Support Vector Machine

By Daniel Rose, Cheyenne Peterson, Zach Bonk, Clara McGrath

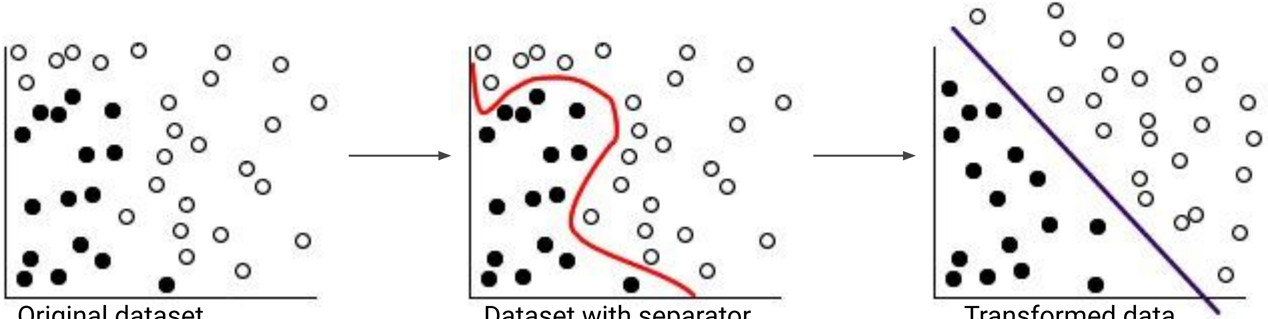
How the algorithm works

- A supervised algorithm used for classification and regression.
- SVM tries to create decision boundaries to separate groups of data. The boundaries can be straight lines, polynomials, or circles.
- Example:

1. Dataset is shown with data points in two different categories.

2. Categories are separated by a curved line.

3. Once transformed using a kernel function, the boundary between the categories can be defined by a hyperplane.



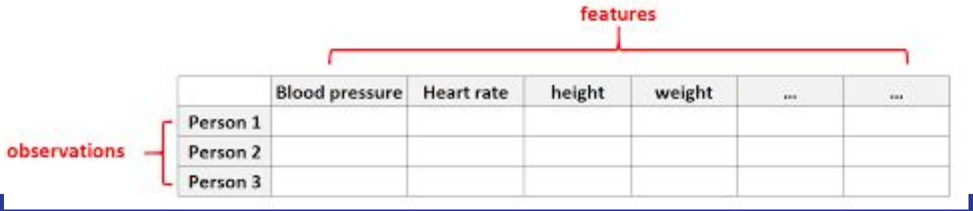
Original dataset

Dataset with separator

Transformed data

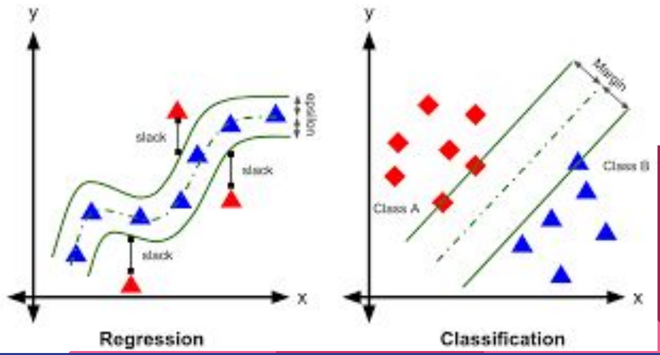
Advantages of SVM

- Many use cases
  - Classification or regression (possible, but uncommonly used for clustering)
  - Images, text, and audio classification
- Handles high-dimensional data
  - Still effective when there are more dimensions than observations
- Very effective when there is a noticeable margin between target classes
- Stable model, small changes have little effect
- Uses L2 (ridge) regularization to avoid overfitting
- Very efficient for small to medium sized datasets
  - Especially with high dimensions



