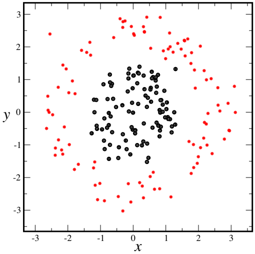
Statistical Machine Learning

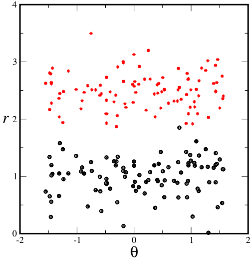
Lecture 9: Feature Space and Kernels

2022-23

Feature Space

- Some time changing data representation $\boldsymbol{x} \in \mathscr{X} \subseteq \mathbb{R}^d$ is a fact of convenience, $\phi : \mathscr{X} \to \mathscr{F} \subseteq \mathbb{R}^M$.
- If $M < d$ there is a dimensionality reduction (like in PCA).
- It may simplify the classification boundary.





Mapping into Feature Space

- In the previous slide the transformation is $\phi : \mathscr{X} \subseteq \mathbb{R}^2 \to \mathscr{F} \subseteq (-\pi, \pi] \times \mathbb{R}^+$, with $\theta = \phi_1(\boldsymbol{x}) = \arctan(y/x)$ and $r = \phi_2(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$.
- A non-linear mapping (ϕ) can be used in order to apply a linear machine in the feature space:, i.e. given $\phi : \mathscr{X} \subseteq \mathbb{R}^d \to \mathscr{F} \subseteq \mathbb{R}^M$, $\boldsymbol{w} \in \Omega \subseteq \mathbb{R}^M$ and $w_0 \in \mathbb{R}$:

$$f(\boldsymbol{x}) = \boldsymbol{w}^T \phi(\boldsymbol{x}) + w_0.$$

- We can also represent this linear machine in the dual representation:

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n y_n \phi(\boldsymbol{x}_n)^T \phi(\boldsymbol{x}) + w_0.$$

Mapping into Feature Space

- It is frequently used the bra-ket notation:

$$\phi(\boldsymbol{x}) = |\phi(\boldsymbol{x})\rangle \;\text{ column vector},$$
$$\phi(\boldsymbol{x})^T = \langle\phi(\boldsymbol{x})| \;\text{ row vector},$$
$$\phi(\boldsymbol{x})^T\phi(\boldsymbol{y}) = \langle\phi(\boldsymbol{x}), \phi(\boldsymbol{y})\rangle \;\text{ inner product}.$$

- The dual representation becomes:

$$f(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n y_n \langle\phi(\boldsymbol{x}_n), \phi(\boldsymbol{x})\rangle + w_0.$$

- If we have a means to compute the inner product $\langle\phi(\boldsymbol{x}_n), \phi(\boldsymbol{x})\rangle$ in feature space directly as a function of the original input points, it becomes possible to merge the two steps needed to build a non-linear learning machine.

Kernels

- A *kernel* is a function $K : \mathscr{X}^2 \to \mathbb{R}$, $\mathscr{X}$ an inner product space, that for all $\boldsymbol{x}, \boldsymbol{z} \in \mathscr{X}$:
$$K(\boldsymbol{x}, \boldsymbol{z}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{z}) \rangle ,$$
where $\phi : \mathscr{X} \to \mathscr{F}$.
- The use of kernels make the representation into the feature space implicit. The feature vectors $\phi(\boldsymbol{x})$ do not need to be explicitly computed.
- The dual representation induces that the training examples appear in the form of inner products, stored into a Gramm matrix $\boldsymbol{G}$ (symmetric and positive definite).

Kernels

- Equivalently, the inner product between feature vectors involving the elements of the data set, can be stored in a kernel matrix:
$$[\boldsymbol{K}]_{m,n} = K(\boldsymbol{x}_m, \boldsymbol{x}_n).$$
- The kernel function is also used to represent the linear machine:
$$f(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n y_n K(\boldsymbol{x}_n, \boldsymbol{x}) + w_0.$$
- Observe that by the use of an appropriate kernel, knowledge about the feature map ϕ is not needed.

Examples of Kernels

- The identity:
$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle \,.$$
- Linear: given $\boldsymbol{A} \in \mathbb{R}^{m \times d}$
$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{A}\boldsymbol{x}, \boldsymbol{A}\boldsymbol{y} \rangle = (\boldsymbol{A}\boldsymbol{x})^T \boldsymbol{A}\boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{y} = \boldsymbol{x}^T \boldsymbol{B}\boldsymbol{y},$$
with $\boldsymbol{B} = \boldsymbol{A}^T \boldsymbol{A}$.

Examples of Kernels

- Square:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \boldsymbol{x}, \boldsymbol{y} \rangle^2 = \sum_{i=1}^{d} \sum_{j=1}^{d} x_i x_j y_i y_j = \sum_{(i,j)=(1,1)}^{(d,d)} (x_i x_j)(y_i y_j),$$

  which is equivalent to the inner product between the feature vectors:

$$[\phi(\boldsymbol{x})]_{(i,j)} = (x_i x_j).$$

- Generalized square: $(\langle \boldsymbol{x}, \boldsymbol{y} \rangle + c)^2$. And other powers.

