Introduction to Non-Linear Models

Máster Universitario en Ciencia de Datos - Métodos Avanzados en Aprendizaje Automático

Carlos María Alaíz Gudín

Escuela Politécnica Superior
Universidad Autónoma de Madrid

Academic Year 2021/22

Universidad Autónoma de Madrid

Contents

Introduction

Limitation of the Linear Models

- Linear models are based on a strong assumption about the data:

  Regression There is a linear relation between inputs and output.

  Classification The classes are linearly separable.

- If such a relation is real, they are a good choice.

- The expressivity of linear models is very limited.
- The number of degrees of freedom corresponds to the number of input features d (plus the bias).
  - They are complex enough if d is large, or if the number of samples N is small.

- In many situations their underlying assumption is not true, and their expressivity is not enough.

Notebook

Limitation of Linear Models: Regression
Classification

Limitation of the Linear Models: Not Always Trivial

- It is not always easy to determine if linear models are appropriate or not for a particular dataset:
  - In a multidimensional context plotting the dataset is not enough.
  - Even if $N \gg d$, maybe there exists a linear relation (perhaps masked by the noise).
  - Even if $d \gg N$, maybe there is a lot of noise, and the effective dimension is small.

- It is always a good idea to start with a linear model and **check the performance**.

Notebook

Limitation of Linear Models: Not Always Trivial

Generalized Linear Models

Generalized Linear Models

Key Idea

- Instead of building the model over the original features, expand the data in a non-linear way.
- A non-linear mapping $\phi : \mathbb{R}^d \to \mathbb{R}^D$ is used.
- A linear model is built using as samples $\phi(\mathbf{x}_i)$ instead of $\mathbf{x}_i$.

- Formally, the model becomes:

$$f(\mathbf{x}) = \mathbf{w}^\intercal \phi(\mathbf{x}) = \sum_{i=1}^{D} w_i \phi_i(\mathbf{x}),$$

with $\mathbf{w} \in \mathbb{R}^D$ and $\mathbf{x} \in \mathbb{R}^d$, and where $\phi_i : \mathbb{R}^d \to \mathbb{R}$ is the i-th component of the mapping ϕ.

Generalized Linear Models - Exercise

Exercise

Given the following input data:

x_i
1
4

1. Compute the extended features for the mapping: $\phi(x) = \left(x, x^2, \sqrt{x}\right)$.
2. Compute the output of a generalized linear model with the mapping above, and with weights $\{b = 0, w_1 = 1, w_2 = 1, w_3 = 2\}$.

Solution

Extended features and estimated output:

$\phi_1(x_i)$	$\phi_2(x_i)$	$\phi_3(x_i)$	y_i
1	1	1	4
4	16	2	24

Data Matrix and Optimization

- The data matrix becomes $\mathbf{\Phi} \in \mathbb{R}^{N \times D}$:

$$\mathbf{\Phi} = \begin{pmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \dots & \phi_D(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \dots & \phi_D(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \dots & \phi_D(\mathbf{x}_N) \end{pmatrix}.$$

- The resultant optimization problem is hence:

$$\min_{\mathbf{w} \in \mathbb{R}^D} \left\{ \frac{1}{2} \|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 \right\},$$

with solution:

$$\boxed{\mathbf{w}^\star = \left(\mathbf{\Phi}^\intercal \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\intercal \mathbf{y}}.$$

- The mapping will be crucial for the performance of the model.

Feature Construction

- The features are carefully crafted by experts.

Advantages

- If there is expert knowledge, this approach can improve the performance.
- It does not depend (necessarily) on d or N, but on the nature of the problem.

Disadvantages

- It requires expert knowledge.
- It requires an intuition about the problem, which is difficult for d large.

Notebook

Generalized Linear Models: Feature Construction

Set of Basis Functions

- Another approach is to define a mapping general enough for any problem.

Advantages

- It is an automatic method.
- It does not require any expert knowledge or intuition.