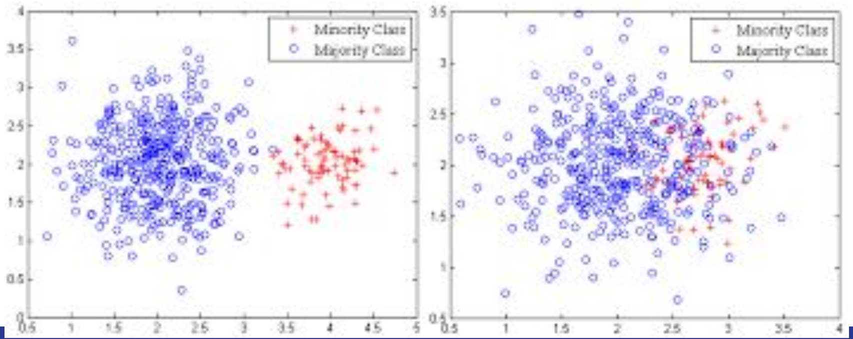
features
observations

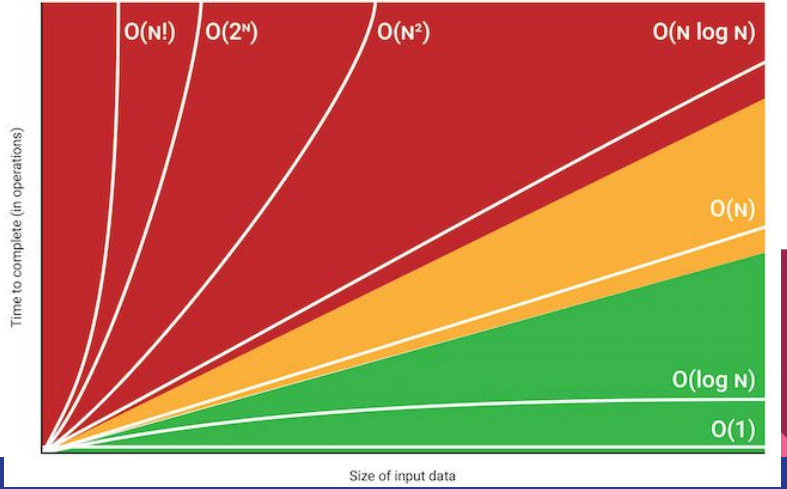
	Blood pressure	Heart rate	height	weight	...	...
Person 1						
Person 2						
Person 3						



Disadvantages of SVM

- Computational inefficient on large datasets
  - Executes at O(n^2) normally, but becomes O(n^3) on large data
  - Can either have a lot of dimensions or a lot of observations
- Executes poorly when dataset has a lot of noise or overlapping target classes
- There is no probability model associated with SVM classification
  - Can't set acceptance thresholds
- Performs poorly on unbalanced datasets





Standardization in SVMs

Recommended to standardize data for SVM

- Scale of input features affects the model and its performance
- Influences the distance from closest points to largest possible margin
    - Data may be skewed towards one side

Missing data

- Difficult for SVM to handle with learning/classification
- Eliminate Missing Values
- Replace MIssing Values
  a. Attribute mean or mode
  b. KNN - distance between example
  c. Train the SVM to impute values (training set has no missing values)

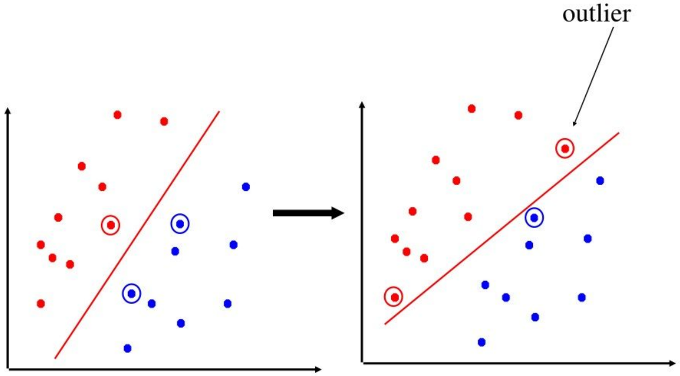
Addressing Outliers

SVM: very sensitive to outliers

- Especially true in training sets
- Certain data points may fall into wrong class because of where the margin falls

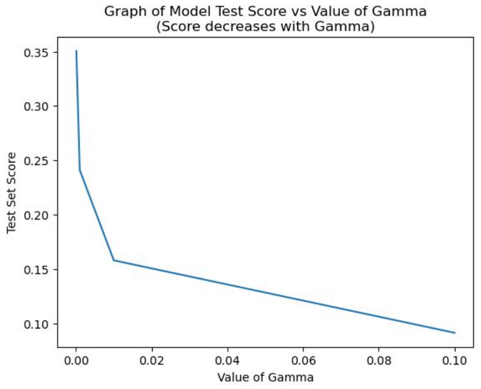
Identify outliers before training

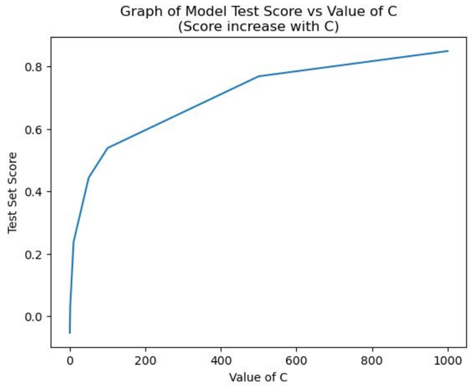
- Remove outliers
- Implement part of SVM to prevent reduction of accuracy based on outliers (i.e., ignore outliers)



What hyperparameters can you tune?

- Kernels - Changes the space of the data
  - The types are linear, polynomial, rbf(radial), and sigmoid
- C - Adds a penalty for misclassified data points.
  - The default value for C is 1
  - The larger the C value increases the penalty but might result in more misclassifications
- Gamma - Changes how far data points affect each other in a radial space
  - Only used in the rbf mode
  - The lower the gamma the less distinct the groups of data are





Appendix/Links

Support vector machine in Machine Learning - GeeksforGeeks

Monkey Learn - Intro to svm

API Reference — scikit-learn 1.2.0 documentation

Support Vector Machine (SVM) in 2 minutes - YouTube