Practical Guide to Support Vector Machines

Tingfan Wu
MPLAB, UCSD
Outline

• Data Classification
• High-level Concepts of SVM
• Interpretation of SVM Model/Result
• Use Case Study
What does it mean to learn?

• Acquire new skills?

• Make predictions about the world?
Making predictions is fundamental to survival

Will that bear eat me?

Is there water in that canyon?

Is that person a good mate?

These are all examples of classification problems
Boot Camp Related

Motion classification

face recognition / speaker identification

Brain Computer Interface / Spikes Classification
Driver Fatigue Detection from Facial Expression
Data Classification

- Given training data (class labels known)
- Predicts test data (class labels unknown)
- Not just fitting → generalization

Sensor → 
Data Preprocessing

Classifier
- SVM
- Adaboost
- Neural Network

→ Prediction
Many possible classification models
Which one generalize better?
Generalization
Why SVM ? (my opinion)

• With careful data preprocessing, and properly use of SVM or NN \(\rightarrow\) similar performance.

• SVM is easier to use properly.

• SVM provides a reasonable good baseline performance.
Outline

• Data Classification
• High-level Concepts of SVM
• Interpretation of SVM Model/Result
• Use case study
A Simple Dilemma

Who do I invite to my birthday party?
Problem Formulation

- training data as vectors: $x_i$
- binary labels [ +1, -1]

<table>
<thead>
<tr>
<th>Name</th>
<th>Gift?</th>
<th>Income</th>
<th>Fondness</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Yes</td>
<td>3k</td>
<td>3/5</td>
</tr>
<tr>
<td>Mary</td>
<td>No</td>
<td>5k</td>
<td>1/5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>class</th>
<th>feature vector</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y_1=+1$</td>
<td>$x_1 = [3000, 0.6]$</td>
</tr>
<tr>
<td>$y_2=-1$</td>
<td>$x_2 = [5000, 0.2]$</td>
</tr>
</tbody>
</table>
Vector space

\[ x_2 \text{ (Disposable Income)} \]

\[ x_1 \text{ (Fondness)} \]

- No Gift

+ Gift
A Line

The line: $w^T x + b = 0$

“Hyperplane” in high dimensional space
The inequalities and regions

$w^T x + b = 0$

$w^T x_i + b < 0$

$w^T x_i + b > 0$

Decision function $f(x) = \text{sign}(w^T x_{new} + b)$
Large Margin

A separating hyperplane: \( w^T x + b = 0 \)

\[
(w^T x_i) + b > 0 \quad \text{if } y_i = 1 \\
(w^T x_i) + b < 0 \quad \text{if } y_i = -1
\]
Maximal Margin

Distance between $w^T x + b = 1$ and $-1$:

$$2/\|w\| = 2/\sqrt{w^T w}$$

$$\max 2/\|w\| \equiv \min w^T w / 2$$

$$\min_{w,b} \frac{1}{2} w^T w$$

subject to $y_i((w^T x_i) + b) \geq 1, \quad i = 1, \ldots, l.$
Data not linearly separable

Case 1

Case 2
Trick 1: Soft-Margin

These points are usually outliers. The hyperplane should not bias too much.

\[ \begin{align*}
\min_{w, b} & \quad \frac{1}{2} w^T w + C \sum_i \xi_i \\
\text{subject to} & \quad y_i (w^T x_i + b) \geq 1 - \xi_i \\
& \quad \xi_i \geq 0
\end{align*} \]

\( C \) is a large penalty parameter, most \( \xi_i \) are zero
Soft-margin

[Ben-Hur & Weston 2005]
Support vectors

More important data that support (define) the hyperplane
Trick 2: Map to Higher Dimension

\[ x = [x_1; x_2] \]

Mapping: \( \phi(x) = [x_1^2; x_2] \)

\[
\begin{align*}
\min_{w,b} & \quad \frac{1}{2} w^T w + C \sum_i \xi_i \\
\text{subject to} & \quad y_i (w^T \phi(x)_i + b) \geq 1 - \xi_i \\
& \quad \xi_i \geq 0
\end{align*}
\]
Mapping to Infinite Dimension

• Is it possible to create a universal mapping?
• What if we can map to infinite dimension? Every problem is separable!

