AIML - CS 337

Lecture 13: Kernel Methods II

September 29, 2022

Lecturer: Abir De *Scribe: Swayam, Koustubh, Teja, Parth*

1 Kernel over sets

1.1 Problem Statement

A kernel $K(x_1, x_2)$ is defined as $\langle \phi(x_1), \phi(x_2) \rangle$ for some ϕ. Suppose x_1, x_2 are being sampled from two distributions namely P_1, P_2 then $K(A, B)$ is defined as

$$K(A,B) = \iint_{x_1 \in A, x_2 \in B} \langle \phi(x_1), \phi(x_2) \rangle \, dp(x_1, x_2) \tag{1}$$

Where A, B are two sets and $p(x_1, x_2)$ is the density of the joint probability distribution of P_1, P_2 such that

$$\iint_{\mathbb{R}^2} dp(x_1, x_2) = 1$$

Devise a ϕ such that $K(A, B) = P(A \cap B) - P(A)P(B)$

1.2 Solution

Consider

$$\phi(x) = \begin{cases} \mathbb{1}_A(x) - P(A) & x \sim P1 \\ \mathbb{1}_B(x) - P(B) & x \sim P2 \end{cases}$$

where $\mathbb{1}$ is the indicator function i.e for some set S

$$\mathbb{1}_S(x) = \begin{cases} 1 & x \in S \\ 0 & \text{otherwise} \end{cases}$$

Substitute the proposed ϕ in equation (1)

$$\begin{aligned} K(A,B) &= \iint_{\mathbb{R}^2} \langle \mathbb{1}_A(x) - P(A), \mathbb{1}_B(x) - P(B) \rangle \, dp(x_1, x_2) \\ &= \iint_{\mathbb{R}^2} (\mathbb{1}_A(x) - P(A))(\mathbb{1}_B(x) - P(B)) \, dp(x_1, x_2) \\ &= \iint_{\mathbb{R}^2} (\mathbb{1}_A(x)\mathbb{1}_B(x) + P(A)P(B) - P(A)\mathbb{1}_B(x) - P(B)\mathbb{1}_A(x)) \, dp(x_1, x_2) \\ &= P(A \cap B) + P(A)P(B) - P(A)P(B) - P(A)P(B) \\ &= P(A \cap B) - P(A)P(B) \end{aligned}$$

2 Homework Problem

Problem. Show that the following kernel is positive semidefinite:

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$$

Solution. Note the following equality:

$$\int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{\pi}} \exp\left(-\frac{(x-z)^2}{\sigma^2}\right) \frac{1}{\sigma\sqrt{\pi}} \exp\left(-\frac{(y-z)^2}{\sigma^2}\right)\, dz = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-y)^2}{2\sigma^2}\right)$$

which is equivalent to the following (assuming a Euclidean norm):

$$\int_{\mathbb{R}^n} \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^n \exp\left(-\frac{\|\mathbf{x}-\mathbf{z}\|^2}{\sigma^2}\right) \exp\left(-\frac{\|\mathbf{y}-\mathbf{z}\|^2}{\sigma^2}\right)\, d\mathbf{z} = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)$$

Let $\{\mathbf{x}_i\}_{i\in\mathcal{D}}$ be a set of data points. Then, for any sequence of real numbers $\{c_i\}_{i\in\mathcal{D}}$, we have

$$\begin{aligned}
&\sum_{i\in\mathcal{D}}\sum_{j\in\mathcal{D}} c_i c_j \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{x}_j\|^2}{2\sigma^2}\right) \\
&= \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^n \int_{\mathbb{R}^n} \sum_{i,j\in\mathcal{D}\times\mathcal{D}} c_i c_j \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{z}\|^2}{\sigma^2}\right) \exp\left(-\frac{\|\mathbf{x}_j-\mathbf{z}\|^2}{\sigma^2}\right)\, d\mathbf{z} \\
&= \left(\frac{1}{\sigma}\sqrt{\frac{2}{\pi}}\right)^n \int_{\mathbb{R}^n} \left[\sum_{i\in\mathcal{D}} c_i \exp\left(-\frac{\|\mathbf{x}_i-\mathbf{z}\|^2}{\sigma^2}\right)\right]^2 d\mathbf{z} \\
&\geq 0
\end{aligned}$$

which is obviously non-negative. This completes the proof. ■

3 Neural Tangent Kernel

We use the Neural Tangent Kernel to compute the similarity as the training progresses. Consider a vector set $\mathcal{V}$, where $v_i \in \mathcal{V}$ for all $i \in \mathcal{D}$ is given as follows:

$$v_i = \begin{bmatrix} \text{loss}(x_i \mid t=0) \\ \text{loss}(x_i \mid t=1) \\ \text{loss}(x_i \mid t=2) \\ \vdots \end{bmatrix}$$

which records the loss of each data point during training.

