Machine Learning and Neural Networks (MATH3431) Epiphany term, 2023

# Draft Handout 7: Kernel methods

Lecturer & author: Georgios P. Karagiannis georgios.karagiannis@durham.ac.uk

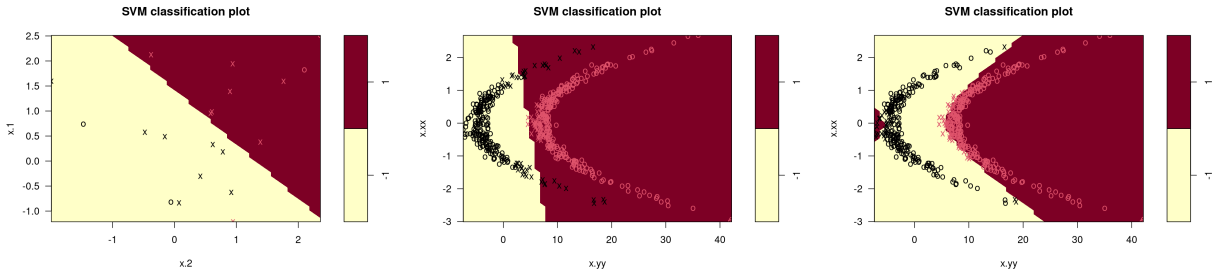
**Aim.** To introduce the ideas of learning machines by introducing data into high-dimensional feature spaces for computational and accuracy gains; introduce the kernel trick, kernel functions, extend SVM into the kernel framework.

**Reading list & references:**

(1) Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
- Ch. 16.2 Support Vector Machine

(2) Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, No. 4, p. 738). New York: Springer.
- Ch. 6 Kernel methods

## 1. Intro and motivation




(A) Linearly separable data set, with linear separator

(B) Non-linearly separable data set, with linear separator

(C) Non-linearly separable data set, with quadratic separator

Figure 1.1. Soft SVM from Computer practical 4

## 2. Improving expressive power

*Note* 1. To make the class of hypotheses more expressive with purpose to improve accuracy, we can first map the original instance space $x \in \mathcal{X}$ into another space (possibly of a higher dimension) and then learn a hypothesis in that space.

**Definition 2.** A Hilbert space is a vector space, with an inner product, which is also complete[1].

[1]Going deeper to completeness is out of the scope.

**Lemma 3.** *If $\mathcal{X}$ is a linear subspace of a Hilbert space, then every $x \in \mathcal{X}$ can be written as $x = u+v$ where $u \in \mathcal{X}$ and $\langle v, w \rangle = 0$ for all $w \in \mathcal{X}$.*

*Summary* 4. The basic paradigm is as follows:

(1) Choose a mapping $\psi : \mathcal{X} \to \mathcal{F}$ with $\psi(x) := (\psi_1(x), ..., \psi_n(x))$ for some feature space $\mathcal{F}$.
(2) Create the image sequence $\tilde{\mathcal{S}} = \left\{ z_i^{\psi} = (\psi(x_i), y_i) \right\}_{i=1}^{m}$ from the original training set $\mathcal{S}$.
(3) Train a linear predictor $h$ against $\tilde{\mathcal{S}}$.
(4) Predict the label or the output of a new point $x^{\text{new}}$ by $h^{\psi}(x^{\text{new}}) := h \circ \psi(x^{\text{new}})$

*Note* 5. Feature space $\mathcal{F}$ can be a Hilbert space, e.g a Euclidean space such as $\mathcal{F} \subseteq \mathbb{R}^n$ for some $n$. That includes infinite dimensional spaces.