Disadvantages

- The number of required basis functions grows rapidly due to the **curse of dimensionality**.
- It can generate a high redundancy.
- The resultant dimension D can be much larger than needed.

Basis Functions: Polynomial Regression (I)

Example (Polynomial 1-Dimensional Regression)

- The mapping transforms the input $x \in \mathbb{R}$ to a polynomial of degree M, $\phi_i(x) = x^{i-1}$, for $i = 1, \dots, M+1$.
- The model becomes:

$$f(\mathbf{x}) = w_1 + w_2 x + w_3 x^2 + \cdots + w_{M+1} x^M.$$

- The corresponding data matrix is the well-known Van der Monde matrix:

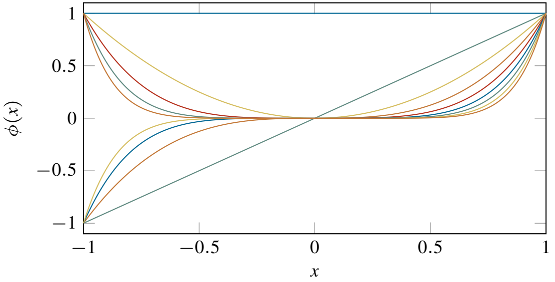
$$\mathbf{\Phi} = \mathbf{V} = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & \mathbf{x}_1^M \\ 1 & x_2 & x_2^2 & \dots & \mathbf{x}_2^M \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_N & x_N^2 & \dots & \mathbf{x}_N^M \end{pmatrix}.$$

- The optimum hyperplane is hence:

$$\mathbf{w}^\star = (\mathbf{V}^\intercal \mathbf{V})^{-1} \mathbf{V}^\intercal \mathbf{y}.$$

Basis Functions: Polynomial Basis

- Polynomial regression can be extended to multidimensional problems, using polynomial combinations of the original inputs up to order M.



Basis Functions: Polynomial Basis - Exercise

Exercise

Given the following input data:

x_i
1
2
3

1. Compute the extended features for the polynomial basis of degree $M = 3$.

Solution

Extended features:

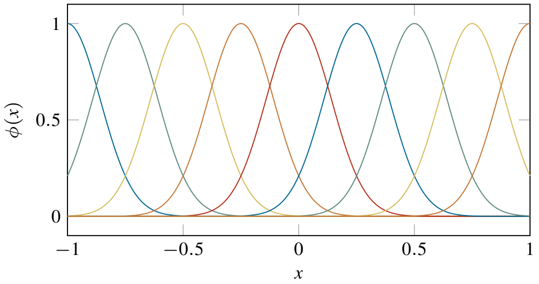
$\phi_1(x_i)$	$\phi_2(x_i)$	$\phi_3(x_i)$	$\phi_4(x_i)$
1	1	1	1
1	2	4	8
1	3	9	27

Basis Functions: Gaussian Basis

- The mapping is:

$$\phi_i(\mathbf{x}) = \exp\left(-\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|_2^2}{\sigma_i^2}\right),$$

  with $\boldsymbol{\mu}_i \in \mathbb{R}^d$ and $\sigma_i \in \mathbb{R}$.
- A Gaussian function is centred at $\boldsymbol{\mu}_i$ with deviation σ_i.



Basis Functions: Gaussian Basis - Exercise

Exercise

Given the following input data:

x_i
1
2
3

1. Compute the extended features for a Gaussian basis with three elements, with means $\mu_1 = 1$, $\mu_2 = 2$ and $\mu_3 = 3$, and deviation $\sigma_1 = \sigma_2 = \sigma_3 = 1$.

