c_i, c_j \in \mathbb{R}$$

  If we define a matrix A such that $A_{ij} = K(x_i, x_j)$. Then from the previous inequality, we have

$$\sum_{i=0,j=0}^{n,n} c_i c_j K(x_i, x_j) \geq 0$$
$$\sum_{i=0,j=0}^{n,n} c_i c_j A_{ij} \geq 0$$
$$c^T A c \geq 0 \ \forall \ c \in \mathbb{R}^n$$

  Therefore, the matrix A will be a positive semi-definite matrix. Since, it is a positive semi-definite matrix, we can write $A = Q \wedge Q^T$, where Q is an orthogonal matrix.

12.3 Homework exercises:

Q1. **Suppose $K_1(x - x')$ is the form of the kernel $K(x, x')$. Then show that we can decompose $K_1(x - x')$ as $\Phi(x)^T.\Phi(x')$.**

Q2. **If $K(x, x')$ is a positive semi-definite kernel satisfying $\sum_{j=1}^{n} \sum_{i=1}^{n} C_i.C_j.K(x_i, x_j) \geq 0$, then can we prove that $K(x, x') = \Phi(x)^T.\Phi(x')$?**
(If needed, assume additional constraints on $K(x, x')$)

Q. **If $K_1(x,x')$ and $K_2(x,x')$ are kernels, then would $K(x,x') = K_1(x,x').K_2(x,x')$ be a kernel?**

Ans. Yes, it would be. The proof proceeds as follows,

$$\begin{aligned}
K(x,x') &= K_1(x,x').K_2(x,x') \\
&= \Phi_1(x)^T.\Phi_1(x').\Phi_2(x)^T.\Phi_2(x') \\
&= \Phi_1(x)^T.\Phi_1(x').\Phi_2(x')^T.\Phi_2(x) \\
&= Tr[\Phi_1(x)^T.\Phi_1(x').\Phi_2(x')^T.\Phi_2(x)] \\
&= Tr[\Phi_2(x).\Phi_1(x)^T.\Phi_1(x').\Phi_2(x')^T] \\
&= Tr[A(x).A^T(x')]
\end{aligned} \tag{12.8}$$

This is of the form $\Phi(x)^T.\Phi(x')$ as required.

(Similar to the proof of the result: $(x_ix_j)^2$ is a kernel)

Note: If the kernel is real:

$$\Rightarrow K(x_i,x_j) = (K(x_i,x_j))^* \tag{12.9}$$
$$\Rightarrow \Phi(x_i) * \Phi(x_j) = (\Phi(x_i)^*\Phi(x_j))^* \tag{12.10}$$
$$\Rightarrow \Phi(x_i)^T\Phi(x_j) = \Phi(x_j)^T\Phi(x_i) \tag{12.11}$$
$$\Rightarrow K(x_i,x_j) = K(x_j,x_i) \tag{12.12}$$

Therefore, if the kernel is real, it is symmetric.

12.4 Regularising Kernel

The original optimization objective function that we started with for finding an optimal w was:

$$\min_w \sum_{i=1}^{n} l(w^T\Phi(x_i,y_i)) + \lambda \|w\|^2$$

In most deep learning applications, w is of very large dimension leading to an infinite dimension kernel which is not ideal for implementation.

Hence, we look for a method for regularising kernel which would serve the same purpose as of regularising w. This means we need to look for a function R such that **"R(f(x))"** acts as a regulariser where "$F(x) = w^T\Phi(x)$"

Suppose,

$$\Phi(x_i) = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 1 \\ . \\ . \\ 0 \end{bmatrix} \tag{12.13}$$

where 1 is in the i^{th} position

$$\begin{aligned} F(x_1) &= w_1 \\ F(x_2) &= w_2 \\ &.. \\ &.. \end{aligned} \tag{12.14}$$

and so on

then, $F^2(x_1) + F^2(x_2) + ...$ would act as $R(F(x))$

Therefore, $R(F(x)) = \int_{\mathbb{R}^d} C(x)F^2(x)dx$

where $C(x_i) = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 1 \\ . \\ . \\ 0 \end{bmatrix}$

where 1 is in the i^{th} position

Note: Although we solved the classification problem posing as a dual problem, we prefer primal as $\alpha_i's$ are not super stable and doesn't perform good in test case.

12.5 Group Details

200070007	Anmol Saraf
200070087	Vadapalli Arvind Narasimha
180040100	Shubham Agrawal
180260029	Saanika Choudhary
190070023	Garweet Sresth