1 Support Vector Machines

Support vector machine is a kernelized optimal margin linear classifier (original paper [1] and a nice summary [2]). It distinguishes itself from classifiers minimizing empirical risk as it favors classifier which makes confident predictions. For binary classification problem $\mathcal{Y} = \{-1, +1\}$, we are interested in finding a linear decision bondary, parameterized by $w \in \mathbb{R}^d, b \in \mathbb{R}$, that separates the training data points by maximizing the worst case distance (margin) of each data point to the decision boundary. We first assume that training set can be linearly separated. Given dataset $\{(x_i, y_i)\}_{i=1}^n$, we are interested in solving the following quadratic programming problem,

$$\begin{aligned} &\min_{w,b} \frac{1}{2} \|w\|^2 \\ &\text{s.t. } y_i(w^T x_i + b) \geq 1 \quad i = 1, 2, \cdots, n \end{aligned}$$

To derive the dual problem, we write the Lagrangian,

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} \|w\|^2 + \sum_{i=1}^n \alpha_i \left[1 - y_i(w^T x_i + b)\right] \tag{1}$$

where $\alpha = \{\alpha_i\}_{i=1}^n$ are the dual variables. Solve for $\inf_{w,b} \mathcal{L}(w, b, \alpha)$ to arrive at the dual objective. In particular, first order optimality condition gives $w = \sum_{i=1}^n \alpha_i y_i x_i$ and it must be that $0 = \sum_{i=1}^n \alpha_i y_i$. Therefore,

$$\begin{aligned} &\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j \\ &\text{s.t. } \alpha_i \geq 0 \quad i = 1, 2, \cdots, n && \text{(dual feasibility)} \\ &\sum_{i=1}^n \alpha_i y_i = 0 && \text{(from } \nabla_b \mathcal{L} = 0\text{)} \end{aligned}$$

The dual can be solved more efficiently than the primal problem using coordinate descent. The decision rule is linear w.r.t support vectors (those x_i right on margin with $\alpha_i > 0$)

$$\hat{y}(x) = \text{sgn}\left(\sum_{i=1}^n \alpha_i y_i x_i^T x + b\right) \quad \text{for} \quad b = y_i - \sum_{j=1}^n \alpha_j y_j x_j^T x_i \tag{2}$$

for any support vector x_i. We observe that optimization as well as prediction uses input vectors via dot products only. We are motivated to use feature mapping $\phi : \mathcal{X} \to \mathcal{H}$ to map input vectors to a higher dimensional possibly infinite feature space in hope that the lifted space is linearly separable. The kernel function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ allows us to compute dot products $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$ efficiently and represents a notion of similarity over two instance of aribtrary objects, e.g. vectors in $\mathbb{R}^n$, graphs, texts. We can substitute k whenever inner product is used and arrive at a optimal margin classifier over implicitly defined nonlinear feature mapping ϕ. In case when training dataset is not linearly separable, we can introduce slack variable $\{\xi_i\}_{i=1}^n$ where $x_i \geq 0$ to relax the inequality constraints and penalize misclassified or within margin points with $C \sum_{i=1}^n \xi_i$ for some $C \in \mathbb{R}$. In this case, we have the following Lagrangian,

$$\mathcal{L}(w, \xi, \alpha, \beta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha \left(1 - y_i(\langle w, \phi(x_i) \rangle_{\mathcal{H}} + b) - \xi_i\right) + \sum_{i=1}^n \beta_i \xi_i \tag{3}$$

First order condition $0 = \frac{\partial}{\partial \xi_i}\mathcal{L} = C - \alpha_i + \beta_i$ together with dual feasibility $\beta_i \geq 0$ yield $\alpha_i \leq C$ for all $i = 1, 2, \cdots, n$. Therefore, we optimize for the following dual problem,

$$\begin{aligned} \min_{\alpha} \ & \frac{1}{2}\alpha^T Q \alpha - 1^T \alpha \\ \text{s.t. } & y^T \alpha = 0 \\ & 0 \leq \alpha_i \leq C \quad i = 1, 2, \cdots, n \end{aligned}$$

where $Q \in \mathbb{R}^{n \times n}$ with $Q_{ij} = \alpha_i \alpha_j k(x_i, x_j)$, with optimal decision rule as

$$\hat{y}(x) = \text{sgn}\left(\sum_{i=1}^{n} \alpha_i y_i k(x_i, x) + b\right) \quad \text{for} \quad b = y_i - \sum_{j=1}^{n} \alpha_j y_j k(x_j, x_i) \tag{4}$$

for any support vector i, i.e. $0 < \alpha_i < C$.

