Biometric Authentication of Smartphone Users with Support Vector Machines

Nico Espinosa Dice

April 6, 2020

Presentation Outline

- The Security-Convenience Trade-off
- Background on Biometric Authentication
- Dataset: Biometric Data from Mobile Devices
  - Exploratory Data Analysis
- Mathematics of Support Vector Machine Model
  - Feature Engineering: Hyperparameter Selection
  - Mathematical Reasoning for Selected Hyperparameters
- Support Vector Machine Model Implementation
- Results
- Discussion and Future Research

The Security-Convenience Trade-off

- **Security**: Essential for the safety of users and companies.
  - Security failures pose existential threats to reputation-oriented companies.
- **Convenience**: Crucial for maintaining a high yield of returning users.
  - Lack of convenience drives users to competitors.
- *General Trend*: High security is less convenient.

Background on Biometric Authentication

- **Biometrics**: Metrics (data) acquired from human characteristics.
  - Body measurements and human behavior.
- **Biometric Authentication**: Collect biometric data and use machine learning techniques to authenticate (distinguish) users.
- *Advantages*:
  - Requires little effort from users.
  - High security with convenience.

Dataset: Biometric Data from Mobile Devices

- Provided by: IDSeal, a cybersecurity company.
- Year: 2014.
- Type: Acceleration data.
- Source: Inertial measurement units of smartphones.
- Data collected during "normal device usage" over a period of several months.

Exploratory Data Analysis

Features of Data

- T = time (miliseconds).
- X = acceleration (g) in x direction.
- Y = acceleration (g) in y direction.
- Z = acceleration (g) in z direction.
- DeviceId (Training) = Unique ID of device.
- SequenceId (Testing) = Unique number assigned to each test sample.

Data Shape

- Total Samples: 60,000,000.
- Total Devices: 387.
- Test Data: 90,000 consecutive samples per device.
- Every device had more than 6000 samples.
- Zero-movement periods lasting 10+ seconds were removed.

Support Vector Machine

Chosen because of recommendations in literature (Sitova et al. 2016).

Equation 14.59 of Murphy 2012

$$\hat{y}(\boldsymbol{x}) = sgn\left(\hat{w}_0 + \sum_{i=1}^{N} \alpha_i k(\boldsymbol{x}_i, \boldsymbol{x})\right), \text{ where}$$

$\boldsymbol{x}_i$: support vector (when $\alpha_i > 0$),

$\alpha_i = \lambda_i y_i$,

k : kernel function.

Kernelized SVMs: $O(n_{\text{features}} \times n^2_{\text{observations}})$ complexity (Murphy 2012; Sitova et al. 2016).

Support Vector Machine: Kernel Function

Kernel: Radial Basis Function

$$
\begin{aligned}
k(\boldsymbol{x}, \boldsymbol{x}') &= exp(-\frac{||\boldsymbol{x} - \boldsymbol{x}'||^2}{2\sigma^2}) \\
&= exp(-\gamma||\boldsymbol{x} - \boldsymbol{x}'||^2), \text{ where} \\
\gamma &= \frac{1}{2\sigma^2}, \\
\sigma^2 &: \text{“bandwidth.”}
\end{aligned}
$$

Support Vector Machine: Kernel Function (cont.)

Advantages:

- Outliers have less impact.
- Effective in higher dimensions.
  - Important when adding more features: gyroscope and magnetometer data.
- **Excels when intersection of classes is trivial**.
  - Biometric authentication: no overlap between classes.

Disadvantages:

- Long fitting time.
  - Poses difficulties when cross-validating.
- Difficult to visualize and interpret.

Feature Engineering: Hyperparameter Selection

Hyperparameters of Support Vector Machine

γ (Gamma) $= \frac{1}{\text{Number of Features} \times \text{Variance of Z-Acceleration}}$
Kernel = Radial Basis Function (RBF)
C (Regularization Parameter) = 1
Tolerance for Stopping Criterion $= 1 \cdot 10^{-3}$
Shrinking Heuristic = True

Cross-Validation for C and γ

2-dimensional grid consisting of values:

$$C \in \{2^{-5}, 2^{-3}, ..., 2^{15}\}$$
$$\gamma \in \{2^{-15}, 2^{-13}, ..., 2^{3}\}$$

Cross-validation range empirically determined in literature (Hsu et al. 2009; Murphy 2012).

SVM Model Implementation

Model Implementation Pseudocode

Import data.
Partition dataset into training and testing data.
Scale the data:
    Create a scaler.
    Fit scaler to the training data.
    Transform the training and testing data.
Build the SVM with relevant characteristics.
Fit the SVM to the training data.
Predict the device corresponding to the test samples.
Compute and output accuracy.

Results

Because of the size of the dataset, model was trained and tested on subset of the dataset: 5 devices.

- Model produced 80% accuracy.
- Accuracy:
  - For each test sample: predicted its corresponding device.
  - 'Accuracy' is the percentage of correct predictions.

```
In [85]: svm2.getResults()
Out[85]: 'Accuracy: 0.7904519072701044'
```

Figure 1: Accuracy of support vector machine using radial basis function kernel.

Discussion and Future Research

Discussion:
Support vector machine model shows potential for biometric authentication applications.

Future Research:

- Empirically compare results for SVMs using other kernels: polynomial and sigmoid.
- Empirically optimize Gamma hyperparameter through cross-validation.
- Use dataset with gyroscope and magnetometer sensors.
  - Conduct Principal Component Analysis on features.
- Derive alternative forms of incorporating time series data with support vector machines.
  - Exponential moving average.
  - Arrays of multiple consecutive samples.