→ Support Vector Machine

- **What is Support Vector Machine**

SVM is a supervised machine learning algorithm used for classification and regression, which finds an optimal decision boundary (hyperplane) that maximizes the margin between different classes. For regression, it is called Support Vector Regression (SVR).

- **Purpose (When)**

To achieve maximum generalization by finding the decision boundary with the largest margin, making the model robust to noise and overfitting.

- **Why it is used**
  - Works well with high-dimensional data
  - Effective when number of features > number of samples.
  - Strong theoretical guarantees
  - Can model non-linear relationships using kernels
  - Robust to overfitting with proper regularization.

- **How it works**
  - Core ideas:
    - Maximum margin classifier
    - Support Vectors
    - Kernel Trick
  - Algorithm:
    1. Map input data into feature space
    2. Find the hyperplane that maximizes margin
    3. Identify support vectors (closest points)
    4. Allow soft margin using slack variables
    5. Optimize using quadratic programming.

Formula

- Hyperplane : ~~ax~~ $w.x + b = 0$
- Margin : $\frac{2}{\|w\|}$
- Optimization objective : $\min \frac{1}{2}\|w\|^2 + C \sum \xi_i$
- Subject to : $y_i(w.x_i + b) \geq 1 - \xi_i$

Kernel Trick (Non-linearity) [kernel : formula : use case]

- Linear : $x_i . x_j$: Linearly seperable
- Polynomial : $(\gamma x_i . x_j + r)^d$: Curved boundaries
- RBF (Gaussian) : $e^{-\gamma\|x_i - x_j\|^2}$: Complex non-linear
- Sigmoid : $\tanh(\gamma x_i . x_j + r)$: Neural-like

Technical Details

- Types : • Linear SVM
  - Support vector regression (SVR)
  - Non-linear SVM
- Loss Function : • Hinge Loss $L = \max(0, 1 - y(w.x))$
- Optimization : • Quadratic Programming
  - Sequential Minimal Optimization (SMO)

Parameters

- Model Parameters (Learned) : • Weight vector w
  - Bias b
  - Support vectors
  - Dual coefficients
- Hyperparameters : • C : Regularization strength
  - kernel : Kernel type
  - gamma : Influence of single point
  - degree : Polynomial degree
  - epsilon : Margin width (SVR)

- **Assumptions**
  - Data is seperable (or approx separable)
  - Correct kernel choice captures structure
  - Features are scaled.

- **Pros**
  - Strong generalization
  - Works in high dimensional spaces
  - Effective with small datasets
  - Flexible using kernels
  - Robust to overfitting

- **Cons**
  - Computationally expensive
  - Memory intensive
  - Hard to interpret
  - Sensitive to hyperparameters
  - Not scalable to very large datasets.

- Real World Applications (where)
  - Text & NLP :
    - spam detection
    - sentiment analysis
    - Document classification
  - Image processing :
    - Face recognition
    - Handwritten digit recognition
  - Bioinformatics :
    - Cancer classification
    - Gene expression analysis
  - Finance :
    - credit risk prediction
    - Fraud detection

- Suppor vector Regression: Fits a function within an ε-tube, ignoring small errors. ~~Ide~~ Objective: $\min \frac{1}{2}\|w\|^2 + C\sum(\xi_i + \xi_i^*)$