Support Vector Machines: A Useful Tool for Classification

Shibasish Dasgupta
Department of Statistics
University of Florida
Aim

- To look at the technique of Support Vector Machine (SVM) for classification from the Machine Learning as well as Statistical point of view

- Search for the theoretical results that might demonstrate specific quantitative advantages of SVM

- Compare the performances of SVM and the traditional statistical techniques on different types of datasets and try to draw conclusions
Outline

- Problem of Classification
- Traditional Statistical Techniques for Classification
- Machine Learning Approach to SVM
- SVM from a statistician’s perspective
- Experimental Analysis
- Results of Experiments
- Conclusion
Problem of Classification

- Classification
  - Allocating new objects to previously defined groups on the basis of the measurements taken
  - Training and Test data
  - Two-class classification problem

- Statistical Approach
  - Problem of “statistical decision functions”
  - Probability distributions of the measurements are assumed completely known or the parameters of the distribution must be estimated from a sample from that population

- A statistician always wants to minimize the errors that he/she makes while classifying a new object
Standards of Good Classification

A good classification procedure is one that minimizes in some sense or other the costs of misclassification.

Given the probability distributions and the prior probabilities of the two populations, a procedure that minimizes the Expected Cost of Misclassification is called a “Bayes Procedure.”
The Bayes rule that minimizes the expected misclassification rate is $\text{sign}(p(x)-1/2)$, where $p(x)=P[Y=1\mid X=x]$

- The Bayes rule is the optimal classification rule if the underlying distribution of the data is known

Classification into one of two multivariate normal populations with equal dispersion matrices
- Linear Discriminant Analysis (LDA)

Classification into one of two multivariate normal populations with unequal dispersion matrices
- Quadratic Discriminant Analysis (QDA)

LDA and QDA are optimal classification procedures in the Bayes sense

In practice, we do not know the underlying distribution, and need to “learn” classification rules from the data
Machine Learning Approach to SVMs

- Linear And Separable SVM
  - SVM classification problem can be defined as:
    \[
    \min_{w,b} \frac{1}{2} \|w\|^2
    \]
    \[
    s.t \quad y_i (w \cdot x_i + b) \geq 1, \quad \forall i
    \]

- Geometrically the problem is finding a separating hyperplane such that
  - Data from one class has positive sign and the other class has negative sign, when evaluated on the hyperplane
  - The distance of the hyperplane from the closest training point is maximized
Linearly Nonseparable Case

- Data may be such that it is not separable by a linear hyperplane
  - Can be handled by allowing and penalizing classification error ($C$)
  - Using nonlinear separator boundary

Find a line that penalizes points on “the wrong side”.
The Lagrangian trick

Reformulate the optimization problem:
A “trick” often used in optimization is to do an Lagrangian formulation of the problem. The constraints will be replaced by constraints on the Lagrangian multipliers and the training data will occur only as dot products.

\[
\text{Max } L = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j x_i \cdot x_j,
\]
Subject to:

\[
\begin{align*}
    w &= \sum \alpha_i y_i x_i \\
    \sum \alpha_i y_i &= 0
\end{align*}
\]

What we need to see: $x_i$ and $x_j$ (input vectors) appear only in the form of dot product – we will soon see why that is important.
Nonlinear SVMS

- Nonlinear separations handled well by using kernel techniques
  - Map the original dataset into a higher dimensional set
  - Mapping to high enough dimensions can produce good separation
**Kernel Function**

- **Definition:** A kernel is a function $K$, such that for all $x, z \in X$,
  $$K(x, z) = \langle \Phi(x) \cdot \Phi(z) \rangle$$
  where $\Phi$ is a mapping from the input space $X$ to an feature space $F$.

- **Types Of Kernel:** Four common kernels are
  1. **Linear:** $K(x, y) = x \cdot y$
  2. **Polynomial:** $K(x, y) = (\gamma x \cdot y + r)^d$, $\gamma > 0$, $r \in \mathbb{R}$, $d = 1, 2, \ldots$
  3. **Radial Basis Function (RBF):** $K(x, y) = e^{-\gamma \|x - y\|^2}$, $\gamma > 0$
  4. **Sigmoid:** $K(x, y) = \tanh(\gamma x \cdot y + r)$, $\gamma, r \in \mathbb{R}$

Here, $\gamma, r$ and $d$ are **kernel parameters**.