Properties of Kernels

- Symmetry:

$$K(\boldsymbol{x}, \boldsymbol{y}) = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle = \langle \phi(\boldsymbol{y}), \phi(\boldsymbol{x}) \rangle = K(\boldsymbol{y}, \boldsymbol{x}).$$

- Cauchy-Schwarz inequality:

$$\begin{aligned}(K(\boldsymbol{x}, \boldsymbol{y}))^2 &= \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{y}) \rangle^2 \\ &\leq \|\phi(\boldsymbol{x})\|_2^2 \, \|\phi(\boldsymbol{y})\|_2^2 = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}) \rangle \langle \phi(\boldsymbol{y}), \phi(\boldsymbol{y}) \rangle \\ &\leq K(\boldsymbol{x}, \boldsymbol{x}) K(\boldsymbol{y}, \boldsymbol{y})\end{aligned}$$

.

Properties of Kernels

- Suppose that the input space $\mathscr{X} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ is finite ($M < \infty$). Suppose $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ is symmetric. Consider the matrix $\boldsymbol{K}$ with entries $[\boldsymbol{K}]_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$.
- Because of the symmetry there exist $\boldsymbol{\Lambda}, \boldsymbol{V} \in \mathbb{R}^{M\times M}$ such that $\boldsymbol{V}\boldsymbol{V}^T = \boldsymbol{V}^T\boldsymbol{V} = \boldsymbol{I}$ and $\boldsymbol{\Lambda}$ diagonal, and $\mathrm{diag}(\boldsymbol{\Lambda}) = \{\lambda_1, \ldots, \lambda_M\}$ eigenvalues of $\boldsymbol{K}$ with eigenvectors given by the columns of $\boldsymbol{V}$.
- Assume that $\lambda_1 \geq \cdots \geq \lambda_M \geq 0$. Let us define the matrix $\boldsymbol{X} \in \mathbb{R}^{M\times M}$ with columns $\boldsymbol{x}_k$. Consider the transformation $\phi(\boldsymbol{X}) = \boldsymbol{\Lambda}^{1/2}\boldsymbol{V}^T$ such that $\phi(\boldsymbol{X})^T\phi(\boldsymbol{X}) = \boldsymbol{K}$.

Properties of Kernels

- In particular $\phi(\boldsymbol{x}_k) = \mathrm{col}(\boldsymbol{\Lambda}^{1/2}\boldsymbol{V}^T)_k$ and $\phi(\boldsymbol{x}_j)^T = \mathrm{row}(\boldsymbol{V}\boldsymbol{\Lambda}^{1/2})_j$,therefore $\langle\phi(\boldsymbol{x}_j), \phi(\boldsymbol{x}_k)\rangle = \sum_{\ell=1}^{M} \lambda_\ell v_{\ell,j} v_{\ell,k} = K(\boldsymbol{x}_j, \boldsymbol{x}_k)$.
- Let as assume now that the feature space $\mathscr{F}$ is infinite Hilbert space (a space that is complete and its norm is given by an inner product) (remember $\phi : \mathscr{X} \to \mathscr{F}$) we can generalize the definition of the inner product as:

$$\langle\phi(\boldsymbol{x}), \phi(\boldsymbol{z})\rangle = \sum_{\ell=1}^{\infty} \lambda_\ell \phi_\ell(\boldsymbol{x})\phi_\ell(\boldsymbol{z}) = K(\boldsymbol{x}, \boldsymbol{z}),$$

with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. (Mercer's Theorem).

.

Properties of Kernels

- Observe that this definition requires that for all $\boldsymbol{\psi} \in \mathscr{F}$ $\sum_{\ell=1}^{\infty} \lambda_\ell \psi_\ell^2 < \infty$, so $\mathscr{F}$ is the space that contains all sequences $\boldsymbol{\psi} = (\psi_1, \psi_2, \dots)$ with finite inner product norm given by $\{\lambda_\ell\}_{\ell=1}^{\infty}$.
- The linear machine will be represented by:

$$f(\boldsymbol{x}) = \sum_{\ell=1}^{\infty} \lambda_\ell \psi_\ell \phi_\ell(\boldsymbol{x}) + w_0 = \sum_{n=1}^{N} \alpha_n t_n K(\boldsymbol{x}_n, \boldsymbol{x}) + w_0,$$

where

$$\boldsymbol{\psi} = \sum_{n=1}^{N} \alpha_n t_n \boldsymbol{\phi}(\boldsymbol{x}_n).$$

.

