AIML - CS 337

Lecture 14: Kernel Methods

3rd October 2022

Lecturer: Abir De *Scribe: Gurpreet,Harshvardhan,Sai Pavan,Yash*

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

1 Mathematics

1.1 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 $\|x\| = \sqrt{\langle x, x \rangle}$

1.2 Hilbert Space

A Hilbert Space is a complete inner product space. A space is called complete if all Cauchy Sequences in the space converge.

1.3 Projection Theorem & Properties

Theorem 1.1. *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 1.2. *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 $m_0 = x$.

2 Representer Theorem

2.1 Generalised Objective Function

Definition 2.1. Generalised Objective Function is given by

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

where

- $w, \psi(x_1), ..., \psi(x_n) \in \mathcal{H}$ for some Hilbert space $\mathcal{H}$.
- $\|\cdot\|$ is the norm corresponding to the inner product of $\mathcal{H}$.(i.e. $\|w\| = \sqrt{\langle w, w\rangle}$)
- $R : [0, \infty] \to \mathcal{R}$ is non-decreasing
- $L : \mathcal{R}^n \to \mathcal{R}$ is arbitrary

Note that Ridge regression and SVM of this form but lasso regression is not.

2.2 Representer Theorem

Theorem 2.2. *Let*

$$J(w) = R(\|w\|) + L(\langle w, \psi(x_1)\rangle, ..., \langle w, \psi(x_n)\rangle)$$

with the properties of generalised objective function as defined above. Now, if $J(w)$ has a minimizer, then it has a minimizer of the form $w^* = \sum_{i=1}^n \alpha_i \psi(x_i)$.

Proof. Let w be a minimizer of $J(w')$ and $\mathcal{M} = \text{span}(\psi(x_i), ..., \psi(x_n))$.Let w^* be the projection of w on $\mathcal{M}$. Then we know that $\|w^*\| \leq \|w\|$. Since R is nondecreasing, we have $R(\|w^*\|) \leq R(\|w\|)$. Since $w - w^* \perp \psi(x_i)$, this implies $\langle w - w^*, \psi(x_i)\rangle = 0$ or $\langle w, \psi(x_i)\rangle = \langle w^*, \psi(x_i)\rangle$. Thus there is no change in the value of L. Hence $J(w^*) \leq J(w)$,therefore $w^* = \sum_{i=1}^n \alpha_i \psi(x_i)$ is also a minimizer. □

3 Kernel Matrix & Prediction function

3.1 Kernel Matrix

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

Definition 3.1. We define the kernel matrix for a kernel k on a set $\{x_1, ..., x_n\}$ is

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

3.2 Prediction Function

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

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

4 Different forms of Objective Function

4.1 In terms of Kernel Matrix and α

Consider $w = \sum_{i=1}^{n} \alpha_i \psi(x_i)$. 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 $\psi(x_1), ..., \psi(x_n)$

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

This is the kernelized objective function

4.2 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 $\exists f(x) = \sum_i \alpha_i k(x_i, x)$

4.3 Need for such substitution

If $\psi(x)$ 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 $\psi(x)$. This is a useful tool for analysing the correctness of RBF kernel where $\psi(x)$ is of infinite dimension.

5 Mercer's Theorem

Theorem 5.1. *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).