Solution

Extended features:

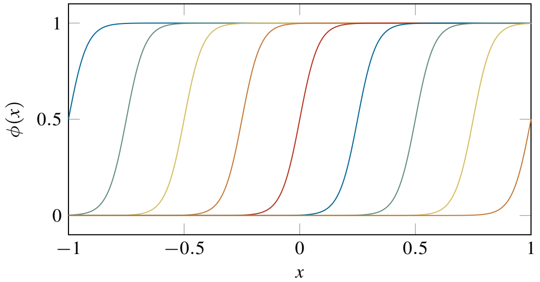
$\phi_1(x_i)$	$\phi_2(x_i)$	$\phi_3(x_i)$
1	0.37	0.02
0.37	1	0.37
0.02	0.37	1

Basis Functions: Sigmoidal Basis

- The mapping is:

$$\phi_i(\mathbf{x}) = \frac{1}{1 + \exp(-(\mathbf{a}_i^\intercal \mathbf{x} - b_i))},$$

with $\mathbf{a}_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$.



Basis Functions: Sigmoidal Basis - Exercise

Exercise

Given the following input data:

x_i
1
2
3

1. Compute the extended features for a sigmoidal basis with three elements, with means $b_1 = 1$, $b_2 = 2$ and $b_3 = 3$, and coefficients $a_1 = a_2 = a_3 = 1$.

Solution

Extended features:

$\phi_1(x_i)$	$\phi_2(x_i)$	$\phi_3(x_i)$
0.5	0.27	0.12
0.73	0.5	0.27
0.88	0.73	0.5

Basis Functions: Conclusions

- There are many other choices of basis functions:
  - Fourier basis (sinusoidal functions).
  - Wavelets.
  - Spline basis (piecewise polynomials; usually of degree 3).

- In the end, they require a partition of the space.
  - Affordable for d small.
  - Prohibitive for d large.

Notebook

Generalized Linear Models: Sets of Basis Functions

Other Approaches

Adaptive Basis Functions

- The mapping is also learned.
- It is automatically adapted to the data.
- Example: Neural Networks.

Kernel Trick

- Maybe it is not necessary to know explicitly ϕ...

Kernel Ridge Regression

The Model

Key Idea

- Ridge Regression applied over an extended feature space:

$$\min_{\mathbf{w}\in\mathbb{R}^D} \left\{ \frac{1}{2}\|\mathbf{y} - \mathbf{\Phi}\mathbf{w}\|_2^2 + \frac{\gamma}{2}\|\mathbf{w}\|_2^2 \right\}.$$

- A particular case are the previous generalized linear models.

- Ridge Regression admits a **dual formulation**.
- It turns out that the solution can be expressed using only scalar products between the vectors.

Primal Problem

- The standard Ridge Regression solution can be used to solve the optimization problem:

$$\mathbf{w}^\star = \left(\boldsymbol{\Phi}^\intercal \boldsymbol{\Phi} + \gamma \mathbf{I}\right)^{-1} \boldsymbol{\Phi}^\intercal \mathbf{y}.$$

- Procedure:
  1. Define the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^D$.
  2. Transform the data matrix explicitly from $\mathbf{X} \in \mathbb{R}^{N \times d}$ to $\boldsymbol{\Phi} \in \mathbb{R}^{N \times D}$.
  3. Solve the standard Ridge Regression problem by inverting a $D \times D$ matrix.
  4. Predict using $(\mathbf{w}^\star)^\intercal \phi(\mathbf{x})$.

- An alternative solution can be derived thanks to a constrained formulation of the problem and the **Lagrangian Duality**.

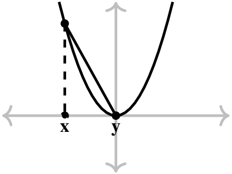
Lagrangian Duality: Convexity (I)

Convex Function

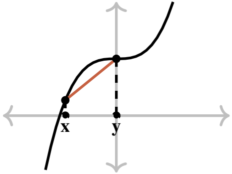
- Convex functions are specially suited for optimization.
- Formally, a real function f is called **convex** if its domain is a convex set (i.e. the line segment joining two points of the set belongs to the set), and $\forall \mathbf{x}, \mathbf{x}'$ and $\forall t \in [0, 1]$,

$$f(t\mathbf{x} + (1-t)\mathbf{x}') \leq t f(\mathbf{x}) + (1-t) f(\mathbf{x}').$$

- Conceptually, the line segment joining two points of the graph of f lies above or on the graph.



Convex function.