• Consider “Radial Basis Function (RBF)”: 
  \[ \phi(x) = e^{-\gamma x^2} [1, \sqrt{\frac{2\gamma}{1!}} x, \sqrt{\frac{(2\gamma)^2}{2!}} x^2, \sqrt{\frac{(2\gamma)^3}{3!}} x^3, \ldots]^T \]

• \[ \phi(x)^T \phi(y) = e^{-\gamma \|x_i - x_j\|^2} = \text{Kernel}(x, y) \]

w : infinite number of variables!

\[
\min_{w,b} \quad \frac{1}{2} w^T w + C \sum_i \xi_i \\
\text{subject to} \quad y_i (w^T \phi(x)_i + b) \geq 1 - \xi_i \\
\xi_i \geq 0
\]
Dual Problem

- **Primal**

  \[
  \min_{w, b} \quad \frac{1}{2} w^T w + C \sum_i \xi_i \\
  \text{s.t.} \quad y_i (w^T \phi(x)_i + b) \geq 1 - \xi_i \\
  \xi_i \geq 0
  \]

- **Dual**

  \[
  \min_{\alpha} \quad \frac{1}{2} \alpha^T Q \alpha - \sum_i \alpha_i \\
  \text{where} \quad Q_{ij} = y_i y_j \phi(x_i)^T \phi(x_j) \\
  \text{s.t.} \quad \sum \alpha_i y_i = 0 \\
  0 \leq \alpha_i \leq C
  \]

  \[
  \phi(x_i)^T \phi(x_j) = e^{-\gamma |x_i - x_j|}
  \]

  finite calculation

  \[
  w = \sum_{i=1}^l \alpha_i y_i \phi(x_i)
  \]
Gaussian/RBF Kernel

\[ \phi(x_i)^T \phi(x_j) = e^{-\gamma|x_i - x_j|} = e^{-dist(x_i, x_j)} = \text{similarity}(x_i, x_j) \]

\[
\begin{align*}
\gamma &= 10^0 \\
\gamma &= 10^1 \\
\gamma &= 10^4 
\end{align*}
\]

\[
\begin{array}{c}
\sim \text{linear kernel} \\
\text{Overfitting nearest neighbor?}
\end{array}
\]
Recap

\[
\begin{align*}
\text{min}_{w, b} & \quad \frac{1}{2} w^T w + C \sum_i \xi_i \\
\text{s.t.} & \quad y_i (w^T \phi(x)_i + b) \geq 1 - \xi_i \\
& \quad \xi_i \geq 0
\end{align*}
\]

\[\phi(x_i)^T \phi(x_j) = e^{-\gamma |x_i - x_j|}\]
Checkout the SVMToy

  - -c (cost control softness of the margin/#SV)
  - -g (gamma controls the curvature of the hyperplane)
Cross Validation

• What is the best \((C, \gamma)\)? \(\rightarrow\) Date dependent
• Need to be determined by “testing performance”

• Split training data into pseudo “training, testing” sets

- Exhausted grid search for best \((C, \gamma)\)
Outline

• Machine Learning → Classification
• High-level Concepts of SVM
• Interpretation of SVM Model/Result
• Use Case Study
(1) Decision value as strength

Decision function \( f(x) = \text{sign}(w^T x_{\text{new}} + b) \)
Facial Movement Classification

- Classes: brow up(+) or down(-)
- Features: pixels of Gabor filtered image
Decision value as strength

Probability estimates from decision values also available
(2) Weight as feature importance

- Magnitude of weight: feature importance
- Similar to regression

1C Inner brow raise

au1
(3) Weights as profiles

Fluorescent image of cells of various dosage of certain drug

Various image-based features

Clustering the weights shows the primal and secondary effect of the drug

Lit-Hsin Loo, Lani F Wu & Steven J Altschuler
Outline

• Machine Learning → Classification
• High-level Concepts of SVM
• Interpretation of SVM Model/Result
• User Case Study
The Software

• SVM requires an constraint quadratic optimization solver
  → not easy to implement.