*Note* 6. The introduction of mapping $\psi : \mathcal{X} \to \mathcal{F}$ induces

(1) probability distribution $G^{\psi}$ over domain $\mathcal{X} \times \mathcal{F}$ with $G^{\psi}(A) = G\left(\psi^{-1}(A)\right)$ for every set $A \subseteq \mathcal{X} \times \mathcal{F}$.
(2) predictive rule $h^{\psi}(\cdot) := h \circ \psi(\cdot)$, where $h \circ \psi(\cdot) = h(\psi(\cdot))$
(3) risk function $R_{G^{\psi}}(h) := R_G(h \circ \psi)$, as

$$R_G(h \circ \psi) = \int \ell(h \circ \psi, z = (x, y)) \, \mathrm{d}G(z) = \int \ell\left(h, z^{\psi}\right) \mathrm{d}G^{\psi}(x, y) = R_{G^{\psi}}(h)$$

*Note* 7. Any feature mapping $\psi$ that maps the original instances $\mathcal{X}$ into some Hilbert space $\mathcal{F}$ can be used. The Euclidean space $\mathcal{F} \subseteq \mathbb{R}^n$ for some $n$ is a Hilbert space. There are also infinite dimensional Hilbert spaces.

*Note* 8. The success of this learning paradigm in Summary 4 depends on choosing a good $\psi$ for a given learning task. Eg, in SVM, $\psi$ will make the image of the data distribution (close to being) linearly separable in the feature space $\mathcal{F}$, thus making the resulting learning algorithm a good learner for a given task. This requires prior knowledge of the problem. However, some popular choices are the polynomial kernel and the Gaussian (or Radial Basis Functions) kernels.

*Note* 9. Using a $\psi(x) := (\psi_1(x), ..., \psi_n(x))$ that is high dimensional ($n$ is too large) may improve accuracy (expressiveness) of the learner (e.g. recall in polynomial regression increasing the polynomial degree). However this increases the computational effort/cost required to perform calculations to minimize the associated risk function in the high dimensional space, as well as we need more data.

## 3. The Kernel trick

*Note* 10. Kernel function $K$ is defined as $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with $K(x, x') = \langle \psi(x), \psi(x') \rangle$ given an embedding $\psi(x)$ of some domain space $\mathcal{X}$ into some Hilbert space $\mathcal{F}$. Hence, Kernel functions are here are used to describe inner products in the feature space. $K$ can be considered as specifying similarity between instances and of the embedding $\psi(\cdot)$ as mapping the domain set $\mathcal{X}$ into a space where these similarities are realized as inner products.

**Problem 11.** (General learning problem) Consider a prediction rule $h : \mathcal{X} \to \mathcal{Y}$ with $h_w(x) = \langle w, \psi(x) \rangle$ which is trained against a training sample $\{z_i = (x_i, y_i)\}_{i=1}^m$ with the following general optimization problem

$$\underset{w}{\text{minimize}} \left( f(\langle w, \psi(x_1) \rangle, ..., \langle w, \psi(x_m) \rangle) + R(\|w\|) \right), \tag{3.1}$$

where $f : \mathbb{R}^m \to \mathbb{R}$ is an arbitrary function and $R : \mathbb{R}_- \to \mathbb{R}$ is a monotonically non-decreasing function.

**Example 12.** In Soft SVM (Proposition 24), it is $f(a_1, ..., a_m) = \frac{1}{m} \sum_{i=1}^m \max\{1 - y_i a_i\}$ and $R(a) = \lambda a^2$.

*Note* 13. The following result states a duality of a general learning problem 11 to be able to be indirectly extended to a possibly high dimensional feature space with purpose to improve the expressiveness (accuracy of the predictive rule) by using kernel ideas mitigating the associated computational cost and data requirements.

**Theorem 14.** *(Representation theorem) Assume mapping $\psi : \mathcal{X} \to \mathcal{F}$ where $\mathcal{F}$ is a Hilbert space. There exists a vector $\alpha \in \mathbb{R}^m$ such that $w = \sum_{i=1}^m \alpha_i \psi(x_i)$ is the optimal solution of (3.1) in Problem 11.*

*Proof.* Let $w^*$ be the optimal solution of (3.1). Because $w^*$ is element of Hilbert space, it can be written as $w^* = \sum_{i=1}^m \alpha_i \psi(x_i) + u$ where $\langle u, \psi(x_i) \rangle = 0$ for all $i = 1, ..., m$. Set $w := w^* - u$.