Non-convex function.

Lagrangian Duality: Convexity (II)

Proposition (Local Minima of Convex Functions)

Any local minimum of a convex function is a global minimum.

Proof.

1. Let $\mathbf{x}$ be a local minimum of f, and let $\mathbf{x}'$ be any other point on the domain of f.
2. Since $\mathbf{x}$ is a local minimum, there exists $t \in [0, 1]$ such that:

$$f(\mathbf{x}) \leq f(t\mathbf{x} + (1 - t)\mathbf{x}').$$

3. Hence, using the convexity:

$$\begin{aligned} & & f(\mathbf{x}) &\leq f(t\mathbf{x} + (1 - t)\mathbf{x}') \\ &\Longrightarrow & f(\mathbf{x}) &\leq tf(\mathbf{x}) + (1 - t)f(\mathbf{x}') \\ &\Longrightarrow & (1 - t)f(\mathbf{x}) &\leq (1 - t)f(\mathbf{x}') \\ &\Longrightarrow & f(\mathbf{x}) &\leq f(\mathbf{x}'). \end{aligned}$$

□

Lagrangian Duality: Convex Programming (I)

Definition (Convex Programming)

The standard formulation of a **Convex Programming (CP)** problem is:

$$\min_{\mathbf{x}} \ \{f(\mathbf{x})\}$$
$$\text{s.t.} \begin{cases} g_i(\mathbf{x}) \leq 0, \\ h_j(\mathbf{x}) = 0. \end{cases}$$

- $f(\mathbf{x})$ is the **convex** objective function.
- $g_i(\mathbf{x})$ are the **convex** inequality constrains.
- $h_j(\mathbf{x})$ are the **linear** equality constrains.

Lagrangian Duality: Convex Programming (II)

Definition (Quadratic Programming)

The standard formulation of a **Quadratic Programming (QP)** problem is:

$$\min_{\mathbf{x}} \left\{\mathbf{x}^\intercal \mathbf{Q}\mathbf{x} + \mathbf{c}^\intercal \mathbf{x})\right\}$$
$$\text{s.t.} \begin{cases} g_i(\mathbf{x}) \leq 0, \\ h_j(\mathbf{x}) = 0. \end{cases}$$

- $\mathbf{Q}$ is a **positive semidefinite** matrix.
- $g_i(\mathbf{x})$ are **linear** inequality constrains.
- $h_j(\mathbf{x})$ are **linear** equality constrains.

Lagrangian Duality: The Dual Problem (I)

$$\min_{\mathbf{x}} \ \{f(\mathbf{x})\} \ \text{ s.t. } \begin{cases} g_i(\mathbf{x}) \leq 0, \\ h_j(\mathbf{x}) = 0. \end{cases}$$

Lagrangian

$$\mathcal{L}(\mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\mathbf{x}) + \sum_i \alpha_i g_i(\mathbf{x}) + \sum_j \beta_j h_j(\mathbf{x}).$$

Saddle-Point Problem

$$\min_{\mathbf{x}} \left\{ \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}} \ \{\mathcal{L}(\mathbf{x}; \boldsymbol{\alpha}, \boldsymbol{\beta})\} \ \text{ s.t. } \ \boldsymbol{\alpha} \geq \mathbf{0} \right\}. \qquad ?$$

Lagrangian Duality: The Dual Problem (II)

- Focusing on the inner maximization problem:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ f(\mathbf{x}) + \sum_i \alpha_i g_i(\mathbf{x}) + \sum_j \beta_j h_j(\mathbf{x}) \right\} \text{ s.t. } \boldsymbol{\alpha} \geq \mathbf{0}.$$

- The problem is separable.

$$\max_{\alpha_i \geq 0} \{\alpha_i g_i(\mathbf{x})\} = \begin{cases} 0 & \text{if } g_i(\mathbf{x}) \leq 0, \\ \infty & \text{if } g_i(\mathbf{x}) > 0. \end{cases}$$

$$\max_{\beta_j} \{\beta_j h_j(\mathbf{x})\} = \begin{cases} 0 & \text{if } h_j(\mathbf{x}) = 0, \\ \infty & \text{if } h_j(\mathbf{x}) \neq 0. \end{cases}$$