In general RBF is a reasonable choice.
Non-linear svm

The function we end up optimizing is:
\[ \text{Max Ld} = \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j K(x_i \cdot x_j), \]
Subject to:
\[ w = \sum \alpha_i y_i x_i \]
\[ \sum \alpha_i y_i = 0 \]
Classification Using SVM

The SV machine maps the input space into a high-dimensional feature space and then constructs an Optimal hyperplane in the feature space.
SVM from a Statistician’s Perspective

- In practice, we do not know the underlying probability distribution, but are only given a sample from it
  - Bayes rule can not be implemented
  - Find a classification rule whose performance is close to that of the Bayes rule, or to apply the Bayes rule approximately

- Features $x_i \in \mathbb{R}^d$, class labels $y_i \in \{-1, 1\}$, $H$ a reproducing kernel Hilbert space (RKHS), then find $f(x) = h(x) + b$ with $h \in H$ minimizing

$$
\frac{1}{n} \sum_{i=1}^{n} (1 - y_i f(x_i))_+ + \lambda \|h\|^2
$$

The classification rule is $\Phi(x) = \text{sign}(f(x))$
For random variables $X$ and $Y$ representing features and labels, respectively, the classification error rate of a given classification rule $\Phi: X \rightarrow Y$ is $E[I(\Phi(X) \neq Y)]$, where $I$ is the indicator function.

Vapnik (1998) showed that SVM minimizes the probability of misclassifying yet-to-be-seen patterns for a fixed but unknown probability distribution of the data based on upper bounds on classification error.

Lin (2002) showed that SVM classifier is asymptotic to the Bayes decision rule, which is the theoretically best classifier in terms of minimizing classification error.
Procedures for SVM

1. Transform data to the format of an SVM software
2. Consider simple scaling on the data
3. Choose a suitable kernel function (Gaussian/RBF in our case)
4. Use cross-validation to find the close to optimal parameters
5. Use the parameters to train the model
6. Classify observations from the test data
Experimental Setup

100 Random Realizations of the data

60% as a Training set

40% as a Test set

Cross-Validation

Linear and Quadratic Classifier

SVM Classifier

TEST

Averaged Misclassification Rate (in %) (with S.D)
Breast-Cancer Dataset

- **Number of examples:** 277
- **Number of attributes:** 9 (age, tumor-size, menopause, etc.)
- **Number of Classes:** 2 (recurrence, no-recurrence)
- **Class Distribution:** no-recurrence events: 196 instances
  recurrence events: 81 instances
- **200 train and 77 test observations**
- **Source:** University Medical Centre, Institute of Oncology, Ljubljana, Yugoslavia.
## Results

<table>
<thead>
<tr>
<th>Classification Techniques</th>
<th>Class. Error Mean ± s.d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>31.80 ± 4.35</td>
</tr>
<tr>
<td>QDA</td>
<td>31.36 ± 4.91</td>
</tr>
<tr>
<td>SVM</td>
<td>27.22 ± 4.74</td>
</tr>
</tbody>
</table>
Comments

- Presence of categorical and qualitative characteristics and hence no question of normality
  - SVM performed better than LDA and QDA

- LDA and QDA are very fast methods compared to SVM
  - SVM needs proper parameter tuning and solving qpp

- Inversion of sample covariance matrices sometimes creates problem in LDA and QDA
  - SVM does not need any inversion of matrix

- SVM performed better compared to others in terms of misclassification rate
German Credit Dataset

- **Number of examples:** 1000

- **Number of attributes:** 20 (Duration in month, Credit amount, Age in years, Number of existing credits at this bank, etc.)

- **Number of Classes:** 2 (good & bad Customers)
  - Class 1: 700 (good credit)
  - Class 2: 300 (bad credit)

- 700 train and 300 test observations

- **Original Source:**
  Professor Dr. Hans Hofmann
  Institut für Statistik und Ökonometrie Universität Hamburg
Results

<table>
<thead>
<tr>
<th>Classification Techniques</th>
<th>Class. Error Mean</th>
<th>s.d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>28.51 +/- 2.55</td>
<td></td>
</tr>
<tr>
<td>QDA</td>
<td>30.01 +/- 2.79</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>23.79 +/- 2.17</td>
<td></td>
</tr>
</tbody>
</table>

- Comments
  - Example of perfectly non-normal data
  - SVM gives better result compared to others

- What about the two types of error encountered in any classification problem?
Discussion 1: Breast-Cancer Dataset

- No-recurrence class dominates the recurrence class
- Classifying a cancer patient in no-recurrence class is dangerous than the other classification error
- $e(2 \mid 1)$: Average classification error (in %) for classifying a patient in recurrence class when she is actually from no-recurrence class
- $e(1 \mid 2)$: Average classification error (in %) for classifying a patient in no-recurrence class when she is actually from recurrence class
- $e(1 \mid 2)$ is more serious than $e(2 \mid 1)$

<table>
<thead>
<tr>
<th>Methods</th>
<th>$e(2 \mid 1)$</th>
<th>$e(1 \mid 2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>30.35</td>
<td>48.97</td>
</tr>
<tr>
<td>LDA</td>
<td>29.14</td>
<td>38.48</td>
</tr>
<tr>
<td>QDA</td>
<td>24.75</td>
<td>47.86</td>
</tr>
</tbody>
</table>
Discussion 2: German Credit Dataset