• Off-the-shelf Software
  – libsvm by Chih-Jen Lin et. al.
  – svm\textsuperscript{light} by Thorsten Joachims

• Incorporated into many ML software
  – matlab / pyML / R…
Beginners may…

1. Convert their data into the format of a SVM software.
2. May **not** conduct **scaling**
3. Randomly try few parameters and **without cross validation**
4. Good result on training data, but poor in testing.
Data scaling

Without scaling
– feature of large dynamic range may dominate separating hyperplane.

<table>
<thead>
<tr>
<th>label</th>
<th>X</th>
<th>Height</th>
<th>Gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>y1=0</td>
<td>x1</td>
<td>150</td>
<td>2</td>
</tr>
<tr>
<td>y2=1</td>
<td>x2</td>
<td>180</td>
<td>1</td>
</tr>
<tr>
<td>y3=1</td>
<td>x3</td>
<td>185</td>
<td>1</td>
</tr>
</tbody>
</table>
Parameter Selection

Contour of cross validation accuracy.
User case : Astroparticle scientist

• User:
  I am using libsvm in a astroparticle physics application .. First, let me congratulate you to a really easy to use and nice package. Unfortunately, it gives me astonishingly bad test results...

• OK. Please send us your data We are able to get 97% test accuracy. Is that good enough for you?

• User:
  You earned a copy of my PhD thesis
Dynamic Range Mismatch

• A problem from astroparticle physics

```
<label> <index>::<value> <index>::<value> ...
1 1:2.6173e+01 2:5.88670e+01 3:-1.89469e-01 4:1.25122e+02
1 1:5.7073e+01 2:2.21404e+02 3:8.60795e-02 4:1.22911e+02
1 1:1.7259e+01 2:1.73436e+02 3:-1.29805e-01 4:1.25031e+02
1 1:2.1779e+01 2:1.24953e+02 3:1.53885e-01 4:1.52715e+02
1 1:9.1339e+01 2:2.93569e+02 3:1.42391e-01 4:1.60540e+02
1 1:5.5375e+01 2:1.79222e+02 3:1.65495e-01 4:1.11227e+02
1 1:2.9562e+01 2:1.91357e+02 3:9.90143e-02 4:1.03407e+02
```

• #Training set 3,089 and #testing set 4,000
• Large dynamic range of some features.
Overfitting

- Training
  
  $./svm-train train.1  (default parameter used)
  
  optimization finished, #iter = 6131
  
  nSV = 3053, nBSV = 724
  
  Total nSV = 3053

- Training Accuracy
  
  $./svm-predict train.1 train.1.model o
  
  Accuracy = 99.7734% (3082/3089)

- Testing Accuracy
  
  $./svm-predict test.1 train.1.model test.1.out
  
  Accuracy = 66.925% (2677/4000)

  nSV and nBSV: number of SVs and bounded SVs (i = C).

  Without scaling. One feature may dominant the value overfitting

- 3053/3089 training data become support vector \(\rightarrow\) Overfitting
- Training accuracy high, but low testing accuracy \(\rightarrow\) Overfitting
Suggested Procedure

• Data pre-scaling
  – scale range [0 1] or unit variance
• Using (default) Gaussian(RBF) kernel
• Use cross-validation to find the best parameter (C, γ)
• Train your model with best parameter
• Test!

All above done automatically in “easy.py” script provided with libsvm.
Large Scale SVM

• (#training data >> #feature ) and linear kernel
  – Use primal solvers (eg. liblinear)
• To approximated result in short time
  – Allow inaccurate stopping condition
    svm-train –e 0.01
  – Use stochastic gradient descent solvers
    24
Resources

• LIBSVM: http://www.csie.ntu.edu.tw/~cjlin/libsvm
• LIBSVM Tools: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools
• Kernel Machines Forum: http://www.kernel-machines.org
• Hsu, Chang, and Lin: A Practical Guide to Support Vector Classification
• my email: tfwu@ucsd.edu

• Acknowledgement
  – Many slides from Dr. Chih-Jen Lin, NTU