Application of Support Vector Machine In Bioinformatics

V. K. Jayaraman
Scientific and Engineering Computing Group
CDAC, Pune
jayaramanv@cdac.in

Arun Gupta
Computational Biology Group
AbhyudayaTech, Indore
arun@abhyudayatech.com
Introduction continue....

Given a set of sequences

Can we Identify Allergen causing Sequences?
Introduction continue....

Given a set of proteins

Can we do the structure based classification?
SEQUENCES

Extract various features

Implement SVM

Class I

Class II
SUPPORT VECTOR CLASSIFICATION ALGORITHM
SVM classification algorithm

- Introduced by Vapnik and co-workers in 1992
- Rigorously based on
  - Statistical Learning Theory
  - Perceptron

- SVM Classifier can Recognize Patterns efficiently
- Solve Real Life Problems in:
  - Chemo/Bio Informatics
  - Many many other fields
• What is the importance of identifying disordered regions in proteins..?
• Dunker et al. have predicted the disordered regions in proteins only on the basis of hydrophobicity and charge.
• We shall now study how this can be done by supervised and unsupervised classification.
Supervised or Unsupervised

Unsupervised

Supervised
Supervised or Unsupervised

- **Unsupervised**: Provided a set of features algorithm groups data into clusters (does not require class information).

- **Supervised**: Employs class label information of some instances to build a model. Model is validated employing unseen data.
**Unsupervised Classification**

- Algorithm will cluster the data points into groups
- **Input**

<table>
<thead>
<tr>
<th>Mean Hydrophobicity</th>
<th>Net Charge</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.68</td>
<td>-7</td>
</tr>
<tr>
<td>-1.315</td>
<td>1</td>
</tr>
<tr>
<td>-1.464</td>
<td>0</td>
</tr>
<tr>
<td>-1.961</td>
<td>2</td>
</tr>
<tr>
<td>-2.003</td>
<td>3</td>
</tr>
<tr>
<td>-1.594</td>
<td>-7</td>
</tr>
</tbody>
</table>

- Class information need not be provided
## Supervised Classification

### Input

<table>
<thead>
<tr>
<th>Mean Hydrophobicity</th>
<th>Net Charge</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.68</td>
<td>-7</td>
<td>1</td>
</tr>
<tr>
<td>-1.315</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1.464</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>-1.961</td>
<td>2</td>
<td>-1</td>
</tr>
<tr>
<td>-2.003</td>
<td>3</td>
<td>-1</td>
</tr>
<tr>
<td>-1.594</td>
<td>-7</td>
<td>-1</td>
</tr>
</tbody>
</table>

- Algorithm will be trained on a set of given data points.
- This algorithm will classify unseen data points into classes.
- Class information is needed for training.
SVM Classification

• SVM can be employed for both supervised and unsupervised classification
• Supervised classifications is more popular
• Entire lecture Is based on “supervised methodology”
General Approach

Train classifier to give minimum error

Test on the unknown data
Algorithm contd...

<table>
<thead>
<tr>
<th>Mean Hydrophobicity</th>
<th>Net Charge</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.68</td>
<td>-7</td>
<td>1</td>
</tr>
<tr>
<td>-1.315</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>-1.464</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>-1.961</td>
<td>2</td>
<td>-1</td>
</tr>
<tr>
<td>-2.003</td>
<td>3</td>
<td>-1</td>
</tr>
<tr>
<td>-1.594</td>
<td>-7</td>
<td>-1</td>
</tr>
</tbody>
</table>

Test Data

<table>
<thead>
<tr>
<th>Mean Hydrophobicity</th>
<th>Net Charge</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.153</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>-1.464</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>-3.300</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>-2.945</td>
<td>-7</td>
<td></td>
</tr>
</tbody>
</table>

SVM Model

Predictions

1
-1
-1
1
....
Algorithm contd..

SVM employs a linear hyperplane

\[ [w^T x] + b = 0 \]

- **W** - weight vector
- **b** - bias

**W** - vector : having dimensions equivalent to number of features

**X** : \(x_1, x_2, x_3, \ldots\) Input Vector

**y** : +1 class 1

**y** : -1 class 2

<table>
<thead>
<tr>
<th>Mean Hydrophobicity</th>
<th>Net Charge</th>
<th>...</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.68</td>
<td>-7</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>-1.315</td>
<td>1</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>-1.464</td>
<td>0</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>-1.961</td>
<td>2</td>
<td></td>
<td>-1</td>
</tr>
<tr>
<td>-2.003</td>
<td>3</td>
<td></td>
<td>-1</td>
</tr>
<tr>
<td>-1.594</td>
<td>-7</td>
<td></td>
<td>-1</td>
</tr>
</tbody>
</table>

Disordered

Normal
Algorithm contd....

Which Hyperplane is Better?

(a) (b)

(c) (d)
Algorithm contd..

