### **1.5 SVR (Support Vector Regression)**

Là một biến thể của **SVM (Support Vector Machine)** được mở rộng để giải quyết bài toán **hồi quy (regression)** thay vì phân loại. SVR tìm một hàm f(x) sao cho sai số giữa giá trị dự đoán và giá trị thực nằm trong một ngưỡng cho trước ε (epsilon), và đồng thời đảm bảo mô hình có **tính đơn giản (flatness)**. [Nguồn]

(**Nguồn:**

* **Vapnik, V. N. (1995)**. *The Nature of Statistical Learning Theory*. Springer.  
   https://doi.org/10.1007/978-1-4757-2440-0
* **Smola, A. J., & Schölkopf, B. (2004)**. *A Tutorial on Support Vector Regression*. *Statistics and Computing*, 14(3), 199–222.  
   https://doi.org/10.1023/B:STCO.0000035301.49549.88

)

SVR là một trong những thuật toán mạnh mẽ, hiệu quả và được ứng dụng nhiều trong các bài toán dự báo thời gian như dự đoán giá cổ phiếu, thời tiết...

### Ý tưởng: Tìm siêu phẳng tối ưu sao cho độ lệch của dữ liệu nằm trong khoảng ϵ.

### **Mục tiêu của SVR:**

### Tìm một **hàm hồi quy f(x)** sao cho:

### với phần lớn dữ liệu, và mô hình có tính **phẳng** (flatness) nhất có thể.

### Cho phép một vài điểm nằm ngoài ε bằng cách sử dụng **slack variables**: ξ và ξ\*

**Mã latex:**

**\subsection\*{Objective of SVR}**

**\begin{itemize}**

**\item Find a regression function \( f(x) \) such that:**

**\[**

**|f(x\_i) - y\_i| \leq \varepsilon**

**\]**

**for most of the data points, while keeping the model as \textbf{flat} as possible.**

**\item Allow some data points to fall outside the \( \varepsilon \)-insensitive zone by introducing \textbf{slack variables} \( \xi \) and \( \xi^\* \).**

**\end{itemize}**

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

### 

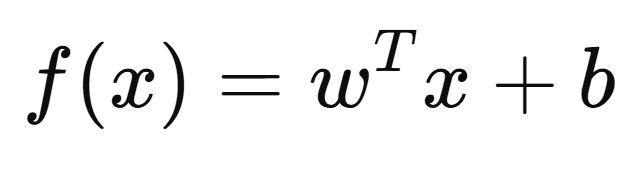
### 

### **Phương trình SVR cơ bản:**

Giả sử một hàm hồi quy dạng tuyến tính:

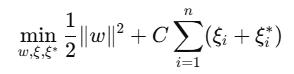
 [Nguồn]

Hàm hồi quy trong **Support Vector Regression (SVR)** là:

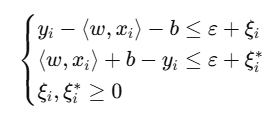


|  |  |
| --- | --- |
| x | **Vector đặc trưng đầu vào** (feature vector), ví dụ: [Open, High, Low, Volume] của một ngày |
| w | **Vector trọng số** tương ứng với mỗi đặc trưng – cái mà mô hình học được trong quá trình huấn luyện |
| w^T\*x | **Tích vô hướng** giữa vector trọng số và đầu vào – đây là giá trị dự đoán “thô” |
| b | **Hằng số dịch (bias term)** – giúp mô hình điều chỉnh đường dự đoán để khớp với dữ liệu thực tế hơn |

Tối ưu:

[Nguồn]

Ràng buộc:

[Nguồn]

**(Nguồn:**

* Smola & Schölkopf (2004), mục “Support Vector Regression: Formulation”, trang 200–201  
  <https://link.springer.com/article/10.1023/B:STCO.0000035301.49549.88>)

### **Các thành phần chính** [Nguồn]**:**

* **ε (epsilon):** Biên sai số chấp nhận được (dải ống ε-insensitive).
* **C (regularization parameter):** Hệ số điều chỉnh mức độ cho phép mô hình học quá sát dữ liệu huấn luyện (tránh overfitting).
* **ξ, ξ\*** (slack variables): Đại diện cho sai số vượt quá ε.
* **Kernel function:** Hàm nhân giúp SVR mở rộng cho các bài toán phi tuyến bằng cách ánh xạ dữ liệu vào không gian đặc trưng cao hơn.

(**Nguồn:**

* **Scikit-learn documentation** – SVR (Support Vector Regression):  
   https://scikit-learn.org/stable/modules/svm.html#svr
* **Bishop, C. M. (2006)**. *Pattern Recognition and Machine Learning*. Springer.  
   Mục 7.1.2: ε-insensitive loss function và regularization in SVR.

