Maximum Mean Discrepancy

Mengxi Wu (mengxiwu@usc.edu)

1 Kernel Baiscs

1.1 Motivation

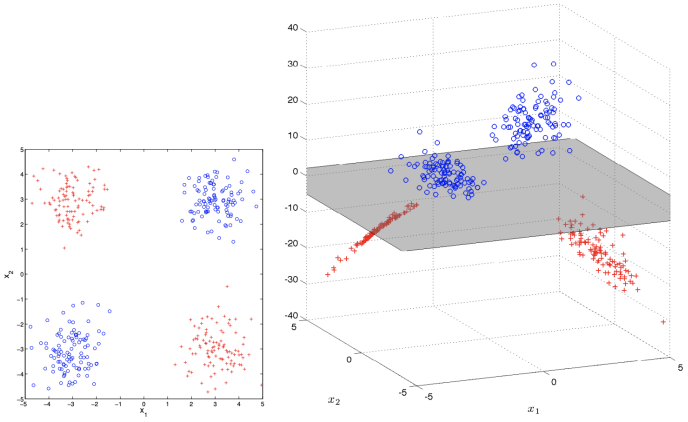

Figure 1: Example to illustrate motivation.

As shown in Figure 1, there is no linear classifier that can separate the red points from the blue points in $\mathbb{R}^2$. However, if we use a mapping function Φ to map the points to a higher dimensional feature space, $\mathbb{R}^3$, we obtain a linear decision boundary, the gray plane.

$$\Phi(x) = [x_1, x_2, x_1 x_2] \tag{1.1}$$

In above example, we can take $k(x, x') = \langle \Phi(x), \Phi(x') \rangle$ as a kernel. The term, kernel, refers to a dot product between features. Besides its ability to map features to different dimensional spaces, it can also control the smoothness of a function used in regression or classification.

1.2 Basic Definitions and Properties

Definition 1.1 (Inner Product) *Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ is said to ba an inner product on $\mathcal{H}$ if*
1. $\langle \alpha_1 f_1 + \alpha_2 f_2, g \rangle_{\mathcal{H}} = \alpha_1 \langle f_1, g \rangle_{\mathcal{H}} + \alpha_2 \langle f_2, g \rangle_{\mathcal{H}}$
2. $\langle f, g \rangle_{\mathcal{H}} = \langle g, f \rangle_{\mathcal{H}}$
3. $\langle f, f \rangle \geq 0$ and $\langle f, f \rangle_{\mathcal{H}} = 0$ if and only if $f = 0$

Definition 1.2 (Metric) *A metric is a nonnegative function $d(x, y)$ describing the "distance" between neighboring points for a given set. A metric should satisfy*
1.$d(x, y) = 0$ if and only if $x = y$
2.$d(x, y) = d(y, x)$
3.The triangle inequality, $d(x, y) + d(y, z) \geq d(x, z)$

Definition 1.3 (Metric Space) *A metric space is a set S with a global distance function (the metric d) that, for every two points x, y in S, gives the distance between them as a nonnegative real number*

$d(x, y)$. A metric space must also satisfy
1.$d(x, y) = 0$ if and only if $x = y$
2.$d(x, y) = d(y, x)$
3.The triangle inequality, $d(x, y) + d(y, z) \geq d(x, z)$

Definition 1.4 (Cauchy Sequence) *A sequence $\{s_n\}_{n=1}^{\infty}$ of elements in a normed space $\mathcal{H}$ with metric $d_{\mathcal{H}}$ is said to be a Cauchy sequence if for every $\epsilon > 0$, there exists $N \in \mathbb{N}$, such that for all $n, m \geq N$, $d_{\mathcal{H}}(s_n, s_m) < \epsilon$.*

Definition 1.5 (Complete Metric Space) *A complete metric space is a metric space in which every Cauchy sequence is convergent.*

Definition 1.6 (Hilbert Space) *A Hilbert space is a vector space $\mathcal{H}$ with an inner product $\langle f, g\rangle$, such that the norm defined by*

$$\|f\|_{\mathcal{H}} = \sqrt{\langle f, f\rangle_{\mathcal{H}}}$$

and this norm (the metric) makes $\mathcal{H}$ a complete metric space.

Definition 1.7 (Kernel) *Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a kernel if there exists an $\mathbb{R}$-Hilbert space and a map $\Phi : \mathcal{X} \to \mathcal{H}$ such that $\forall x, x' \in X$,*

$$k(x, x') = \langle \Phi(x), \Phi(x')\rangle_{\mathcal{H}}$$

Note that there is almost no condiction on $\mathcal{X}$. We allow no inner product defined on the $x \in \mathcal{X}$.

Lemma 1.1 (Sums of Kernels are Kernels) *Given $\alpha > 0$ and k, k_1 and k_2 all kernels on $\mathcal{X}$, then αk and $k_1 + k_2$ are kernels on $\mathcal{X}$.*

Lemma 1.2 (Products of Kernels are Kernels) *Given k_1 on $\mathcal{X}_1$ and k_2 on $\mathcal{X}_2$, then $k = k_1 \times k_2$ is a kernel on $\mathcal{X}_1 \times \mathcal{X}_2$. If $\mathcal{X} = \mathcal{X}_1 = \mathcal{X}_2$, then k is a kernel on $\mathcal{X}$.*

Lemma 1.3 (Kernels are Positive Definite) *Let $\mathcal{H}$ be any Hilbert space, $\mathcal{X}$ a non-empty set and $\Phi : \mathcal{X} \to \mathcal{H}$. Then $k(x, y) = \langle \Phi(x), \Phi(y)\rangle_{\mathcal{H}}$ is a positive definite function.*

2 The Reproducing Kernel Hilbert Space

In previous section, we introduce the kernels on feature spaces and the Hilbert space. The kernels defined in Hilbert space are positive definite. In this section, we introduce the functions determined by kernels on feature space $\mathcal{X}$. The space of such functions is known as a **reproducing kernel Hilbert space**.