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ f(\mathbf{x}) + \sum_i \alpha_i g_i(\mathbf{x}) + \sum_j \beta_j h_j(\mathbf{x}) \right\} \text{ s.t. } \boldsymbol{\alpha} \geq \mathbf{0} = \begin{cases} f(\mathbf{x}) & \text{if } g_i(\mathbf{x}) \leq 0 \text{ and } h_j(\mathbf{x}) = 0, \\ \infty & \text{otherwise.} \end{cases}$$

Lagrangian Duality: The Dual Problem (III)

- Therefore, the saddle-point problem is equivalent to the original one:

$$\min_{\mathbf{x}} \left\{ \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \{\mathcal{L}(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta})\} \text{ s.t. } \boldsymbol{\alpha} \geq \mathbf{0} \right\} \equiv \min_{\mathbf{x}} \{f(\mathbf{x})\} \text{ s.t. } \begin{cases} g_i(\mathbf{x}) \leq 0, \\ h_j(\mathbf{x}) = 0. \end{cases}$$

- Furthermore, the order of the problems can be inverted:

$$\min_{\mathbf{x}} \left\{ \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \{\mathcal{L}(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta})\} \text{ s.t. } \boldsymbol{\alpha} \geq \mathbf{0} \right\} \equiv \max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \left\{ \min_{\mathbf{x}} \{\mathcal{L}(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta})\} \right\} \text{ s.t. } \boldsymbol{\alpha} \geq \mathbf{0}.$$

- The **dual function** is defined as:

$$\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{\mathbf{x}} \{\mathcal{L}(\mathbf{x};\boldsymbol{\alpha},\boldsymbol{\beta})\}.$$

Lagrangian Duality: The Dual Problem (IV)

Dual Problem

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \{\mathcal{D}(\boldsymbol{\alpha},\boldsymbol{\beta})\} \text{ s.t. } \boldsymbol{\alpha} \geq \mathbf{0}.$$

- Both problems are equivalent if **strong duality** holds.
- In that case, the **duality gap** (different between the optimum of both problems) is zero.
- Sufficient conditions:
  - The primal problem is strictly feasible.
  - The constraints are linear.

Dual Problem: Lagrangian Duality (I)

- The Lagrangian duality can be used to get an alternative problem for Kernel Ridge Regression.
- The starting point is a constrained formulation of the original problem:

$$\min_{\mathbf{w}\in\mathbb{R}^D} \left\{ \frac{1}{2}\|\mathbf{y}-\mathbf{\Phi}\mathbf{w}\|_2^2 + \frac{\gamma}{2}\|\mathbf{w}\|_2^2 \right\} \equiv \min_{\substack{\mathbf{w}\in\mathbb{R}^D\\ \mathbf{e}\in\mathbb{R}^N}} \left\{ \frac{1}{2\gamma}\sum_{i=1}^N e_i^2 + \frac{1}{2}\|\mathbf{w}\|_2^2 \right\} \text{ s.t. } e_i = y_i - \mathbf{w}^\intercal \phi(\mathbf{x}_i).$$

- This constrained problem can be rewritten using the Lagrangian:

$$\mathcal{L}(\mathbf{w},\mathbf{e};\boldsymbol{\alpha}) = \frac{1}{2\gamma}\sum_{i=1}^N e_i^2 + \frac{1}{2}\|\mathbf{w}\|_2^2 + \sum_{i=1}^N \alpha_i(y_i - \mathbf{w}^\intercal\phi(\mathbf{x}_i) - e_i).$$

