Lecture 18: Multiclass Support Vector Machines

Hao Helen Zhang
Outlines

- Traditional Methods for Multiclass Problems
  - One-vs-rest approaches
  - Pairwise approaches
- Recent development for Multiclass Problems
  - Simultaneous Classification
  - Various loss functions
- Extensions of SVM
Multiclass Classification Setup

- Label: \( \{-1, +1\} \rightarrow \{1, 2, \ldots, K\} \).
- Classification decision rule:
  \[
  f : \mathbb{R}^d \mapsto \{1, 2, \ldots, K\}.
  \]
- Classification accuracy is measured by
  - Equal-cost: Generalization Error (GE)
    \[
    \text{Err}(f) = P(Y \neq f(X)).
    \]
  - Unequal-cost: the risk
    \[
    R(f) = E_{Y,X} C(Y, f(X)).
    \]
Main ideas:

(i) Decompose the multiclass classification problem into multiple binary classification problems.

(ii) Use the majority voting principle (a combined decision from the committee) to predict the label

Common approaches: simple but effective

- One-vs-rest (one-vs-all) approaches
- Pairwise (one-vs-one, all-vs-all) approaches
One-vs-rest Approach

One of the simplest multiclass classifier; commonly used in SVMs; also known as the one-vs-all (OVA) approach

(i) Solve $K$ different binary problems: classify “class $k$” versus “the rest classes” for $k = 1, \cdots, K$.

(ii) Assign a test sample to the class giving the largest $f_k(x)$ (most positive) value, where $f_k(x)$ is the solution from the $k$th problem

Properties:

- Very simple to implement, perform well in practice
- Not optimal (asymptotically): the decision rule is not Fisher consistent if there is no dominating class (i.e. $\arg \max p_k(x) < \frac{1}{2}$).

Pairwise Approach

Also known as all-vs-all (AVA) approach

(i) Solve $\binom{K}{2}$ different binary problems: classify “class $k$” versus “class $j$” for all $j \neq k$. Each classifier is called $g_{ij}$.

(ii) For prediction at a point, each classifier is queried once and issues a vote. The class with the maximum number of (weighted) votes is the winner.

Properties:

- Training process is efficient, by dealing with small binary problems.
- If $K$ is big, there are too many problems to solve. If $K = 10$, we need to train 45 binary classifiers.
- Simple to implement; perform competitively in practice.

One Single SVM approach: Simultaneous Classification

- Label: $\{-1, +1\} \rightarrow \{1, 2, \ldots, K\}$.
- Use one single SVM to construct a decision function vector
  \[ f = (f_1, \ldots, f_K). \]
- Classifier (Decision rule):
  \[ f(x) = \arg\max_{k=1,\ldots,K} f_k(x). \]
- If $K = 2$, there is one $f_k$ and the decision rule is $\text{sign}(f_k)$.
- In some sense, multiple logistic regression is a simultaneous classification procedure.
Overview of Multiclass Learning
Simultaneous Classification by MSVMs
Extensions of SVM
Various Loss Functions
Generalized Functional Margin

Hao Helen Zhang
Lecture 18: Multiclass Support Vector Machines
Overview of Multiclass Learning
Simultaneous Classification by MSVMs
Extensions of SVM
Various Loss Functions
Generalized Functional Margin

Ployhedron one
\[ f_1 - f_2 = 1 \]
\[ f_1 - f_2 = 0 \]
\[ f_2 - f_1 = 1 \]
\[ f_1 - f_3 = 1 \]
\[ f_1 - f_3 = 0 \]
\[ f_3 - f_1 = 1 \]
\[ f_2 - f_3 = 1 \]
\[ f_3 - f_2 = 1 \]

Ployhedron two
\[ f_2 - f_3 = 0 \]

Ployhedron three
\[ f_3 - f_2 = 1 \]
Multiclass SVM: solving one single regularization problem by imposing a penalty on the values of $f_y(x) - f_l(x)$’s.

- Weston and Watkins (1999)
- Cramer and Singer (2002)
- Lee et al. (2004)
- Liu and Shen (2006); multiclass $\psi$-learning: Shen et al. (2003)
Various Multiclass SVMs

- Weston and Watkins (1999):
  a penalty is imposed only if \( f_y(x) < f_k(x) + 2 \) for \( k \neq y \).
  
    - Even if \( f_y(x) < 1 \), a penalty is not imposed as long as \( f_k(x) \) is sufficiently small for \( k \neq y \);
    - Similarly, if \( f_k(x) > 1 \) for \( k \neq y \), we do not pay a penalty if \( f_y(x) \) is sufficiently large.

\[
L(y, f(x)) = \sum_{k \neq y} [2 - (f_y(x) - f_k(x))]_+. 
\]

- Lee et al. (2004):  \( L(y, f(x)) = \sum_{k \neq y} [f_k(x) + 1]_+ \).

  \[
  L(y, f(x)) = [1 - \min_k \{f_y(x) - f_k(x)\}]_+. 
  \]

To avoid the redundancy, a sum-to-zero constraint \( \sum_{k=1}^{K} f_k = 0 \) is sometimes enforced.
Linear Multiclass SVMs

For linear classification problems, we have

\[ f_k(x) = \beta_k x + \beta_{0k}, \quad k = 1, \ldots, K. \]

The sum-to-zero constraint can be replaced by

\[ \sum_{k=1}^{K} \beta_{0k} = 0, \quad \sum_{k=1}^{K} \beta_k = 0. \]

The optimization problem becomes

\[
\min_{f} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \sum_{k=1}^{K} \|\beta_k\|^2
\]

subject to the sum-to-zero constraint.
Nonlinear Multiclass SVMs

To achieve the nonlinear classification, we assume

\[ f_k(x) = \beta'_k \phi(x) + \beta_{k0}, \quad k = 1, \ldots, K. \]

where \( \phi(x) \) represents the basis functions in the feature space \( \mathcal{F} \).