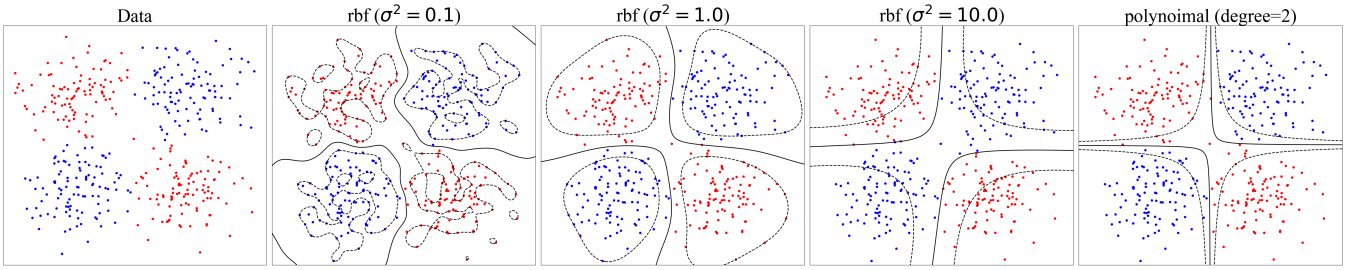

Figure 1: SVM on simulated 2D Gaussian with degree two polynomial kernel and radial basis function kernel with varying bandwidth. Larger bandwidth corresponds to smoother decision boundary and in the limit approaches the decision boundary of a linear kernel

2 Reproducing Kernel Hilbert Space

[3] provides a rigorous introduction to RKHS while [4] provides pretty good intuitions. Intuitively, RKHS can be considered as (1) a space of well-behaved functions whose smoothness is determined by its kernel. This view is useful when trying to think about regularization in terms of $\|f\|_{\mathcal{H}}^2$ (2) an inner product space for features. This view is useful when trying to apply kernel trick. (3) a space of functions spanned by representors $\{k(x, \cdot)\}_{x \in \mathcal{X}}$. We think of functions in RKHS as a simple function class, i.e. linear with respect to $\phi(x)$, and also flexible due to various choice of kernels.

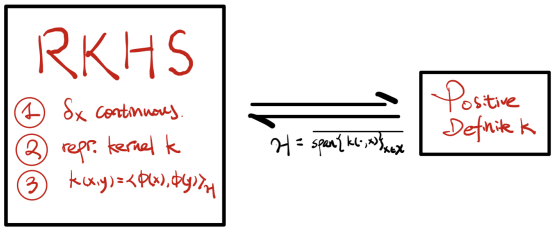

Figure 2: Equivalent views of RKHS

Definition 1. *(Hilbert space) A Hilbert space is a complete inner product space, i.e.* $(\mathcal{H}, \langle \cdot, \cdot \rangle_{\mathcal{H}})$

Definition 2. *(Reproducing kernel Hilbert space) A Hilbert space* $\mathcal{H}$ *of functions* $f : \mathcal{X} \to \mathbb{R}$ *is a RKHS if its evaluation functional,* $\delta_x : \mathcal{H} \to \mathbb{R}$ *where* $\delta_x(f) = f(x)$ *is continuous* $\forall\, x \in \mathcal{X}$.

Intuitively, RKHS is a space of well-behaved functions. In particular, norm convergence in $\mathcal{H}$ yield pointwise convergence, i.e. $\lim_{n\to\infty} \|f_n - f\| = 0$ implies $f_n \to f$. Note evaluation functional is linear. The condition that δ_x is continuous is equivalent to δ_x be bounded [3]. Regularization of the form $\|f\|_{\mathcal{H}}$ leads to regularization on function values.

Definition 3. *(Reproducing kernel) Let* $\mathcal{H}$ *be Hilbert space defined on non-empty* $\mathcal{X}$*, then a function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is called a reproducing kernel of* $\mathcal{H}$ *if*

1. $\forall\, x \in \mathcal{X}$, $k(x, \cdot) \in \mathcal{H}$
2. $\forall\, x \in \mathcal{X}$, $\forall\, f \in \mathcal{H}$, $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$ *(reproducing property)*

In particular, for any $x, y \in \mathcal{X}$, $k(x, y) = \langle k(y, \cdot), k(y, \cdot) \rangle_{\mathcal{H}}$.