Mercer's Theorem

- Theorem: Let $\mathscr{X}$ be a compact subset of $\mathbb{R}^d$ (compact is similar to require that all sequences of elements in $\mathscr{X}$ converge in $\mathscr{X}$.) Suppose K is a continuous symmetric function such that the integral operator $T_K : L_2(\mathscr{X}) \to L_2(\mathscr{X})$,

$$(T_K f)(\boldsymbol{y}) = \int_{\mathscr{X}} \mathrm{d}\boldsymbol{x} K(\boldsymbol{y}, \boldsymbol{x}) f(\boldsymbol{x}),$$

- is positive, i.e.

$$\int_{\mathscr{X} \times \mathscr{X}} \mathrm{d}\boldsymbol{x} \mathrm{d}\boldsymbol{z} K(\boldsymbol{x}, \boldsymbol{z}) f(\boldsymbol{x}) f(\boldsymbol{z}) \geq 0, \qquad \forall f \in L_2(\mathscr{X}).$$

Mercer's Theorem

- Then we can expand $K(\boldsymbol{x}, \boldsymbol{z})$ in a uniformly convergent series (on $\mathscr{X} \times \mathscr{X}$) in terms of T_K eigenfunctions $\phi_j \in L_2(\mathscr{X})$, normalised in such a way that $\|\phi_j\|_{L_2} = 1$, and positive associated eigenvalues $\lambda_j > 0$,

$$K(\boldsymbol{x}, \boldsymbol{z}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\boldsymbol{x}) \phi_j(\boldsymbol{z}).$$

- Observe that we do not need to compute the Mercer's features $\boldsymbol{\phi}(\boldsymbol{x}) \in \mathscr{F}$. We only need to compute their inner products through K, remember that

$$f(\boldsymbol{x}) = \sum_{\ell \in \mathrm{sv}} \alpha_\ell \boldsymbol{t}_\ell K(\boldsymbol{x}_\ell, \boldsymbol{x}).$$

Reproducing Kernel Hilbert Spaces

- Let us define the set $\mathscr{H} = \{\sum_{j=1}^{\ell} \alpha_j K(\boldsymbol{x}_j, \cdot) : \ell \in \mathbb{N}, \boldsymbol{x}_j \in \mathscr{X}, \alpha_j \in \mathbb{R}, j = 1, \ldots, \ell\}$. (Observe these are functions $\mathscr{H} \ni h : \mathscr{X} \to \mathbb{R}$).
- Observe that if $f(\cdot), g(\cdot) \in \mathscr{H}$ and $a, b \in \mathbb{R}$ then $af(\cdot) + bg(\cdot) \in \mathscr{H}$. Thus $\mathscr{H}$ is a vector space.
- Suppose $f(\cdot) = \sum_{j=1}^{\ell_f} \alpha_j K(\boldsymbol{x}_j, \cdot)$ and $g(\cdot) = \sum_{j=1}^{\ell_g} \beta_j K(\boldsymbol{z}_j, \cdot)$ both in $\mathscr{H}$. We define the inner product of $\mathscr{H}$ as

$$\begin{aligned}
\langle f, g \rangle_{\mathscr{H}} &= \sum_{j=1}^{\ell_f} \sum_{k=1}^{\ell_g} \alpha_j \beta_k K(\boldsymbol{x}_j, \boldsymbol{z}_k) \\
&= \sum_{j=1}^{\ell_f} \alpha_j g(\boldsymbol{x}_j) = \sum_{k=1}^{\ell_g} \beta_k f(\boldsymbol{z}_k).
\end{aligned}$$

Reproducing Kernel Hilbert Spaces

- The kernel matrices are positive semidefinite, i.e. for any collection of input space vectors $\{\boldsymbol{x}_j \in \mathscr{X}\}_{j=1}^{N}$, $\boldsymbol{K} \in \mathbb{R}^{N\times N}$ defined as $[\boldsymbol{K}]_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ is symmetric, real, and therefore positive semidefinite. Thus

$$\langle f, f\rangle_{\mathscr{H}} = \sum_{j=1}^{\ell_f}\sum_{k=1}^{\ell_f} \alpha_j K(\boldsymbol{x}_j, \boldsymbol{x}_k)\alpha_k = \boldsymbol{\alpha}^T \boldsymbol{K}\boldsymbol{\alpha} \geq 0.$$

- Observe also that $K(\boldsymbol{x}, \cdot) \in \mathscr{H}$ (with $\ell_K = 1$, $\alpha_1 = 1$, and $\boldsymbol{x}_1 = \boldsymbol{x}$). Thus

$$\langle f, K(\boldsymbol{x}, \cdot)\rangle_{\mathscr{H}} = \sum_{j=1}^{\ell_f} \alpha_j K(\boldsymbol{x}_j, \boldsymbol{x}) = f(\boldsymbol{x}),$$

which is known as the *reproducing property* of the kernel.

Reproducing Kernel Hilbert Spaces

- $\mathscr{H}$ is known as the *Reproducing Kernel Hilbert Space*.
- Learning implies to find a suitable function $f \in \mathscr{H}$ that separates classes. In the dual representation $f(\boldsymbol{x}) = \sum_{j \in \mathrm{sv}} \alpha_j K(\boldsymbol{x}_j, \boldsymbol{x}) \in \mathscr{H}$.