AIML - CS 337

Lecture 14: Kernel Methods

3rd October 2022

Lecturer: Abir De *Scribe: Group 29 and Group 30*

1 Recap : SVM formulation

Recall the discussion of earlier classes

$$w^*_{svm} = \frac{\sum_i^{|D|} \alpha_i y_i x_i}{2\lambda}$$

This is linear in x, and the $dim(x) = dim(w) < \infty$ To generalise this, suppose we make it non-linear in x, but it's linear in some $\phi(x)$ which can be ∞-dimensional. Previously the similarity mechanism involved $x_i^T x$. The new similarity mechanism uses the kernel formulation $K(x_i, x)$ for e.g $K(x_i, x) = e^{-||x_i - x||^2}$. Formally, this new "similarity measure" must have some properties, which are discussed later.

2 Mathematics

We continue with our discussion on kernel methods/tricks in this lecture with more rigorous mathematics.

2.1 Inner Product Space

An inner product space (over reals) is a vector space $\mathcal{V}$ and an inner product, which is a mapping

$$\langle \cdot, \cdot \rangle : \mathcal{V} \times \mathcal{V} \to \mathcal{R}$$

that has the following properties $\forall x, y, z \in \mathcal{V}$ and $a, b \in \mathcal{R}$:

- Symmetry: $\langle x, y \rangle = \langle y, x \rangle$
- Linearity: $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$
- Positive-definiteness: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0 \iff x = 0$

For an inner product space, we define norm as $\|x\| = \sqrt{\langle x, x \rangle}$

2.2 Hilbert Space

A *Hilbert Space* is an inner product space that is complete and seperable with respect to the norm defined by the inner product.A space is called complete if all Cauchy Sequences in the space converge. Examples of Hilbert spaces include :

1. $\mathbb{R}^n$ is an Hilbert space for the Euclidean norm. The dot-product is defined as with $\langle a, b \rangle = a^T b$, the vector dot product of a and b.

2. The space l_2 of square summable sequences, with inner product $\langle x, y \rangle = \sum_{i=0}^{\infty} x_i y_i$

Definition 2.1. Kernel
Let $\mathcal{X}$ be a non-empty set. A function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if there exists a Hilbert space $\mathcal{H}$ and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in \mathcal{X}$,

$$K(x, x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$$

If we are given a function of two arguments, $K(x, x')$, the following can be used to determine if it is a valid kernel.

1. Find a feature map. But this may not be obvious sometimes, and the feature map may not be unique.

2. Can use a direct property of the function which is positive definiteness. The following lemma gives a sufficient and necessary condition.

Lemma 2.2. *Let $\mathcal{H}$ be a Hilbert space, $\mathcal{X}$ a non-empty set and $\phi : \mathcal{X} \to \mathcal{H}$. A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ implements an inner product in $\mathcal{H}$ if and only if it is positive semidefinite; namely $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$, the Gram matrix $G_{i,j} = K(x_i, x_j)$, is a positive semidefinite matrix.*

Proof. $\implies$ (If K implements an inner product then the Gram matrix is positive semidefinite)

$$\begin{aligned}
\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j K(x_i, x_j) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle a_i \phi(x_i), a_j \phi(x_j) \rangle_{\mathcal{H}} \\
&= || \sum_{i=1}^{n} a_i \phi(x_i) ||_{\mathcal{H}}^2 \geq 0
\end{aligned}$$

$\impliedby$ For this direction, define the space of functions over $\mathcal{X}$ as $\mathbb{R} = \{f : \mathcal{X} \to \mathbb{R}\}$ For each $x \in \mathcal{X}$ let $\phi(x)$ be the function $x \mapsto K(\cdot, x)$. Define a vector space by taking all linear combinations of elements of the form $K(\cdot, x)$. Define an inner product on this vector space to be

$$\langle \sum_i \alpha_i K(\cdot, x_i), \sum_j \beta_j K(\cdot, x'_j) \rangle = \sum_{i,j} \alpha_i \beta_j K(x_i, x'_j)$$