Intuitively, the definition of reproducing kernel implies that $k(x, \cdot)$ is (1) a high dimensional representer of x, and (2) as a representer of evaluation for any function in $\mathcal{H}$ on data point x. We also note that we can always find a feature map associated with a reproducing kernel, namely the canonical feature map $\phi : x \to k(x, \cdot)$, and represent k as an inner product in feature space $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. It turns out that RKHS has $\mathcal{H}$ is a RKHS if and only if $\mathcal{H}$ has a reproducing kernel. Definition of RKHS based on continuous evaluation functional is equivalent to existence of a (unique) reproducing kernel.

Definition 4. *(Positive definite functions) A symmetric function* $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is positive definite if for all* $n \geq 1$ *and for all* $(a_1, \cdots, a_n) \in \mathbb{R}^n$ *for all* $(x_1, \cdots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j h(x_i, x_j) \geq 0 \tag{5}$$

If you consider a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, k is positive definite if any kernel matrix $K \in \mathbb{R}^{n \times n}$ where $K_{ij} = k(x_i, x_j)$ is positive definite, i.e. $K \succeq 0$.

Definition 5. *(Kernel defined via feature map) Let* $\mathcal{X}$ *be non-empty set. The function* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is a kernel if there exists a real Hilbert space* $\mathcal{H}$ *and a map* $\phi : \mathcal{X} \to \mathcal{H}$ *such that* $\forall\, x, y \in \mathcal{X}$, $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$. ϕ *is said to be the feature map and* $\mathcal{H}$ *be the feature space.*

Note there maybe more than one feature map yield for any one kernel. Intuitively, a kernel is a function that can be represented as inner product. It turns out that RKHS with reproducing kernel k is positive definite,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \langle k(x_i, \cdot), k(x_j, \cdot) \rangle = \left\langle \sum_{i=1}^{n} a_i \phi(x_i), \sum_{i=1}^{n} a_i \phi(x_i) \right\rangle_{\mathcal{H}} \geq 0 \tag{6}$$

and, conversely, we can show that for every positive definite function k there is an unqiue RKHS whose reproducing kernel is k (Moore-Aronsajn). Intuitively, we construct a RKHS $\mathcal{H} = \overline{\mathcal{H}_0}$, the completion of a pre-RKHS space $\mathcal{H}_0 = \text{span}(\{k(x, \cdot)\}_{x \in \mathcal{X}})$. In particular, the choice of inner product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_i \beta_j k(x_i, y_j) \tag{7}$$

for $f = \sum_i \alpha_i k(x_i, \cdot), g = \sum_j \beta_j k(y_j, \cdot)$ makes $\mathcal{H}_0$ a valid pre-RKHS. Intuitively, this construction implies that RKHS is a space spanned by representers $\{k(x, \cdot)\}_{x \in \mathcal{X}}$.

3 Kernel Mean Embedding of Distributions

[5] chapter 3 provides a really clear generalization from feature map over points $x \in \mathcal{X}$ to measures over the measurable space $(\mathcal{X}, \mathcal{F})$ where $\mathcal{F}$ is sigma algebra of $\mathcal{X}$. Let $\mathcal{H}$ be a reproducing kernel Hilbert space with reproducing kernel k. Consider the mean map $\mu : \mathcal{P} \to \mathcal{H}$

$$\mu(\mathbb{P}) = \int k(x, \cdot)\mathbb{P}(x) \tag{8}$$

and write $\mu_{\mathbb{P}} := \mu(\mathbb{P})$. Intuitively, the mean map $\mu_{\mathbb{P}}$ is a representation of probability measure $\mathbb{P}$ in $\mathcal{H}$. Not rigorously, we noticed that we can compute $\mathbb{E}_{\mathbb{P}}[f(X)]$ as inner products in $\mathcal{H}$,

$$\int f\mathbb{P} = \int \langle f, k(x, \cdot)\rangle_{\mathcal{H}} \, d\mathbb{P}(x) = \left\langle f, \int k(x, \cdot) d\mathbb{P}(x) \right\rangle_{\mathcal{H}} = \langle f, \mu_{\mathbb{P}} \rangle$$