**Maximum Margin Classifier**

- Constrain the data belonging to two different classes to be at least distance ‘1’ from the separating hyperplane
- Minimize the risk of overfitting by choosing the maximal margin hyperplane in feature space

\[
[w^T x] + b = +1
\]

\[
[w^T x] + b = 0
\]

\[
[w^T x] + b = -1
\]

Margin = \[ \frac{1}{2}||w||^2 \]

Maximize Margin
Limitations of Linear Classifier

- Linear classifier is not always the winner.
Map the data into a feature space where they are linearly separable.
Learning in the Feature Space

<table>
<thead>
<tr>
<th>Mean Hydrophobicity</th>
<th>Net Charge</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.68</td>
<td>-7</td>
<td>1</td>
</tr>
<tr>
<td>-1.464</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>-1.961</td>
<td>2</td>
<td>-1</td>
</tr>
<tr>
<td>-2.003</td>
<td>3</td>
<td>-1</td>
</tr>
</tbody>
</table>

Can we have a function that transforms lower dimensional non-Separable data into higher dimensional separable data

\[ x \overset{\phi}{\rightarrow} \phi(x) \]

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>6.01</td>
<td>3.9</td>
<td>2.5</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>0.17</td>
<td>2.0</td>
<td>8.0</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>3.10</td>
<td>5.6</td>
<td>3.6</td>
<td>-1</td>
</tr>
<tr>
<td></td>
<td>5.0</td>
<td>4.3</td>
<td>4.7</td>
<td>-1</td>
</tr>
</tbody>
</table>
Algorithm contd..

Learning in the Feature Space

\[ x \xrightarrow{\phi} \phi(x) \]
SVM Linear Classifier in High Dimensional Space

Algorithm contd..

\[ [w^T \phi(x)] + b = -1 \]
\[ [w^T \phi(x)] + b = 0 \]
\[ [w^T \phi(x)] + b = 1 \]

Support Vectors
H Dim. feature space

• Real life problems have several input features
  – Gene Expression Profiles $\rightarrow$ Thousands of Genes
  – Drug Discovery $\rightarrow$ More than 100 thousand Descriptors

• Eg. 256 dimensional data, polynomial of degree 5 gives the feature space dimension $\approx 10^{10}$
Introduction to Kernel functions

- Working in high dimensional feature spaces solves the problem of expressing complex functions
  BUT....
  - Computationally intractable
  - Data Dimensionality increases exponentially in the feature space

SOLUTION.......

- Introduce kernel functions for simplification
The “Kernel Trick”

- The linear classifier relies on dot product between vectors
  \[ K(x_i, x_j) = x_i^T x_j \]

- If every data point is mapped into high-dimensional space via some transformation \( \Phi: x \rightarrow \varphi(x) \), the dot product becomes:
  \[ K(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) \]

- A kernel function is some function that corresponds to an inner product in some expanded feature space.
Commonly-used kernel functions

- Linear kernel: \[ K(x_i, x_j) = x_i^T x_j \]

- Polynomial kernel: \[ K(x_i, x_j) = (1 + x_i^T x_j)^p \]

- Gaussian (Radial-Basis Function (RBF)) kernel:
  \[ K(x_i, x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \]

- Sigmoid:
  \[ K(x_i, x_j) = \tanh(\beta_0 x_i^T x_j + \beta_1) \]
Choosing the Kernel Function

- Probably the most tricky part of using SVM.
- The kernel function is important because it creates the kernel matrix, which summarizes all the data.
- In practice, a *low degree polynomial kernel* or *RBF kernel* with a reasonable width is a good initial try.
- Note that SVM with RBF kernel is closely related to RBF neural networks, with the centers of the radial basis functions automatically chosen for SVM.
Applications of SVM in BI

Identification of protein functions
Gene functions
Micro array Classification
Identification of protein function

• Secondary structure prediction
• Identification of binding sites
• Sub nuclear localization of proteins
• Sub cellular localization
• Protein-protein interaction prediction
• Prediction of protein disorder
Identification of gene functions

- Promoter prediction
- Prediction of tissue specific localization of genes
- Prediction of DNA methylation sites
- DNA hot spots prediction
Microarray Classification

- Lukemia prediction
- Colon cancer prediction
- Prediction of several genetic disorders
- No of examples less & No Of Features very Large.
- Employ Feature Selection
Domain features extraction

Protein function Identification :

- Numerical representation of the sequence.
- Amino acid frequencies
- Dipeptide frequencies
- Tripeptide frequencies
- K-mer frequencies
- Homology information in terms of Blast and Psi-blast profiles.
- Remote Homology using PSSM
- Motive information
- Secondary Structure information
- Physical properties like hydrophobic, hydrophilic, charge, etc.
- Structural features like surface accessibility, co-ordinates of atoms, contact order.
Features extraction

Protein function Identification:

Numerical representation of the sequence.

Alanine → 1 0 0 0 0 0 . . . .

