Outline

Mathematical Background

- Lecture 1: Linear regression: A basic data analytic tool
- Lecture 2: Regularization: Constraining the solution
- Lecture 3: Kernel Method: Enabling nonlinearity

Lecture 3: Kernel Method

- Kernel Method
  - Dual Form
  - Kernel Trick
  - Algorithm

- Examples
  - Radial Basis Function (RBF)
  - Regression using RBF
  - Kernel Methods in Classification
Why Another Method?

- Linear regression: Pick a **global** model, best fit globally.
- Kernel method: Pick a **local** model, best fit locally.
- In kernel method, instead of picking a line / a quadratic equation, we pick a **kernel**.
- A kernel is a measure of **distance** between **training samples**.
- Kernel method buys us the ability to handle nonlinearity.
- Ordinary regression is based on the **columns** (features) of $A$.
- Kernel method is based on the **rows** (samples) of $A$. 
Pictorial Illustration

**goal:** learn the surface

**prediction:** when new sample comes, interpolate on the surface
Overview of the Method

Model Parameter:
- We want the model parameter $\hat{\theta}$ to look like: (How? Question 1)
  \[ \hat{\theta} = \sum_{n=1}^{N} \alpha_n x^n. \]
- This model expresses $\hat{\theta}$ as a combination of the samples.
- The trainable parameters are $\alpha_n$, where $n = 1, \ldots, N$.
- If we can make $\alpha_n$ local, i.e., non-zero for only a few of them, then we can achieve our goal: localized, sample-dependent.

Predicted Value
- The predicted value of a new sample $x$ is
  \[ \hat{y} = \hat{\theta}^T x = \sum_{n=1}^{N} \alpha_n \langle x, x^n \rangle. \]
- We want this model to encapsulate nonlinearity. (How? Question 2)
Dual Form of Linear Regression

Goal: Addresses Question 1: Express $\hat{\theta}$ as

$$\hat{\theta} = \sum_{n=1}^{N} \alpha_n x^n.$$

We start by listing out a technical lemma:

**Lemma**

*For any $A \in \mathbb{R}^{N \times d}$, $y \in \mathbb{R}^d$, and $\lambda > 0$,*

$$\left(A^T A + \lambda I\right)^{-1} A^T y = A^T (AA^T + \lambda I)^{-1} y.$$  \hspace{1cm} (1)

Proof: See Appendix.

Remark:
- The dimensions of $I$ on the left is $d \times d$, on the right is $N \times N$.
- If $\lambda = 0$, then the above is true only when $A$ is invertible.
Using the Lemma, we can show that

\[ \hat{\theta} = (A^T A + \lambda I)^{-1} A^T y \]  

(Primal Form)

\[ = A^T (AA^T + \lambda I)^{-1} y \]  

(Dual Form)

\[ \text{def} \equiv \alpha \]

\[
\begin{bmatrix}
- (x^1)^T \\
- (x^2)^T \\
\vdots \\
- (x^N)^T
\end{bmatrix}
\begin{bmatrix}
\alpha_1 \\
\vdots \\
\alpha_N
\end{bmatrix}
= \sum_{n=1}^{N} \alpha_n x^n, \quad \alpha_n \text{ def } [(AA^T + \lambda I)^{-1} y].
\]
The Kernel Trick

**Goal:** Addresses Question 2: Introduce nonlinearity to

\[ \hat{y} = \hat{\theta}^T x = \sum_{n=1}^{N} \alpha_n \langle x, x^n \rangle. \]

**The Idea:**
- Replace the inner product \( \langle x, x^n \rangle \) by \( k(x, x^n) \):

\[ \hat{y} = \hat{\theta}^T x = \sum_{n=1}^{N} \alpha_n k(x, x^n). \]

- \( k(\cdot, \cdot) \) is called a **kernel**.
- A kernel is a measure of the **distance** between two samples \( x^i \) and \( x^j \).
- \( \langle x^i, x^j \rangle \) measure distance in the ambient space, \( k(x^i, x^j) \) measure distance in a **transformed** space.
- In particular, a valid kernel takes the form \( k(x^i, x^j) = \langle \phi(x^i), \phi(x^j) \rangle \) for some nonlinear transforms \( \phi \).
A kernel typically lifts the ambient dimension to a **higher** one.

For example, mapping from $\mathbb{R}^2$ to $\mathbb{R}^3$

$$x^n = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \text{and} \quad \phi(x_n) = \begin{bmatrix} x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix}$$
Consider the following kernel \( k(u, v) = (u^T v)^2 \). What is the transform?

- Suppose \( u \) and \( v \) are in \( \mathbb{R}^2 \). Then \( (u^T v)^2 \) is

\[
(u^T v)^2 = \left( \sum_{i=1}^{2} u_i v_i \right) \left( \sum_{j=1}^{2} u_j v_j \right) = \sum_{i=1}^{2} \sum_{j=1}^{2} (u_i u_j)(v_i v_j) = \begin{bmatrix} u_1^2 & u_1 u_2 & u_2 u_1 & u_2^2 \end{bmatrix} \begin{bmatrix} v_1^2 \\ v_1 v_2 \\ v_2 v_1 \\ v_2^2 \end{bmatrix}.
\]

- So if we define \( \phi \) as

\[
\begin{bmatrix} u_1 \\ u_2 \end{bmatrix} \mapsto \phi(u) = \begin{bmatrix} u_1^2 \\ u_1 u_2 \\ u_2 u_1 \\ u_2^2 \end{bmatrix}
\]

then \( (u^T v)^2 = \langle \phi(u), \phi(v) \rangle \).
Radial Basis Function

A useful kernel is the **radial basis kernel** (RBF):

\[
k(u, v) = \exp \left\{ -\frac{\|u - v\|^2}{2\sigma^2} \right\}.
\]

- The corresponding nonlinear transform of RBF is **infinite dimensional**. See Appendix.
- \(\|u - v\|^2\) measures the distance between two data points \(u\) and \(v\).
- \(\sigma\) is the std dev, defining “far” and “close”.
- RBF enforces **local** structure; Only a few samples are used.
Kernel Method

Given the choice of the kernel function, we can write down the algorithm as follows.

1. Pick a kernel function $k(\cdot, \cdot)$.
2. Construct a kernel matrix $K \in \mathbb{R}^{N \times N}$, where $[K]_{ij} = k(x^i, x^j)$, for $i = 1, \ldots, N$ and $j = 1, \ldots, N$.
3. Compute the coefficients $\alpha \in \mathbb{R}^N$, with
   \[ \alpha_n = [(K + \lambda I)^{-1} y]_n. \]
4. Estimate the predicted value for a new sample $x$:
   \[ g_\theta(x) = \sum_{n=1}^{N} \alpha_n k(x, x^n). \]

Therefore, the choice of the regression function is shifted to the choice of the kernel.