Formally,

Lemma 1. *If* $\mathbb{E}\left[\sqrt{k(X, X)}\right] < \infty$*, then* $\mu_{\mathbb{P}} \in \mathcal{H}$ *and* $\mathbb{E}_{\mathbb{P}}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$ *for all* $f \in \mathcal{H}$*.*

Proof. For any $\mathbb{P} \in \mathcal{P}$, define a linear functional $T_{\mathbb{P}} : \mathcal{H} \to \mathbb{R}$ where $T_{\mathbb{P}}(f) = \mathbb{E}_{\mathbb{P}}[f(X)]$ as the operation of taking expectation with respect to function $f \in \mathcal{H}$ over $\mathbb{P}$. We show $T_{\mathbb{P}}$ is bounded (to use Rietz)

$$|T_{\mathbb{P}} f| = |\mathbb{E}[f(X)]| \leq \mathbb{E}[|f(X)|] = \mathbb{E}[|\langle f, k(X, \cdot)\rangle|] \leq \mathbb{E}\left[\sqrt{k(X, X)}\, \|f\|_{\mathcal{H}}\right] < \infty$$

By Rietz representation theorem, exists $g \in \mathcal{H}$ such that $T_{\mathbb{P}} f = \langle f, g \rangle_{\mathcal{H}}$. The choice of $f = k(x, \cdot)$ implies.

$$g(x) = \langle k(x, \cdot), g\rangle_{\mathcal{H}} = T_{\mathbb{P}}[k(x, \cdot)] = \int k(x, x') d\mathbb{P}(x') = \mu_{\mathbb{P}}(x)$$

Therefore, $\mu_{\mathbb{P}} \in \mathcal{H}$ is guaranteed by Rietz and we can re-write the Rietz's result and considered it as a reproducing property for $T_{\mathbb{P}}$, i.e. $\mathbb{E}[f(X)] = \langle f, \mu_{\mathbb{P}} \rangle_{\mathcal{H}}$. □

In effect, what this lemma states is under some mild assumptions, the mean map $\mu_{\mathbb{P}}$ ends up in RKHS $\mathcal{H}$ and that we can compute the expectation of any function in RKHS with respect to distribution $\mathbb{P}$ by computing inner product between the function f and the mean map $\mu_{\mathbb{P}}$ in $\mathcal{H}$. This is analogous to previous definition of a reproducing kernel, that the cannical feature map ends up in RKHS $k(x, \cdot) \in \mathcal{H}$ and we can evaluate the evaluation functional as an inner product in RKHS, i.e. $\delta_x(f) = \langle f, k(x, \cdot)\rangle_{\mathcal{H}}$.

Furthermore, we note that inner products of mean maps is expectation of kernel, i.e. for $X \sim \mathbb{P}, Y \sim \mathbb{Q}$,

$$\langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \mathbb{E}[\mu_{\mathbb{Q}}(X)] = \mathbb{E}[k(X, Y)] \tag{9}$$

where we have used fact that evaluating the mean map at $x \in \mathcal{X}$ computes the expectation of a kernel,

$$\mu_{\mathbb{Q}}(x) = \langle \mu_{\mathbb{Q}}, k(x, \cdot)\rangle_{\mathcal{H}} = \mathbb{E}[k(x, Y)] \tag{10}$$

For appropriately chosen kernel, specifically a characteristic kernel, the mean map $\mu_{\mathbb{P}}$ completely characterizes a distribution. For example, the choice of $k(x, x') = e^{\langle x, x' \rangle}$ implies that the mean embedding is the moment generating function $\mu_{\mathbb{P}} = \mathbb{E}\left[e^{\langle X, \cdot \rangle}\right]$.

Definition 6. *(Characteristic kernel) A kernel is characteristic if the map* $\mu : \mathbb{P} \mapsto \mu_{\mathbb{P}}$ *is injective*

A characteristic kernel ensures that the induced RKHS is rich enough to represent higher order momemts of $\mathbb{P}$. In particular, it ensures that $\|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} = 0$ if and only if $\mathbb{P} = \mathbb{Q}$. As an example, Gaussian and Laplacian kernels are characteristic. Characteristic kernels are important when trying to distinguish distributions, might be less useful when trying to do predictive tasks with distributional data.