Because $\|w^*\|^2 = \|w\|^2 + \|u\|^2$ it is $\|w\| \leq \|w^*\|$ implying that

$$R(\|w\|) \leq R(\|w^*\|).$$

Because $\langle w, \psi(x_i) \rangle = \langle w^* - u, \psi(x_i) \rangle = \langle w^*, \psi(x_i) \rangle$ for all $i = 1, ..., m$, it is

$$f(\langle w, \psi(x_1) \rangle, ..., \langle w, \psi(x_m) \rangle) = f(\langle w^*, \psi(x_1) \rangle, ..., \langle w^*, \psi(x_m) \rangle)$$

Then the objective function of 3.1 at $w$ is less than or equal to that of the minimizer $w^*$ implying that $w = \sum_{i=1}^m \alpha_i \psi(x_i)$ is an optimal solution. □

*Note* 15. Let $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a kernel function with $K(x, x') = \langle \psi(x), \psi(x') \rangle$. According to the representation Theorem 14, the general learning problem 11, can be equivalently addressed by re-writing the learning predictive rule as

$$h_\alpha(x) = \sum_{i=1}^m \alpha_i K(x_i, x)$$

and learning $\{\alpha_i\}$ as the solutions of

$$\underset{\alpha}{\text{minimize}} \left( f\left( \sum_{i=1}^m \alpha_i K(x_i, x), ..., \sum_{i=1}^m \alpha_i K(x_m, x) \right) + R\left( \sqrt{\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j K(x_i, x_j)} \right) \right), \tag{3.2}$$

This is because

$$\langle w, \psi(x_j) \rangle = \langle \sum_{i=1}^m \alpha_i \psi(x_i), \psi(x_j) \rangle = \sum_{i=1}^m \alpha_i \langle \psi(x_i), \psi(x_j) \rangle = \sum_{i=1}^m \alpha_i K(x_i, x_j)$$

and

$$\|w\|^2 = \langle \sum_{i=1}^{m} \alpha_i \psi(x_i), \sum_{j=1}^{m} \alpha_j \psi(x_j) \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j \langle \psi(x_i), \psi(x_j) \rangle = \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j K(x_i, x_j)$$

*Note* 16. In learning problem 11, direct access to elements $\psi(\cdot)$ in the feature space is not necessary, as equivalently one can calculate or specify the kernel function (that is inner products in the feature space).

**Definition 17.** Gram matrix is called the $m \times m$ matrix $G$ s.t. $[G]_{i,j} = K(x_i, x_j)$.

**Example 18.** In Soft SVM, the predictive rule is $h(x) = \text{sign}\left(\sum_{i=1}^{m} \alpha_i K(x_i, x)\right)$ where $\alpha$ is learned by

$$\underset{\alpha}{\text{minimize}} \left( \lambda \alpha^\top G \alpha + \frac{1}{m} \sum_{i=1}^{m} \max(0, 1 - y_i [G\alpha]_i) \right)$$

that can be solved as quadratic programming.

**Example 19.** (Polynomial Kernels) Let $x \in \mathcal{X} \subseteq \mathbb{R}^n$. Assume we want to extend the linear mapping $x \mapsto \langle w, x \rangle$ to the $k$ degree polynomial mapping $x \mapsto h(x)$. The multivariate polynomial can be written as $h(x) = \langle w, \psi(x) \rangle$, where $\psi : \mathbb{R}^n \to \mathbb{R}^d$ with $\psi(x)$ is a vector of elements $\psi_J(x) = \prod_{i=1}^{r} x_{J_i}$ for $J \in \{1, ..., n\}^r$ and $r \leq k$. This learning problem can be equivalently be addressed with the $k$ degree polynomial kernel

$$K(x, x') = \left(1 + \langle x, x' \rangle\right)^k$$

**Solution.** Set $x_0 = x'_0 = 1$. Then

$$\begin{aligned} K(x, x') = \left(1 + \langle x, x' \rangle\right)^k &= \left( \sum_{j=0}^{n} x_j x'_j \right)^k \\ &= \sum_{J \in \{1,...,n\}^k} \prod_{j=0}^{n} x_{J_i} x'_{J_i} = \sum_{J \in \{1,...,n\}^k} \left( \prod_{j=0}^{n} x_{J_i} \right) \left( \prod_{j=0}^{n} x'_{J_i} \right) \\ &= \langle \psi(x), \psi(x') \rangle \end{aligned}$$

where $\psi(x)$ is as defined.