Glycine → 0 0 1 0 0 0 . . . .

can be done for 20 amino acids
Protein function Identification:

Amino acid frequencies
Features extraction

Protein function Identification:

Secondary structure information:

- Helix
- Sheets
- Turns
Feature Calculation

Training sequences

<table>
<thead>
<tr>
<th>seq1</th>
<th>seq2</th>
<th>seq3</th>
</tr>
</thead>
<tbody>
<tr>
<td>AASQRLAQS</td>
<td>ASSNRADSQ</td>
<td>ADSQRADSQ</td>
</tr>
</tbody>
</table>

Input features (X)

Class labels (Y)

<table>
<thead>
<tr>
<th></th>
<th>A_freq</th>
<th>C_freq</th>
<th>AA_freq</th>
<th>AC_freq</th>
<th>PSSM1</th>
<th>PSSM2</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.33</td>
<td>0.00</td>
<td>...</td>
<td>0.125</td>
<td>0.000</td>
<td>0.754</td>
<td>0.000</td>
</tr>
<tr>
<td>0.22</td>
<td>0.00</td>
<td>...</td>
<td>0.00</td>
<td>0.00</td>
<td>0.754</td>
<td>0.012</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Class

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>-1</th>
<th>...</th>
</tr>
</thead>
</table>

...
Commonly used Methods for Parameter Tuning

- k-fold cross validation
- Leave-one-out error estimation
k-fold Cross Validation

- The training data is randomly split into $k$ mutually exclusive subsets (or the folds) of approximately equal size.
- For $k$ times:
  - SVM decision rule is obtained using $k - 1$ of the subsets
  - Then tested on the subset left out
3-fold Cross Validation

**Total Data**

**Fold 1**

**Fold 2**

**Fold 3**

**SVM**

- **Testing Data**
- **Training Data**
3-fold Cross Validation
3-fold Cross Validation

Total Data

Fold 1

Fold 2

Fold 3

Testing Data

Training Data

SVM

Error
3-fold Cross Validation

Validation error = Average of three errors
Leave-one-out error (LOO) estimate

- Extreme form of $k$-fold cross-validation
  - $k$ is equal to the number of examples, $I$
- For $I$ times
  - SVM decision rule is obtained using $I-1$ of the examples
  - Then tested on the subset left out example
- Good thing about LOO
  - Almost unbiased estimate of the expected generalization error
- Limitation
  - Computationally expensive since computations require running the training algorithm $I$ times.
Model Evaluation Measures

Sensitivity = \frac{TP}{TP + FN}

Specificity = \frac{TN}{TN + FP}

Precision = \frac{TP}{TP + FP}

Matthew’s Correlation Coefficient

\[ MCC(X) = \frac{TPTN - FPFN}{\sqrt{(TN + FN)(TP + FN)(TN + FP)(TP + FP)}} \]
SVM based prediction servers

Cyclin Pred
http://bioinfo.icgeb.res.in/cyclinpred/ CyclinPred is a SVM based prediction method to identify novel cyclins using various features of proteins.

SS PRED

Bayes Server

NRpred
http://www.imtech.res.in/raghava/lgepred/ This server allows user to analyse the expression data (Microarray Data).

UbiPred
http://iclab.life.nctu.edu.tw/ubipred/ UbiPred is a SVM-based prediction server using that detects the presence /absence of ubiquitylation site in a protein sequence.
OUR WORK

• HLA-B27 is found to be associated with the development of variety of autoimmune diseases including Ankylosing spondylitis.

• Several theories have been proposed to explain the association of HLA-B27 with spondyloarthritis.

• HLAB27Pred will be helpful in designing new peptide vaccines through the prediction of corresponding binding peptides.

UnPublished (under Review)
HLAb27Pred is a server designed for the prediction of HLA-B*2705 (MHC class I allele) based nanomeric epitopic binding peptides. Server implements 2 techniques for the purpose of prediction, viz. SVM and PSSM.

- SVM based prediction are deployed by training a set of experimentally validated nanomeric binding and non-binding peptides.
- The performance of the SVM predictions has been tested through 5 cross-validation.
- The **specificity** and **sensitivity** obtained during the development of this server is **84.54%** and **85.57%** respectively.
- Whereas average **precision** and average **MCC** values were observed to be **84.69%** and **0.8%** respectively.
HLAb27Pred

HLA-B27 is found to be associated with the development of variety of autoimmune diseases including Ankylosing spondylitis. Several theories have been proposed to explain the association of HLA-B27 with spondylarthrits. HLAB27Pred will be helpful in designing new peptide vaccines through the prediction of corresponding binding peptides.

Submit protein sequence(s) for prediction

Prediction Name

Upload Sequence file

Sequence: (Type/paste FASTA format amino acid sequences) *Example sequence*

Prediction Settings:
- Method: SVM
- Threshold: 0.6
- Display top: 10 Peptides
- Output: Enriched Tabular

Run Prediction  Reset
Software to try your hands on

- SVMLight
- LibSVM
- WEKA and Bio-Weka
- MATLAB
Questions and/or Comments...?
Thank You...