AIML - CS 337

Lecture 14: Kernel Methods - II

29th Sept 2022

Lecturer: Abir De *Scribe:*

1 Kernel for probability distributions

In the previous lecture we had seen that a Kernel can be defined as an inner product in the feature space as thus:

$$K(x, x') = \langle \phi(x), \phi(x') \rangle \tag{1}$$

Extending on this definition let $x \in A, x' \in B$, and x, x' be drawn from probability distributions P_1, P_2 respectively. We can define a Kernel K over the sets A,B as such:

$$K(A, B) = \int_{-\infty}^{\infty} \phi_A^T(x).\phi_B(x')Pr(x, x') \tag{2}$$

$$= \iint_{x \in A, x' \in B} \phi^T(x)\phi(x')dP(x, x') \tag{3}$$

Now, one possible measure of similitarity between the two sets A and B is $P(A \cap B) - P(A)P(B)$. To show that this indeed is a valid kernel, we can show that there exists some ϕ such that $K(A, B) = P(A \cap B) - P(A)P(B)$ and $K(A, B)$ satisfies Equation (3).

Consider $\phi_A(x) = \mathbb{I}_A(x) - P(A)$ where $\mathbb{I}_A(x)$ is the indicator function:

$$\mathbb{I}_A(x) = \begin{cases} 1 & x \in A \\ 0 & \text{otherwise} \end{cases}$$

Substituting in Equation (2) we get:

$$\begin{aligned} K(A, B) &= \int_{-\infty}^{\infty} (\mathbb{I}_A(x) - P(A))(\mathbb{I}_B(x') - P(B))Pr(x, x') \\ &= \int_{-\infty}^{\infty} (\mathbb{I}_A(x)\mathbb{I}_B(x'))Pr(x, x') - P(A) \int_{-\infty}^{\infty} (\mathbb{I}_B(x'))Pr(x, x') \\ &- P(B) \int_{-\infty}^{\infty} (\mathbb{I}_A(x'))Pr(x, x') + P(A)P(B) \int_{-\infty}^{\infty} Pr(x, x') \\ &= P(A \cap B) - P(A)P(B) - P(A)P(B) + P(A)P(B) \\ &= P(A \cap B) - P(A)P(B) \end{aligned}$$

2 Inner Product of Functions

We define inner product of two functions f,g as:

$$\langle f, g \rangle \doteq \int_{x,y} f(x)g(y)dP(x,y)$$

Under this inner product definition, K is Kernel for ϕ.
Properties of inner product:

1. Positive Semidefinite: $\langle f, f \rangle \ \geq 0$
2. Symmetric: $\langle f, g \rangle = \langle g, f \rangle$
3. Linearity: $< \langle c_1 f_1 + c_2 f_2, g \rangle = c_1 \langle f_1, g \rangle + c_2 \langle f_2, g \rangle$

We define norm of function using this inner product definition as:

$$\|f\| \doteq \sqrt{\langle f, f \rangle}$$

3 Finding similarity from loss

Until now, we assumed that Kernel was given to us and we used the kernel as a measure of similarity. But what if we have to find the Kernel if the loss is given.
Let F(w) be defined as follows:

$$F(w) = \sum_{i \in D} l(h_w(x_i), y_i)$$

Given this loss, we now need to find the similarity between points in Dataset.

$$\begin{bmatrix} Loss(x_i|t=0) \\ Loss(x_i|t=1) \\ . \\ . \\ . \end{bmatrix}$$

We make vectors of all x_i's and compare their similarity. We use this to select batches of data which have different training curve. If loss of two points is similar, we can not say that the points themselves are also similar. This is because weights are randomly initialized, so loss cannot be a good measure of similarity. We can instead use gradient of loss to find similarity.
If $X_i \sim X_j$ then $\nabla_w l(h_w(x_i), y_i) \approx \nabla_w l(h_w(x_j), y_j)$ but not the other way round.
Therefore, we can define the kernel as follows:

$$K(x_i, x_j) = E_w[\nabla_w^T l(h_w(x_i), y_i).\nabla_w l(h_w(x_j), y_j)]$$

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 $\alpha=\begin{bmatrix}\alpha_1 & \cdots & \alpha_{|\mathcal{D}|}\end{bmatrix}^T$.

5 Homework Problem

Problem. Show that the following kernel is positive semidefinite:

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

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):

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

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

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

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