- The saddle-point problem is:

$$\min_{\substack{\mathbf{w}\in\mathbb{R}^D\\ \mathbf{e}\in\mathbb{R}^N}} \left\{ \max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \{\mathcal{L}(\mathbf{w},\mathbf{e};\boldsymbol{\alpha})\} \right\} \equiv \max_{\boldsymbol{\alpha}\in\mathbb{R}^N} \left\{ \min_{\substack{\mathbf{w}\in\mathbb{R}^D\\ \mathbf{e}\in\mathbb{R}^N}} \{\mathcal{L}(\mathbf{w},\mathbf{e};\boldsymbol{\alpha})\} \right\}.$$

Dual Problem: Lagrangian Duality (II)

- Solving the inner problem (taking derivatives with respect to $\mathbf{w}$ and e_i) leads to:

$$\frac{\partial}{\partial e_i}\mathcal{L}(\mathbf{w},\mathbf{e};\boldsymbol{\alpha}) = \frac{1}{\gamma}e_i - \alpha_i = 0 \implies e_i = \gamma\alpha_i;$$

$$\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w},\mathbf{e};\boldsymbol{\alpha}) = \mathbf{w} - \sum_{i=1}^{N}\alpha_i\phi(\mathbf{x}_i) = 0 \implies \boxed{\mathbf{w} = \sum_{i=1}^{N}\alpha_i\phi(\mathbf{x}_i)}.$$

- Substituting back leads to the dual problem:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^N}\{\mathcal{D}(\boldsymbol{\alpha})\} = \max_{\boldsymbol{\alpha}\in\mathbb{R}^N}\left\{-\frac{\gamma}{2}\|\boldsymbol{\alpha}\|_2^2 - \frac{1}{2}\boldsymbol{\alpha}^\intercal\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha} + \boldsymbol{\alpha}^\intercal\mathbf{y}\right\}.$$

- The solution is hence:

$$\nabla_{\boldsymbol{\alpha}}\mathcal{D}(\boldsymbol{\alpha})\big|_{\boldsymbol{\alpha}^\star} = -\gamma\boldsymbol{\alpha}^\star - \boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal\boldsymbol{\alpha}^\star + \mathbf{y} = 0 \implies \boxed{\boldsymbol{\alpha}^\star = \left(\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal + \gamma\mathbf{I}_N\right)^{-1}\mathbf{y}}.$$

Dual Problem: Procedure

- The dual formulation leads to an alternative approach.

- Procedure:
  1. Define the mapping $\phi : \mathbb{R}^d \to \mathbb{R}^D$.
  2. Transform the data matrix explicitly from $\mathbf{X} \in \mathbb{R}^{N \times d}$ to $\mathbf{\Phi} \in \mathbb{R}^{N \times D}$.
  3. Solve the dual Ridge Regression problem by inverting an $N \times N$ matrix as $\boldsymbol{\alpha}^\star = \left(\mathbf{\Phi}\mathbf{\Phi}^\intercal + \gamma \mathbf{I}_N\right)^{-1}\mathbf{y}$.
  4. Recompose $\mathbf{w}^\star = \mathbf{\Phi}^\intercal \boldsymbol{\alpha}^\star$.
  5. Predict using $(\mathbf{w}^\star)^\intercal \phi(\mathbf{x})$.

Notebook

Kernel Ridge Regression: Ridge Regression vs. Kernel Ridge Regression

The Kernel Trick (I)

- The solution of the dual problem is:

$$\boldsymbol{\alpha}^\star = \left(\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal + \gamma \mathbf{I}_N\right)^{-1} \mathbf{y}.$$

- The data only appears as $\mathbf{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal \in \mathbb{R}^{N\times N}$, with $k_{i,j} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^\intercal \phi(\mathbf{x}_j)$.
- The function $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is known as the **kernel function**.

- The kernel function computes the inner product in a certain Hilbert space.
- It can be defined directly, without a explicit form for ϕ.

