Supervised, Unsupervised and Reinforcement Learning in Finance

Week 1: Supervised Learning

Support Vector Machines

Igor Halperin

NYU Tandon School of Engineering, 2017

Support Vector Machines

- SVM was developed as a classification method by Vapnik et al. in 1992 within the framework of statistical learning theory
- SVM was generalized to handle regression problems by Vapnik et al in mid 1990s (Support Vector Regression, or SVR for short). We will collectively refer to both SVM and SVR as SVM
- SVM is based on simple and beautiful geometric ideas (maximum margin hyperplane classifiers). Computationally, SVM amounts to convex optimization
- SVM handles both linear and non-linear regression. Unlike NN where non-linearity is implicit via the choice of both architecture and parameters, SVM controls non-linearity explicitly via the choice of the kernel
- SVM usually shows better out-of-sample performance than alternative methods including shallow NN
- SVM is widely used nowadays for classification and regression analyses such as object identification, text recognition, bioinformatics, speech recognition, etc. Previous financial applications mostly dealt with stock price predictions or bankruptcy predictions

Sketch of SVM math

- Start with linear regression:

$$f(x) = \langle w, x \rangle + b \ , \ w \in \mathcal{X}, \ b \in \mathbb{R}$$

where $\mathcal{X}$ is the input pattern space (e.g. $\mathcal{X} = \mathbb{R}^d$), $\langle a, b \rangle$ is a dot product in $\mathcal{X}$. Want to fit a function $f(x)$ admitting at most deviation of ε from the data

- Flatness means a small w - can be formulated as quadratic optimization min $\frac{1}{2}||w||^2$ subject to constraints $y_i - \langle w, x_i \rangle - b \leq \varepsilon$ and $\langle w, x_i \rangle + b - y_i \leq \varepsilon$

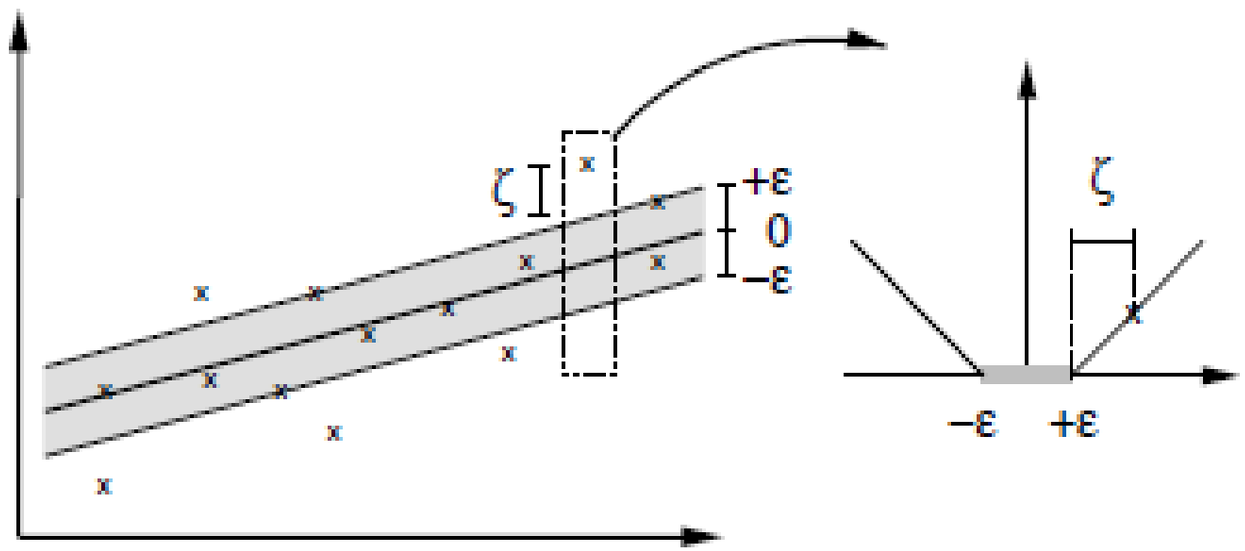
- To make the problem feasible, introduce slack variables ξ_i, ξ_i^*, and reformulate the problem as follows:

$$\text{min } \frac{1}{2}||w||^2 + C \sum_i (\xi_i + \xi_i^*)$$

subject to $y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i$ and $\langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^*$, with $\xi_i, \xi_i^* \geq 0$. Parameter $C > 0$ determines the trade-off between flatness and our tolerance for deviations larger than ε.

Sketch of SVM math (cont-ed)

- Geometric interpretation: ε-insensitive loss function. Deviations exceeding ε are penalized in a linear fashion



- The solution is found in terms of dual variables α_i, α_i^* (Lagrange multipliers) conjugate to the constraints:

$$f(x) = \sum_i (\alpha_i - \alpha_i^*) \langle x_i, x \rangle + b$$

  Inside the ε-tube, α_i, α_i^* vanish (KKT condition), so this is a sparse expansion in *support vectors* - hence the name SVM

- Dependence on x enters only through the dot product

Control question

Select all correct answers

1. Support Vector Machines use vectors of data points that support each other in their probabilistic assignments to one of the predicted classes.
2. An epsilon-insensitive loss function only penalizes small deviations from a model-predicted function.
3. An epsilon-insensitive only penalizes (linearly) large deviations, while assigning zero penalties to points within the ε-tube.
4. Unlike Neural Networks, an objective function in SVM is convex and hence has a unique minimum.

Correct answers: 3, 4