Definition 2.1 (Reproducing Kernel Hilbert Space) *Let $\mathcal{H}$ be a Hilbert sapce of $\mathbb{R}$-valued functions defined on a non-empty set $\mathcal{X}$. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of $\mathcal{H}$, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if k satisfies*
1. $\forall x \in \mathcal{X}$, $k(\cdot, x) = \Phi(x)$ and $\Phi(x) \in \mathcal{H}$
2. $\forall x \in \mathcal{X}$, $\forall f \in \mathcal{H}$, $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (The Reproducing Property)

The reproducing property brings an important result. We define δ_x to be the operator of evaluation at x, i.e.

$$\delta_x f = f(x), \forall f \in \mathcal{H}, \forall x \in \mathcal{X} \tag{2.1}$$

We re-write the definition of the reproducing kernel hilbert space as follows.

Definition 2.2 (Reproducing Kernel Hilbert Space) *$\mathcal{H}$ is a reproducing kernel Hilbert space if for all $x \in \mathcal{X}$, the evaluation δ_x is bounded: there exists a corresponding $\lambda_x \geq 0$ such that $\forall f \in \mathcal{H}$,*

$$|f(x)| = |\delta_x f| \leq \lambda_x \|f\|_{\mathcal{H}}$$

This definition shows that when two functions are identical in the norm of a reproducing kernel Hilbert space, they are identical at every point (Eq.2.2).

$$|f(x) - g(x)| = |\delta_x(f - g)| \leq \lambda_x \|f - g\|, \forall f, g \in \mathcal{H} \tag{2.2}$$

Distance between means in feature space. Suppose we have two distributions p_1 and p_2. We sample $\{x_i\}_{i=1}^m$ from p_1 and $\{y_i\}_{i=1}^n$ from p_2. The following is the derivation of computing distance of means of these two distributions in feature space.

$$
\begin{aligned}
&\|\frac{1}{m}\sum_{i=1}^{m}\Phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\Phi(y_j)\|_{\mathcal{H}}^2 \\
&= \langle \frac{1}{m}\sum_{i=1}^{m}\Phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\Phi(y_j), \frac{1}{m}\sum_{i=1}^{m}\Phi(x_i) - \frac{1}{n}\sum_{j=1}^{n}\Phi(y_j)\rangle_{\mathcal{H}} \\
&= \frac{1}{m^2}\langle \sum_{i=1}^{m}\Phi(x_i), \sum_{i=1}^{m}\Phi(x_i)\rangle_{\mathcal{H}} + ... \\
&= \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m}k(x_i, x_j) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(y_i, y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}k(x_i, y_j)
\end{aligned}
$$

When $\Phi(x) = x$, we can distinguish distributions with different means. When $\Phi(x) = [x, x^2]$ we can distinguish distributions with different means, different variances, or both. More complex feature spaces allow us to distinguish more complex features of the distributions. One thing worth noting is that there exist kernels that can distinguish any two distributions.

3 Maximum Mean Dsicrepancy

In general, MMD is a dffierence between distributions by computing the distance of the mean embeddings of features. Suppose we have two distributions p_1 and p_2 over a feature space $\mathcal{X}$. The MMD is defined by a feature map $\Phi : \mathcal{X} \to \mathcal{H}$. $\mathcal{H}$ is a reproducing kernel Hilbert space.

$$
\text{MMD}(p_1, p_2) = \|E_{x\sim p_1}[\Phi(x)] - E_{y\sim p_2}[\Phi(y)]\|_{\mathcal{H}} \tag{3.1}
$$

The "mean" term in MMD is shown as we take the expectation. How does the "maximum" term come? Here we need to introduce a fact in Hilbert space (Eq.3.2).

$$
\sup_{f:\|f\|\leq 1}\langle f, g\rangle_{\mathcal{H}} = \|g\|, \text{ when } f = \frac{g}{\|g\|} \tag{3.2}
$$

Then, we can write

$$
\begin{aligned}
\text{MMD}(p_1, p_2) &= \|E_{x\sim p_1}[\Phi(x)] - E_{y\sim p_2}[\Phi(y)]\|_{\mathcal{H}} \\
&= \sup_{f:\|f\|\leq 1}\langle f, E_{x\sim p_1}[\Phi(x)] - E_{y\sim p_2}[\Phi(y)]\rangle_{\mathcal{H}} \\
&= \sup_{f:\|f\|\leq 1}E_{x\sim p_1}[\langle f, \Phi(x)\rangle_{\mathcal{H}}] - E_{y\sim p_2}[\langle f, \Phi(y)\rangle_{\mathcal{H}}] \\
&= \sup_{f:\|f\|\leq 1}E_{x\sim p_1}[f(x)] - E_{y\sim p_2}[f(y)]
\end{aligned}
$$

This last equation above explains the "maximum" term. It is the maximum over function f in a unit ball of $\mathcal{H}$.

References

[1] Cauchy Sequence. link.

[2] Maximum Mean Discrepancy. link.

[3] Metric. link.

[4] Metric Space. link.

[5] Arthur Gretton. Introduction to RKHS, and some simple kernel algorithms. link.