The Kernel Trick (II)

- The primal hyperplane can be recovered as:

$$\mathbf{w}^\star = \boldsymbol{\Phi}^\intercal \boldsymbol{\alpha}^\star = \sum_{i=1}^{N} \alpha_i^\star \phi(\mathbf{x}_i).$$

- The prediction is hence:

$$f(\mathbf{x}) = (\mathbf{w}^\star)^\intercal \phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i^\star \phi(\mathbf{x}_i)^\intercal \phi(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i^\star \mathcal{K}(\mathbf{x}_i, \mathbf{x}).$$

- There is **no need to compute explicitly $\mathbf{w}^\star$**.
- Moreover, there is **no need to know ϕ as long as $\mathcal{K}$ is known**.

The Kernel Trick: Kernel Ridge Regression

- Procedure:
  1. Define the kernel function $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.
  2. Solve the dual Ridge Regression problem by inverting an $N \times N$ matrix.
  3. Predict using $\sum_{i=1}^N \alpha_i^\star \mathcal{K}(\mathbf{x}_i, \mathbf{x})$.

- Computing $\mathcal{K}$ has to be efficient, and it should not require to apply the mapping explicitly.

Building Kernel Functions

- A **kernel function** $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a symmetric, positive definite function.

- Given two kernels $\mathcal{K}_1(\mathbf{x}, \mathbf{x}')$ and $\mathcal{K}_2(\mathbf{x}, \mathbf{x}')$, and $c \in \mathbb{R}$, the following new kernels can be defined:
  - $\mathcal{K}_1(\mathbf{x}, \mathbf{x}') + c$.
  - $c\mathcal{K}_1(\mathbf{x}, \mathbf{x}')$, for $c > 0$.
  - $\mathcal{K}_1(\mathbf{x}, \mathbf{x}') + \mathcal{K}_2(\mathbf{x}, \mathbf{x}')$.
  - $\mathcal{K}_1(\mathbf{x}, \mathbf{x}')\mathcal{K}_2(\mathbf{x}, \mathbf{x}')$.

- Examples of kernels:

$$\begin{aligned}
\text{Linear} \quad & \mathcal{K}(\mathbf{x}, \mathbf{x}') = \mathbf{x}^\intercal \mathbf{x}'; \mathcal{K}(\mathbf{x}, \mathbf{x}') = c + \mathbf{x}^\intercal \mathbf{x}'; \mathcal{K}(\mathbf{x}, \mathbf{x}') = (\mathbf{x} - \boldsymbol{\mu})^\intercal \boldsymbol{\Sigma}^{-1} (\mathbf{x}' - \boldsymbol{\mu}). \\
\text{Polynomial (degree } d\text{)} \quad & \mathcal{K}(\mathbf{x}, \mathbf{x}') = (\mathbf{x}^\intercal \mathbf{x}' + c)^d. \\
\text{Gaussian (RBF)} \quad & \mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma \|\mathbf{x} - \mathbf{x}'\|_2^2\right). \\
\text{Exponential} \quad & \mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp(-\gamma \|\mathbf{x} - \mathbf{x}'\|_2).
\end{aligned}$$

- There are many more: Gamma Exponential, Sigmoidal, Matérn, Periodic Kernel...
- The kernel (and its hyper-parameters) has to be carefully selected.

Notebook

Kernel Ridge Regression: Polynomial Kernel
RBF Kernel

Introduction to Non-Linear Models

Carlos María Alaíz Gudín

Additional Material - Alternative Derivation for KRR Dual Problem

Dual Problem: Matrix Identity (I)

- There is an additional derivation of the Kernel Ridge Regression dual problem, based on the following matrix identity that allows to rewrite the solution:

$$\left(\mathbf{P}^{-1} + \mathbf{B}^\intercal \mathbf{R}^{-1} \mathbf{B}\right)^{-1} \mathbf{B}^\intercal \mathbf{R}^{-1} = \mathbf{P}\mathbf{B}^\intercal (\mathbf{B}\mathbf{P}\mathbf{B}^\intercal + \mathbf{R})^{-1}.$$