)

**Mã latex:**

**\subsection\*{SVR Objective}**

**The goal of Support Vector Regression (SVR) is to find a regression function $f(x)$ such that:**

**\[**

**|f(x\_i) - y\_i| \leq \varepsilon**

**\]**

**for most of the training data, while keeping the function as \textbf{flat} as possible.**

**To handle outliers beyond the $\varepsilon$ margin, SVR introduces \textit{slack variables} $\xi\_i$ and $\xi\_i^\*$ to allow some flexibility.**

**\subsection\*{Basic SVR Formulation}**

**Assume a linear regression function of the form:**

**\[**

**f(x) = \langle w, x \rangle + b**

**\]**

**Then the SVR regression function becomes:**

**\[**

**f(x) = w^T x + b**

**\]**

**Where:**

**\begin{itemize}**

**\item $x$: \textbf{Input feature vector}, e.g., \texttt{[Open, High, Low, Volume]} for a day**

**\item $w$: \textbf{Weight vector} associated with each input feature (learned during training)**

**\item $w^T x$: \textbf{Dot product} between weights and input features — this is the raw predicted output**

**\item $b$: \textbf{Bias term} — allows the model to better fit the actual data by shifting the prediction line**

**\end{itemize}**

**\subsection\*{Optimization Problem}**

**SVR solves the following optimization problem:**

**\[**

**\min\_{w,\xi\_i,\xi\_i^\*} \frac{1}{2} \|w\|^2 + C \sum\_{i=1}^{n} (\xi\_i + \xi\_i^\*)**

**\]**

**Subject to the constraints:**

**\[**

**\begin{cases}**

**y\_i - \langle w, x\_i \rangle - b \leq \varepsilon + \xi\_i \\**

**\langle w, x\_i \rangle + b - y\_i \leq \varepsilon + \xi\_i^\* \\**

**\xi\_i, \xi\_i^\* \geq 0**

**\end{cases}**

**\]**

**\subsection\*{Key Components}**

**\begin{itemize}**

**\item $\varepsilon$ (\textbf{epsilon}): Tolerance margin — the width of the $\varepsilon$-tube within which no penalty is given**

**\item $C$ (\textbf{regularization parameter}): Controls the trade-off between model complexity and training error**

**\item $\xi\_i$, $\xi\_i^\*$ (\textbf{slack variables}): Represent the degree of violation of the $\varepsilon$-margin**

**\item \textbf{Kernel function}: Extends SVR to nonlinear problems by mapping data to higher-dimensional feature spaces**

**\end{itemize}**

### **Các Kernel thường dùng** [Nguồn]**:**

* **Linear kernel:** 
* **Polynomial kernel:** 
* **RBF kernel (phổ biến nhất):**

(**Nguồn:**

* Scikit-learn kernel functions documentation:  
   https://scikit-learn.org/stable/modules/svm.html#kernels
* Smola & Schölkopf (2004), mục “Kernels in SVR”.

)

### **Ý nghĩa các metric:**

| **Metric** | **Diễn giải** | **Giá trị tốt** |
| --- | --- | --- |
| **MAE** | Sai số tuyệt đối trung bình | Càng nhỏ càng tốt |
| **MSE** | Trung bình bình phương sai số | Càng nhỏ càng tốt |
| **RMSE** | Căn bậc hai của MSE | Càng nhỏ càng tốt |
| **MAPE** | Phần trăm sai số tuyệt đối trung bình | Càng nhỏ càng tốt |
| **R²** | Hệ số xác định (độ phù hợp mô hình) | Càng gần 1 càng tốt |

### **—----------------------------------------------------------**

**Evaluate metric**

| ARIMA | Train - test - validate | MAE | MSE | RMSE | MAPE | R² |
| --- | --- | --- | --- | --- | --- | --- |
| 75 - 15 -10 | **82.32** | 13539.96 | 116.36 | **0.73%** | 0.9882 |
| 70 - 20 -10 | 84.75 | **12804.20** | **113.16** | 0.74% | 0.9898 |
| 65 - 25 - 10 | 99.03 | 20434.6 | 142.95 | 0.82% | **0.9914** |

So sánh 75/15/10 và 70/20/10 :Dù MSE và RMSE của 70/20/10 có tốt hơn, nhưng **chênh lệch không đáng kể**, chỉ vài đơn vị trên nền hàng chục nghìn. Trong khi MAE chênh lệch lại rõ hơn.

So sánh 75/15/10 và 65/25/10: 65/25/10 chỉ có **R²** cao nhất, nhưng lại đi kèm với **sai số lớn hơn nhiều**, có thể là dấu hiệu overfitting nhẹ.

### **Kết luận:**

* **Tốt nhất tổng thể: Tỷ lệ 75/15/10**
  + Có **MAE và MAPE thấp nhất**, nghĩa là mô hình này **dự đoán sai số trung bình nhỏ hơn**.
  + R² tuy không cao nhất, nhưng vẫn rất sát (0.9882).
  + Mô hình này cân bằng tốt giữa **độ chính xác** và **độ ổn định**.