Definition 7. *(Universal kernsl) A continuous positive definite kernel on compact metric space $\mathcal{X}$ is universal if the corresponding RKHS $\mathcal{H}$ is dense in $C(\mathcal{X})$, space of bounded continous functions over $\mathcal{X}$.*

Universal kernel are characteristic kernels. See [5] for a classification of kernels.

We can derive empirical estimate of the mean map $\hat{\mu}_{\mathbb{P}} = \frac{1}{m}\sum_{i=1}^{m} k(\cdot, x_i)$ for $x_i \overset{iid}{\sim} \mathbb{P}$.

4 Maximm Mean Discrepancy

The kernel mean embedding defines a natural metric for probability distributions, so called maximum mean discrepancy, as $\text{MMD}(\mathbb{P}, \mathbb{Q}) = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$. In particular, evaluation can be deligated to kernels,

$$\begin{aligned} \text{MMD}^2(\mathbb{P}, \mathbb{Q}) &= \langle \mu_{\mathbb{P}}, \mu_{\mathbb{P}} \rangle_{\mathcal{H}} + \langle \mu_{\mathbb{Q}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} - 2 \langle \mu_{\mathbb{P}}, \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} \quad (11) \\ &= \mathbb{E}\left[k(X, X')\right] + \mathbb{E}\left[k(Y, Y')\right] - 2\mathbb{E}\left[k(X, Y)\right] \quad (12) \end{aligned}$$

where $X, X' \overset{iid}{\sim} \mathbb{P}$ and $Y, Y' \overset{iid}{\sim} \mathbb{Q}$. The unbiased estimate is given by

$$\widehat{\text{MMD}}_u^2(X, Y) = \frac{1}{m(m-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{n(n-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{mn} \sum_{i,j} k(x_i, y_j) \quad (13)$$

where $\mathbf{X} = (X_1, \cdots, X_m) \overset{iid}{\sim} \mathbb{P}$ and $\mathbf{Y} = (Y_1, \cdots, Y_m) \overset{iid}{\sim} \mathbb{Q}$. Alternatively, we can consider maximum mean discrepancy as a class of integral probability metric over functions in the unit ball in RKHS $\mathcal{H}$,

$$\text{MMD}(\mathbb{P}, \mathbb{Q}) = \sup_{f \in \mathcal{H}:\, \|f\|_{\mathcal{H}} \leq 1} \left\{ \int f d\mathbb{P} - \int f d\mathbb{Q} \right\} \quad (14)$$

which is equivalent to the view of MMD as distance between kernel mean embeddings

$$\sup_{f \in \mathcal{H}:\, \|f\|_{\mathcal{H}} \leq 1} \left\{ \int f d\mathbb{P} - \int f d\mathbb{Q} \right\} = \sup_{f \in \mathcal{H}:\, \|f\|_{\mathcal{H}} \leq 1} \langle f, \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} \rangle_{\mathcal{H}} = \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}} \quad (15)$$

where optimal witness function is normalized feature mean $f^* = \mu_{\mathbb{P}} - \mu_{\mathbb{Q}} / \|\mu_{\mathbb{P}} - \mu_{\mathbb{Q}}\|_{\mathcal{H}}$.

References

[1] Bernhard Boser, Isabelle Guyon, and Vladimir Vapnik. "A Training Algorithm for Optimal Margin Classifiers". In: (1992).

[2] Jean-Philippe Vert, Koji Tsuda, and Bernhard Scholkopf. "A primer on kernel methods". In: (2004), p. 42.

[3] Arthur Gretton. "What is an RKHS?" In: (2012). URL: http://www.stats.ox.ac.uk/~sejdinov/teaching/atml14/Theory_2014.pdf.

[4] Arthur Gretton. "Introduction to RKHS, and some simple kernel algorithms". In: (2019), p. 33. URL: http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf.

[5] Krikamol Muandet et al. "Kernel Mean Embedding of Distributions: A Review and Beyond". In: *Foundations and Trends® in Machine Learning* 10.1 (2017), pp. 1–141. ISSN: 1935-8237, 1935-8245. DOI: 10.1561/2200000060. arXiv: 1605.09522. URL: http://arxiv.org/abs/1605.09522 (visited on 12/25/2020).