## 4. Construction of kernels

*Note* 20. The suitability of any hypothesis class to a given learning task depends on the nature of that task. Feature mapping $\psi$ can be considered as expanding the class of linear classifiers to a richer class (corresponding to linear classifiers over the feature space). Embedding $\psi$ is a way to express and utilize prior knowledge about the problem at hand.

*Note* 21. Hence, according to the Representation theorem 14, specifying a kernel function is a way to express prior knowledge about the problem with purpose to enrich the hypothesis class and learn a more accurate predictor.

*Claim* 22. If we believe that positive examples can be distinguished by some ellipse, we can define $\psi$ to be all the monomials up to order 2, or use a degree 2 polynomial kernel.

*Note* 23. The question is if the specified function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ (by the practitioner) is indeed a kernel function; namely if $K$ can be written as inner product $K(x, x') = \langle \psi(x), \psi(x') \rangle$ of feature functions $\psi(x)$. Theorem 25 provides sufficient and necessary conditions to check that.

**Definition 24.** A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semi-definite if its Gram matrix $G$ , $[G]_{i,j} = K(x_i, x_j)$ , is a positive semi-definite matrix.

**Theorem 25.** *A symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ implements an inner product in some Hilbert space if and only if it is positive semi-definite i.e. its Gram matrix $G$ , $[G]_{i,j} = K(x_i, x_j)$ , is a positive semi-definite matrix.*

*Proof.* Assume $K$ implements an inner product $K(x, x') = \langle \psi(x), \psi(x') \rangle$ in some Hilbert space. Let $G$ be its Gram matrix with $G = \Psi^\top \Psi$ and $\psi(x_i)$ is the $i$-th column of $\Psi$. You may consider that $\cdot^\dagger$ represents the transpose for simplicity. For any $\xi \in \mathbb{R}^d - \{0\}$

$$\xi^\dagger G \xi = \sum_i \sum_j \xi_i K(x_i, x_j) \xi_j = \sum_i \sum_j \xi_i \langle \psi(x_i), \psi(x_j) \rangle \xi_j = \sum_i \sum_j \langle \xi_i \psi(x_i), \psi(x_j) \xi_j \rangle$$

$$= \langle \sum_i \xi_i \psi(x_i), \sum_j \psi(x_j) \xi_j \rangle = \left\| \sum_i \xi_i \psi(x_i) \right\|_2^2 \geq 0$$

Assume the symmetric function $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive semi-definite. Let $\mathbb{R}^f = \{f : \mathcal{X} \to \mathbb{R}\}$. For $x \in \mathcal{X}$ let function $\psi$ over $\mathbb{R}^f$ with $\psi(x) = K(\cdot, x)$. This allows to define a vector space consisting of all the linear combinations of elements of the form $K(\cdot, x)$, having an inner product

$$\langle \sum_i \alpha_i K(\cdot, x_i), \sum_j \beta_j K(\cdot, x_j) \rangle = \sum_i \sum_i \alpha_i \beta_j \underbrace{\langle K(\cdot, x_i), K(\cdot, x_j) \rangle}_{= K(x_i, x_j)}.$$

This is satisfies all the properties of inner product, s.t. it is symmetric, linearity, positive definite as $K(x, x') \geq 0$. Then there is some feature vector $\psi$ such that $K(x, x') = \langle \psi(x), \psi(x') \rangle$. □

**Example 26.** Consider two kernel functions $K_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Show that $K_3$ with $K_3(x, x') = K_1(x, x') + K_2(x, x')$ is a kernel function.

**Solution.** Let Gram matrix, $G_j$ induced by kernel function $K_j$. For any $\xi \in \mathbb{R}^d - \{0\}$, it is

$$\xi^\top G_3 \xi = \xi^\top (G_1 + G_2) \xi = \xi^\top G_1 \xi + \xi^\top G_2 \xi \geq 0$$

that completes the proof according to Lemma 25.

## 5. Implementations

### 5.1. **Kernel SVM.**

### 5.2. **Kernel PCA.**

### 5.3. **Gaussian process regression.**

# Appendix A.