MACHINE LEARNING FUNDAMENTALS - LS2026
SEMINAR: KERNEL FUNCTIONS

CZECH TECHNICAL UNIVERSITY IN PRAGUE
V. FRANC

Assignment 1. Consider a feature map $\boldsymbol{\phi}\colon \mathbb{R}^d \to \mathbb{R}^{d(d+1)/2}$ whose entries are

$$\begin{array}{rcllllll} \boldsymbol{\phi}(\boldsymbol{x}) & = & (& x_1^2, & \sqrt{2}x_1x_2, & \sqrt{2}x_1x_3, & \dots, & \sqrt{2}x_1x_d, \\ & & & & x_2^2, & \sqrt{2}x_2x_3, & \dots, & \sqrt{2}x_2x_d, \\ & & & & & & & \vdots \\ & & & & & & & x_d^2 \qquad)^T, \end{array}$$

so that the features correspond to all possible products of unordered pairs of entries from $\boldsymbol{x}$, and the products of different entries are multiplied by a constant factor $\sqrt{2}$. For example, if $\boldsymbol{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$ then

$$\boldsymbol{\phi}(x) = (x_1^2, \sqrt{2}x_1x_2, \sqrt{2}x_1x_3, x_2^2, \sqrt{2}x_2x_3, x_3^2)^T \,.$$

This feature map defines a kernel $k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle$ referred to as the homogeneous polynomial kernel of degree 2. Show that the kernel value equals to the square of the dot product of the input vectors, that is prove the identity

$$k(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{\phi}(\boldsymbol{x}), \boldsymbol{\phi}(\boldsymbol{x}') \rangle = \langle \boldsymbol{x}, \boldsymbol{x}' \rangle^2 \,, \qquad \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d \,.$$

Solution 1. By definition of the feature map, the inner product in feature space is

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \sum_{i=1}^{d} x_i^2 x_i'^2 + \sum_{1 \le i < j \le d} (\sqrt{2}\, x_i x_j)(\sqrt{2}\, x_i' x_j').$$

Since $(\sqrt{2})(\sqrt{2}) = 2$, this simplifies to

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \sum_{i=1}^{d} x_i^2 x_i'^2 + 2 \sum_{1 \le i < j \le d} x_i x_j x_i' x_j'. \tag{1}$$

Now consider the square of the standard dot product:

$$\langle \mathbf{x}, \mathbf{x}' \rangle^2 = \left(\sum_{i=1}^{d} x_i x_i' \right)^2 .$$

Using the identity

$$\left(\sum_{i=1}^{d} a_i \right)^2 = \sum_{i=1}^{d} a_i^2 + 2 \sum_{1 \le i < j \le d} a_i a_j,$$

with $a_i = x_i x'_i$, we obtain

$$\left(\sum_{i=1}^{d} x_i x'_i\right)^2 = \sum_{i=1}^{d} (x_i x'_i)^2 + 2 \sum_{1 \le i < j \le d} (x_i x'_i)(x_j x'_j).$$

Hence

$$\langle \mathbf{x}, \mathbf{x}' \rangle^2 = \sum_{i=1}^{d} x_i^2 x_i'^2 + 2 \sum_{1 \le i < j \le d} x_i x_j x'_i x'_j. \tag{2}$$

The right-hand sides of (1) and (2) are identical, therefore

$$\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle = \langle \mathbf{x}, \mathbf{x}' \rangle^2.$$

Thus the homogeneous polynomial kernel of degree 2 is

$$k(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^T \mathbf{x}')^2.$$

Assignment 2. Given a training set $T_m = ((\mathbf{x}_i, y_i) \in \mathbb{R}^d \times \mathbb{R} \mid i = 1, \dots, m)$, consider the linear regressor

$$h(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b,$$

whose parameters $(\mathbf{w}, b) \in \mathbb{R}^{d+1}$ are learned from T_m by solving the least-squares problem. The resulting parameters are given by

$$\begin{bmatrix} \mathbf{w}_m \\ b_m \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

where

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^\top & 1 \\ \vdots & \vdots \\ \mathbf{x}_m^\top & 1 \end{bmatrix} \in \mathbb{R}^{m \times (d+1)}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix} \in \mathbb{R}^m.$$