- The class good credit dominates over the others
- It is worse to class a customer as good when they are bad, than it is to class a customer as bad when they are good
- \( e(2 \mid 1) \): Average classification error (in %) for classifying a customer as bad when they are good
- \( e(1 \mid 2) \): Average classification error (in %) for classifying a customer as good when they are bad
- \( e(1 \mid 2) \) is more serious than \( e(2 \mid 1) \)

<table>
<thead>
<tr>
<th>Methods</th>
<th>( e(2 \mid 1) )</th>
<th>( e(1 \mid 2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>23.09</td>
<td>41.79</td>
</tr>
<tr>
<td>LDA</td>
<td>28.39</td>
<td>28.86</td>
</tr>
<tr>
<td>QDA</td>
<td>29.08</td>
<td>32.15</td>
</tr>
</tbody>
</table>
Comments

- For both the cases LDA’s performance is much better than the other two methods.

- SVM has a tendency to commit the serious error as well as QDA though in terms of total misclassification rate SVM has given better results.

- These types of circumstances can restrict us to use SVM in such situations.
Iris Dataset

- **Description:** The data contains 3 classes of 50 instances each, where each class refers to a type of Iris plant.
- **Attribute Information:** Sepal length, sepal width, petal length, and petal width (all in cm.).
- **Class:** Iris Setosa, Iris Versicolor, and Iris Virginica.

- We consider two classes at a time and perform 2-class classification.
- The sample of these classes came from approximately multivariate normal populations.
- In the three classification problems, the number of training and test samples are 60 and 40 respectively.
Fig. Scatter plot of Iris data with respect to sepal length and sepal width
Comments

- Iris-Setosa is linearly separable than Iris-Virginica as well as Iris-Versicolor
- LDA and QDA performed best over SVMs
- LDA perform well on small datasets
- LDA perform well in a situation when two classes are linearly separable
Iris-Versicolor vs. Iris- Virginica

- **Results:**

<table>
<thead>
<tr>
<th>Classification Techniques</th>
<th>Class. Error</th>
<th>Mean</th>
<th>s.d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>3.52 +/- 2.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>QDA</td>
<td>3.90 +/- 2.84</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td>5.30 +/- 3.32</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Comments**
  - The two classes are not linearly separable
  - LDA is giving better result followed by QDA and SVM respectively
  - LDA performed consistently well over all the 100 realizations (smallest dispersion)

How these techniques will perform when the distribution of 2 classes are perfectly multivariate normal?
Simulated Datasets

- **Case 1: Two Normal Populations with Same Dispersion Matrices**
- **Description**
  - We have generated two samples of 250 observations each from two 9-variate normal distributions of equal covariance matrices
  - Number of train and test observations are 300 and 200 respectively
- **Results:**

<table>
<thead>
<tr>
<th>Classification Techniques</th>
<th>Class. Error</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
</tr>
<tr>
<td>LDA</td>
<td>21.02 +/- 2.52</td>
</tr>
<tr>
<td>QDA</td>
<td>23.33 +/- 2.64</td>
</tr>
<tr>
<td>SVM</td>
<td>21.53 +/- 2.53</td>
</tr>
</tbody>
</table>

- **Comments**
  - This is the situation where LDA is supposed to be the optimal classification rule in the Bayes sense
  - LDA gives the best performance followed by SVM
Simulated Datasets

- **Case 2: Two Normal Populations with Unequal Dispersion Matrices**
- **Description**
  - We have generated two samples of 200 observations each from two different 9-variate Normal distributions of unequal covariance matrices
  - 240 train and 160 test observations
- **Results:**

<table>
<thead>
<tr>
<th>Classification Techniques</th>
<th>Class. Error Mean ± s.d</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>33.49 ± 3.13</td>
</tr>
<tr>
<td>QDA</td>
<td>0.93 ± 0.63</td>
</tr>
<tr>
<td>SVM</td>
<td>2.59 ± 1.04</td>
</tr>
</tbody>
</table>

- **Comments**
  - This is the situation where QDA is supposed to be the optimal classification rule in the Bayes sense
  - QDA simply outperforms all the other methods followed by SVM
Conclusion

- SVMs do not need any distributional assumptions and performs well on non-normal data with reasonably large training sets
  - Most real-life datasets are non-normal
  - In the presence of continuous and some qualitative variables the traditional techniques do not perform satisfactorily
- SVMs need to be modified in non-standard situations
- In high-dimensional spaces LDA and QDA seldom succeed in practice
- QDA is very sensitive to departure from normality
  - Usual test for covariance homogeneity are greatly affected by non-normality
- LDA and QDA is more simple, quick and straightforward than Support Vector Classification
Thank You!