This is a valid inner product since it is symmetric (because K is symmetric), it is linear, and it is positive definite. Clearly,

$$\langle \phi(x), \phi(x') \rangle = \langle K(\cdot, x), K(\cdot, x') \rangle = K(x, x').$$

□

2.3 Projection Theorem & Properties

Theorem 2.3. *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{M}$ be a closed subspace of $\mathcal{H}$. Then for any $x \in \mathcal{H}$, there exists a unique $m_0 \in \mathcal{M}$ for which*

$$\|x - m_0\| \leq \|x - m\| \forall m \in \mathcal{M}$$

This m_0 is called the projection of x onto $\mathcal{M}$. Furthermore, $m_0 \in \mathcal{M}$ is the projection of x onto M iff

$$x - m_0 \perp \mathcal{M}$$

Theorem 2.4. *Let $\mathcal{M}$ be a closed subspace of $\mathcal{H}$. For any $x \in \mathcal{H}$, let m_0 be the projection of x onto $\mathcal{M}$. Then*

$$\|m_0\| \leq \|x\|$$

with equality only when $m_0 = x$.

3 Generalised objective function

Consider

$$\underset{w}{min} \;\; l(\{w^T \phi(x_i)\}_{i \in D}, \{y_i\}_{i \in D}) + \lambda R(||w||) \tag{1}$$

where $l : \mathbb{R}^{|D|} \to \mathbb{R}$ is an arbitrary function and $R : \mathbb{R}_+ \to \mathbb{R}$ is a monotonically non-decreasing function.

Theorem 3.1. ***Representer Theorem***

Assume that ϕ is a mapping from $\mathcal{X}$ to a Hilbert space. Then, there exists a vector $\alpha \in \mathbb{R}^{|D|}$ such that $w = \sum_{i=1}^{|D|} \alpha_i \phi(x_i)$ is an optimal solution of equation 1

Proof. Let w^* be an optimal solution of Equation 1. Because w^* is an element of a Hilbert space, we can rewritew^* as

$$w* = \sum_{i=1}^{|D|} \alpha_i \phi(x_i) + u$$

where $\langle u, \phi(x_i) \rangle = 0$ for all i. Set $w = w^* - u$. Clearly, $||w^*||^2 = ||w||^2 + ||u||^2$, thus $||w|| < ||w^*||$. Since R is non-decreasing we obtain that $R(||w||) < R(||w^*||)$. Additionally, for all i we have that

$$\langle w, \phi(x_i) \rangle = \langle w^* - u, \phi(x_i) \rangle = \langle w^*, \phi(x_i) \rangle,$$

hence

$$l(\{w^T\phi(x_i)\}_{i\in D}, \{y_i\}_{i\in D}) = l(\{w^{*T}\phi(x_i)\}_{i\in D}, \{y_i\}_{i\in D})$$

We have shown that the objective of Equation 1 at w cannot be larger than the objective at w^* and therefore w is also an optimal solution. Since $w = \sum_{i=1}^{|D|} \alpha_i \phi(x_i)$ we conclude our proof. $\square$

Form of f is

$$\begin{aligned} f(x) &= w^{*T}\phi(x_i) \\ &= \sum_{i=1}^{|D|} \alpha_i \phi^T(x_i)\phi(x) \end{aligned}$$

Here $\phi^T(x_i)\phi(x)$ is like a similarity measure If $\phi(\cdot)$ is ∞-dimensional, we can write it as

$$f(x) = \sum_{i=1}^{|D|} \alpha_i \sum_{j=0}^{\infty} \phi(x_i)[j]\phi(x)[j]$$

Hence, if $\phi(\cdot)$ is ∞-dimensional, it is not feasible to code w as it has the same dimension as ϕ. So, we try to represent the objective function in functional form or through the kernel formulation.