Show that linear regression can be expressed solely in terms of dot products between input vectors. In particular, derive a formulation of the predictor that depends only on inner products

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle,$$

and explain how this leads to a kernelized version in which these dot products are replaced by a kernel

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle.$$

Solution 2. To simplify the notation, introduce the augmented vectors

$$\tilde{\mathbf{x}}_i = \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix} \in \mathbb{R}^{d+1}, \qquad \tilde{\mathbf{x}} = \begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix} \in \mathbb{R}^{d+1}, \qquad \tilde{\mathbf{w}} = \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}.$$

Then the regressor can be written as

$$h(\mathbf{x}; \mathbf{w}, b) = \langle \mathbf{w}, \mathbf{x} \rangle + b = \langle \tilde{\mathbf{w}}, \tilde{\mathbf{x}} \rangle.$$

The design matrix becomes

$$\mathbf{X} = \begin{bmatrix} \tilde{\mathbf{x}}_1^\top \\ \vdots \\ \tilde{\mathbf{x}}_m^\top \end{bmatrix},$$

and the least-squares solution is

$$\tilde{\mathbf{w}}_m = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.$$

Therefore the prediction for a new input $\mathbf{x}$ is

$$h_m(\mathbf{x}) = \langle \tilde{\mathbf{w}}_m, \tilde{\mathbf{x}} \rangle = \tilde{\mathbf{x}}^\top \tilde{\mathbf{w}}_m = \tilde{\mathbf{x}}^\top (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}.$$

Now use the identity

$$(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1},$$

which follows from

$$(\mathbf{X}^\top \mathbf{X})\mathbf{X}^\top = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)$$

by multiplying from the left by $(\mathbf{X}^\top \mathbf{X})^{-1}$ and from the right by $(\mathbf{X}\mathbf{X}^\top)^{-1}$. Hence

$$h_m(\mathbf{x}) = \tilde{\mathbf{x}}^\top \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y} = (\mathbf{X}\tilde{\mathbf{x}})^\top (\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{y}.$$

Define the Gram matrix

$$\mathbf{K} = \mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{m\times m}.$$

Its entries are

$$K_{ij} = \tilde{\mathbf{x}}_i^\top \tilde{\mathbf{x}}_j = \langle \mathbf{x}_i, \mathbf{x}_j \rangle + 1.$$

Also define the vector

$$\mathbf{k}(\mathbf{x}) = \mathbf{X}\tilde{\mathbf{x}} = \begin{bmatrix} \tilde{\mathbf{x}}_1^\top \tilde{\mathbf{x}} \\ \vdots \\ \tilde{\mathbf{x}}_m^\top \tilde{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} \langle \mathbf{x}_1, \mathbf{x} \rangle + 1 \\ \vdots \\ \langle \mathbf{x}_m, \mathbf{x} \rangle + 1 \end{bmatrix}.$$

Thus,

$$h_m(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top \mathbf{K}^{-1}\mathbf{y}.$$

If we set

$$\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y},$$

then the predictor becomes

$$h_m(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \big(\langle \mathbf{x}_i, \mathbf{x} \rangle + 1\big).$$

This shows that linear regression depends only on dot products between the input vectors.

Now replace the dot product by a kernel

$$k(\mathbf{x}, \mathbf{x}') = \langle \phi(\mathbf{x}), \phi(\mathbf{x}') \rangle.$$

Then define the kernel matrix

$$K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j),$$

and compute

$$\boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}.$$

The predictor becomes

$$h_m(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \, k(\mathbf{x}_i, \mathbf{x}).$$

Therefore, linear regression admits a dual representation based only on inner products, and this immediately leads to the kernelized form

$$\boxed{h_m(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \, k(\mathbf{x}_i, \mathbf{x}), \qquad \boldsymbol{\alpha} = \mathbf{K}^{-1}\mathbf{y}.}$$

For ordinary linear regression with bias, the corresponding kernel is

$$k(\mathbf{x}, \mathbf{x}') = \langle \mathbf{x}, \mathbf{x}' \rangle + 1.$$