- In particular, this identity can be applied to the expression $\mathbf{w}^\star = (\mathbf{\Phi}^\intercal \mathbf{\Phi} + \gamma \mathbf{I})^{-1} \mathbf{\Phi}^\intercal \mathbf{y}$:

$$\left(\mathbf{P}^{-1} + \mathbf{B}^\intercal \mathbf{R}^{-1} \mathbf{B}\right)^{-1} \mathbf{B}^\intercal \mathbf{R}^{-1} = \mathbf{P}\mathbf{B}^\intercal (\mathbf{B}\mathbf{P}\mathbf{B}^\intercal + \mathbf{R})^{-1}$$

$$\boxed{\mathbf{P} = \gamma^{-1}\mathbf{I}_D} \implies \left(\gamma \mathbf{I}_D + \mathbf{B}^\intercal \mathbf{R}^{-1} \mathbf{B}\right)^{-1} \mathbf{B}^\intercal \mathbf{R}^{-1} = \gamma^{-1}\mathbf{I}_D \mathbf{B}^\intercal \left(\mathbf{B}\gamma^{-1}\mathbf{I}_D \mathbf{B}^\intercal + \mathbf{R}\right)^{-1}$$

$$\boxed{\mathbf{R} = \gamma \mathbf{I}_N} \implies \left(\gamma \mathbf{I}_D + \mathbf{B}^\intercal \gamma^{-1}\mathbf{I}_N \mathbf{B}\right)^{-1} \mathbf{B}^\intercal \gamma^{-1}\mathbf{I}_N = \gamma^{-1}\mathbf{I}_D \mathbf{B}^\intercal \left(\mathbf{B}\gamma^{-1}\mathbf{I}_D \mathbf{B}^\intercal + \gamma \mathbf{I}_N\right)^{-1}$$

$$\boxed{\mathbf{B} = \gamma^{\frac{1}{2}}\mathbf{\Phi}} \implies \left(\gamma \mathbf{I}_D + \gamma^{\frac{1}{2}}\mathbf{\Phi}^\intercal \gamma^{-1}\mathbf{I}_N \gamma^{\frac{1}{2}}\mathbf{\Phi}\right)^{-1} \gamma^{\frac{1}{2}}\mathbf{\Phi}^\intercal \gamma^{-1}\mathbf{I}_N = \gamma^{-1}\mathbf{I}_D \gamma^{\frac{1}{2}}\mathbf{\Phi}^\intercal \left(\gamma^{\frac{1}{2}}\mathbf{\Phi}\gamma^{-1}\mathbf{I}_D \gamma^{\frac{1}{2}}\mathbf{\Phi}^\intercal + \gamma \mathbf{I}_N\right)^{-1}$$

$$\implies \gamma^{-\frac{1}{2}}(\gamma \mathbf{I}_D + \mathbf{\Phi}^\intercal \mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal = \gamma^{-\frac{1}{2}}\mathbf{\Phi}^\intercal (\mathbf{\Phi}\mathbf{\Phi}^\intercal + \gamma \mathbf{I}_N)^{-1}$$

$$\implies \mathbf{w}^\star = (\gamma \mathbf{I}_D + \mathbf{\Phi}^\intercal \mathbf{\Phi})^{-1}\mathbf{\Phi}^\intercal \mathbf{y} = \mathbf{\Phi}^\intercal (\mathbf{\Phi}\mathbf{\Phi}^\intercal + \gamma \mathbf{I}_N)^{-1}\mathbf{y}.$$

Dual Problem: Matrix Identity (II)

- Therefore, there is an equivalent solution for $\mathbf{w}^\star$:

$$\boxed{\mathbf{w}^\star = \boldsymbol{\Phi}^\intercal \boldsymbol{\alpha}^\star = \sum_{i=1}^{N} \alpha_i^\star \phi(\mathbf{x}_i)},$$

  based on the dual coefficients $\boldsymbol{\alpha}^\star \in \mathbb{R}^N$.

- The optimum dual coefficients are computed as:

$$\boxed{\boldsymbol{\alpha}^\star = \left(\boldsymbol{\Phi}\boldsymbol{\Phi}^\intercal + \gamma \mathbf{I}_N\right)^{-1} \mathbf{y}}.$$

- This result is exactly the one obtained using Lagrangian duality.