3.1 Objective in terms of Kernel

Writing $w = \sum_{j=1}^{|D|} \alpha_j \phi(x_j)$, we have that for all i

$$\langle w, \phi(x_i)\rangle = \langle \sum_{j=1}^{|D|} \alpha_j \phi(x_j), \phi(x_i\rangle = \sum_{j=1}^{|D|} \alpha_j \langle \phi(x_j), \phi(x_i)\rangle.$$

Similarly,

$$||w||^2 = \langle \sum_{j=1}^{|D|} \alpha_j \phi(x_j), \sum_{j=1}^{|D|} \alpha_j \phi(x_j)\rangle = \sum_{i,j=1}^{|D|} \alpha_i \alpha_j \langle \phi(x_i), \phi(x_j)\rangle.$$

Let $K(x, x') = \langle \phi(x), \phi(x')\rangle$ be a function that implements the kernel function with respect to the feature space. Hence, instead of solving Equation 1, we can solve the equivalent problem

$$\underset{\alpha\in\mathbb{R}^{|D|}}{min}\, l(\{\sum_{j=1}^{|D|} \alpha_j K(x_j, x_i)\}_{i\in D}, \{y_i\}_{i\in D}) + \lambda R(\sqrt{\sum_{i,j=1}^{|D|} \alpha_i \alpha_j K(x_j, x_i)}) \tag{2}$$

3.2 Objective in terms of functional form

$$f(x) = \sum_i \alpha_i K(x_i, x)$$

Function f forms a vector space

$$f_1, f_2 \in V \implies af_1 + bf_2 \in V$$

$$0 \in V, \text{ by putting } \alpha_i = 0 \quad \forall i$$

Now define an inner product

$$\langle f, g\rangle_H = \sum \alpha_i^f \alpha_j^g K(x_i, x_j) \tag{3}$$

From the properties of inner product space, for 3 to be true, $\sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) \geq 0$ and $K(x_i, x_j) = K(x_j, x_i) \quad \forall i, j$

These are in accordance with the earlier lemma which we proved.

$$\begin{aligned}
||w||^2 &= \langle w, w\rangle \\
&= \langle \sum_{i=1}^{|D|} \alpha_i \phi(x_i), \sum_{i=1}^{|D|} \alpha_j \phi(x_j)\rangle = \sum_{i,j=1}^{|D|} \alpha_i \alpha_j \langle \phi(x_i)\phi(x_j)\rangle \\
&= \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j) = \langle f, f\rangle \\
&= ||f||^2
\end{aligned}$$

Hence, the objective function becomes

$$\min_w \; l(\{f(x_i)\}_{i\in D}, \{y_i\}_{i\in D}) + \lambda R(||f||)$$

4 Kernel Matrix & Prediction function

4.1 Kernel Matrix

We define $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle$.

Definition 4.1. We define the kernel matrix for a kernel k on a set $\{x_1, ..., x_n\}$ is

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

4.2 Prediction Function

Consider the minimizer $w = \sum_{i=1}^{n} \alpha_i \phi(x_i)$ according to the representer theorem. Then for a given x, we define the prediction function as

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

5 Different forms of Objective Function

5.1 In terms of Kernel Matrix and α

Consider $w = \sum_{i=1}^{n} \alpha_i \phi(x_i)$. Then we have for norm

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

Similarily, predictions on the training points have a particular simple form:

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

Hence our generalised objective function can be reduced to using the knowledge that minimizer lies in the span of $\phi(x_1), ..., \phi(x_n)$

$$\min_{\alpha \in \mathcal{R}^n} R(\sqrt{\alpha^T K \alpha}) + L(K\alpha)$$

This is the kernelized objective function

5.2 In terms of prediction function

Recall that $f(\cdot) = \sum_{i=1}^{m} \alpha_i k(\cdot, x_i)$. Now we define a dot product of f and another function $g(\cdot) = \sum_{i}^{m'} \beta_j k(\cdot, x'_j)$ as follows

$$\langle f, g\rangle := \sum_i^m \sum_j^{m'} \alpha_i \beta_j k(x_i, x'_j)$$

Now we try to find the condition on kernel k, such that f belongs to Hilbert space so that we can define norm of f.
Symmetry can be seen as follows:

$$\langle f, g\rangle = \sum_{j=1}^{m'} \beta_j f(x'_j) = \sum_{i=1}^{m} \alpha_i g(x_i)$$