If the gradients at two points are the same, they are said to have higher similarity. We initialize weights randomly and calculate gradients for each initialization and then compute the expected value of the aforementioned gradients and use that as a measure of similarity.

The kernel can be explicitly be written as:

$$K(x_i, x_j) = \mathbb{E}[\ \langle \nabla_w l(h_w(x_i), y_i),\ \ \nabla_w l(h_w(x_j), y_j) \rangle\ ]$$

Where ∇_w is the gradient.

4 Final Problem

Consider now the following optimization objective:

$$\min_{f \in \Lambda} \sum_{i \in \mathcal{D}} (y_i - f(x_i))^2 + \lambda \sum_{i \in \mathcal{D}} f(x_i)^2$$

where f is defined in the vector space Λ of functions generated by the set

$$\{k(x_i, \cdot)\}_{i \in \mathcal{D}}$$

which is a linear subspace of $\mathbb{R}^{\mathcal{X}}$, where $x_i \in \mathcal{X}$ for all $i \in \mathcal{D}$ and $k(\cdot, \cdot)$ is the kernel function defined $\mathcal{X} \times \mathcal{X} \xrightarrow{k} \mathbb{R}$. This vector space is also equipped with the following inner product:

$$\left\langle \sum_{i \in \mathcal{D}} \alpha_i k(x_i, \cdot), \sum_{j \in \mathcal{D}} \beta_j k(x_j, \cdot) \right\rangle = \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}} \alpha_i \beta_j k(x_i, x_j)$$

That the above is an inner product space is easily verified. Indeed, for all $g \in \Lambda$, there are $\alpha_i \in \mathbb{R}$ for all $i \in \mathcal{D}$ such that $g = \sum_{i \in \mathcal{D}} \alpha_i k(x_i, \cdot)$.

$$\langle g, g \rangle = \sum_{i \in \mathcal{D}} \sum_{j \in \mathcal{D}} \alpha_i \alpha_j k(x_i, x_j) \geq 0$$

since $k(\cdot, \cdot)$ is a kernel and therefore is positive semidefinite. Linearity in both operands of $\langle \cdot, \cdot \rangle$ is implicit from the definition and finally symmetry of $\langle \cdot, \cdot \rangle$ follows from the symmetry of $k(\cdot, \cdot)$.

As a result, we may rephrase the objective as

$$\min_{\alpha \in \mathbb{R}^{|\mathcal{D}|}} \sum_{i \in \mathcal{D}} \left(y_i - \sum_{j \in \mathcal{D}} \alpha_j k(x_j, x_i) \right)^2 + \lambda \sum_{i \in \mathcal{D}} \left(\sum_{j \in \mathcal{D}} \alpha_j k(x_j, x_i) \right)^2$$

We have

$$\begin{aligned} \sum_{i \in \mathcal{D}} f(x_i)^2 &= \sum_{i \in \mathcal{D}} \left(\sum_{j \in \mathcal{D}} \alpha_j k(x_j, x_i) \right)^2 \\ &= \sum_{i \in \mathcal{D}} \left(\sum_{j \in \mathcal{D}} \sum_{k \in \mathcal{D}} \alpha_j \alpha_k k(x_j, x_i) k(x_k, x_i) \right) \end{aligned}$$

Finally, we note that the norm of f in the aforementioned inner product space is given by

$$\begin{aligned}\left\langle \sum_{i\in\mathcal{D}} \alpha_i k(x_i,\cdot), \sum_{i\in\mathcal{D}} \alpha_j k(x_j,\cdot) \right\rangle &= \sum_{i\in\mathcal{D}} \sum_{j\in\mathcal{D}} \alpha_i \alpha_j k(x_i,x_j) \\ &= \alpha^T G \alpha\end{aligned}$$

where $G = \Big[k(x_i,x_j)\Big]_{|\mathcal{D}|\times|\mathcal{D}|}$ and $\alpha = \begin{bmatrix}\alpha_1 & \cdots & \alpha_{|\mathcal{D}|}\end{bmatrix}^T$.