This implies $\langle f, g\rangle = \langle g, f\rangle$ if $k(x_i, x_j) = k(x_j, x_i)$.
Positive definiteness can be seen as follows:

$$\langle f, f\rangle = \sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) \geq 0 \; \forall \alpha_i, \alpha_j \in \mathcal{R}$$

This property holds true when the kernel matrix K is positive semi-definite.
Similarly, linearity is also true without any further assumption on the kernel. Hence $\|f\|^2 = \langle f, f\rangle = \sum_i \sum_j \alpha_i \alpha_j k(x_i, x_j) = \|w\|^2$. Hence we can substitute $\|w\|^2$ with $\|f\|^2$ with the given properties of K. Hence our generalised loss function becomes

$$\min_f R(\|f\|) + L(f(x_1), f(x_2), ..., f(x_n))$$

Note: If $\forall x |f(x)| \leq M_x \|f\|_H$ then $\exists f(x) = \sum_i \alpha_i k(x_i, x)$

5.3 Need for such substitution

If $\phi(x)$ has a very large or ∞ dimension, it is impossible to code w as it has the same dimension as $\psi(x)$. So we can either go with the kernel matrix or make our analysis on the prediction function, both of which are independent of the dimension of $\phi(x)$. This is a useful tool for analysing the correctness of RBF kernel where $\phi(x)$ is of infinite dimension.

6 Reproducing kernel Hilbert spaces

For a Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$, and for any point $x \in \mathcal{X}$, the evaluation functional at x is defined as the map $L_x : \mathcal{H} \mapsto \mathbb{R}$ such that for all functions $f \in \mathcal{H}$,

$$L_x(f) = f(x). \tag{4}$$

In this setting, $\mathcal{H}$ is called a reproducing kernel Hilbert space if for all $x \in \mathcal{X}$, L_x is bounded, i.e. there is some finite constant M such that

$$|L_x(f)| = |f(x)| \leq M\|f\|_{\mathcal{H}}. \tag{5}$$

(Equivalently, for all $x \in \mathcal{X}$, L_x is continuous at any $f \in \mathcal{H}$.)

7 Example problem

7.1 Problem Statement

Consider the functions $h : \mathbb{N} \to [1 \ldots m]$ and $\mathcal{E} : \mathbb{N} \to \pm 1$

$$a^{h,\mathcal{E}}(x)[i] = \sum_{j \text{ s.t } h(i)=j} \mathcal{E}(j) x_j$$

Then prove that

$$\mathop{\mathbb{E}}_{h,\mathcal{E}\sim\ \mathcal{U}(.)} [\langle a^{h,\mathcal{E}}(x), a^{h,\mathcal{E}}(x') \rangle] = \langle x, x' \rangle$$

7.2 Solution

$$\mathop{\mathbb{E}}_{h,\mathcal{E}\sim\ \mathcal{U}(.)} [\langle a^{h,\mathcal{E}}(x), a^{h,\mathcal{E}}(x') \rangle] = \mathop{\mathbb{E}}_{h,\mathcal{E}\sim\ \mathcal{U}(.)} [\sum_{j;h(i)=j} \sum_{j';h(i)=j'} \mathcal{E}(j)\mathcal{E}(j') x_j x'_{j'}]$$

Note that since h and $\mathcal{E}$ are sampled from uniform distributions, for $j \neq j'$

$$\mathbb{E}[\mathcal{E}(j)\mathcal{E}(j')] = 0$$

and when $j = j'$, $\mathbb{E}[\mathcal{E}(j)\mathcal{E}(j)] = 1$

Therefore the expectation simplifies to

$$\mathop{\mathbb{E}}_{h,\mathcal{E}\sim\ \mathcal{U}(.)} [\sum_{j=j'} (1) * x_j x'_{j'} + 0] = \mathbb{E}[\sum_j x(j) x'(j)] = \mathbb{E}[\langle x, x' \rangle] = \langle x, x' \rangle$$

8 Mercer's Theorem

Theorem 8.1. *A "symmetric" function $k(x, x')$ can be expressed as an inner product*

$$k(x, x') = \langle \psi(x), \psi(x') \rangle$$

for some ψ if and only if K (kernel matrix) is positive semi-definite (and symmetric).