- Similar to the binary classification, the nonlinear MSVM can be conveniently solved using a kernel function.
Regularization Problems for Nonlinear MSVMs

We can represent the MSVM as the solution to a regularization problem in the RKHS.

- Assume that

\[ f(x) = (f_1(x), \ldots, f_K(x)) \in \prod_{k=1}^{K} (\{1\} + \mathcal{H}_k) \]

under the sum-to-zero constraint.

- Then a MSVM classifier can be derived by solving

\[ \min_{f} \sum_{i=1}^{n} L(y_i, f(x_i)) + \lambda \sum_{k=1}^{K} \|g_k\|_{\mathcal{H}_k}^2, \]

where \( f_k(x) = g_k(x) + \beta_{0k}, g_k \in \mathcal{H}_k, \beta_{0k} \in \mathcal{R}. \)
Generalized Functional Margin

Given \((x, y)\), a reasonable decision vector \(f(x)\) should
- encourage a large value for \(f_y(x)\)
- have small values for \(f_k(x), k \neq y\).

Define the \(K-1\)-vector of relative differences as
\[
g = (f_y(x) - f_1(x), \ldots, f_y(x) - f_{y-1}(x), f_y(x) - f_{y+1}(x), \ldots, f_y(x) - f_K(x))\,.
\]

Liu et al. (2004) called the vector \(g\) the generalized functional margin of \(f\)
- \(g\) characterizes correctness and strength of classification of \(x\) by \(f\).
- \(f\) indicates a correct classification of \((x, y)\) if
\[
g(f(x), y) > 0_{k-1}.
\]
0-1 Loss with Functional Margin

A point \((x, y)\) is misclassified if \(y \neq \arg \max_k f_k(x)\).
Define the multivariate sign function as
\[
\text{sign}(u) = 1 \quad \text{if} \quad u_{\min} = \min(u_1, \ldots, u_m) > 0,
-1 \quad \text{if} \quad u_{\min} \leq 0.
\]

where \(u = (u_1, \ldots, u_m)\). Using the functional margin,

- The 0-1 loss becomes
  \[
  l(\min g(f(x), y) < 0) = \frac{1}{2} [1 - \text{sign}(g(f(x), y))].
  \]
- The GE becomes to
  \[
  R[f] = \frac{1}{2} E \left[1 - \text{sign}(g(f(x), y))\right].
  \]
Generalized Loss Functions Using Functional Margin

A natural way to generalize the binary loss is

$$\sum_{i=1}^{n} \ell(\min \mathbf{g}(\mathbf{f}(\mathbf{x}_i), y_i)).$$

In particular, the loss function $L(y, \mathbf{f}(\mathbf{x}))$ can be expressed as $V(\mathbf{g}(\mathbf{f}(\mathbf{x}), y))$ with

- Weston and Watkins (1999): $V(\mathbf{u}) = \sum_{j=1}^{K-1} [2 - u_j]_+.$
- Lee et al. (2004): $V(\mathbf{u}) = \sum_{j=1}^{K-1} [\frac{\sum_{c=1}^{K-1} u_c}{K} - u_j + 1]_+.$

All of these loss functions are the upper bounds of the 0-1 loss.
Small Round Blue Cell Tumors of Childhood

- Khan et al. (2001) in *Nature Medicine*

- Tumor types: neuroblastoma (NB), rhabdomyosarcoma (RMS), non-Hodgkin lymphoma (NHL) and the Ewing family of tumors (EWS)

- Number of genes: 2308

- Class distribution of data set

<table>
<thead>
<tr>
<th>Data set</th>
<th>EWS</th>
<th>BL(NHL)</th>
<th>NB</th>
<th>RMS</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training set</td>
<td>23</td>
<td>8</td>
<td>12</td>
<td>20</td>
<td>63</td>
</tr>
<tr>
<td>Test set</td>
<td>6</td>
<td>3</td>
<td>6</td>
<td>5</td>
<td>20</td>
</tr>
<tr>
<td>Total</td>
<td>29</td>
<td>11</td>
<td>18</td>
<td>25</td>
<td>83</td>
</tr>
</tbody>
</table>
Figure 3: Three principal components of 100 gene expression levels (circles: training samples, squares: test samples including non SRBCT samples). The tumor types are distinguished by colors.
Characteristics of Support Vector Machines

- High accuracy, high flexibility
- Naturally handle large dimensional data
- Sparse representation of the solutions (via support vectors): fast for making future prediction
- No probability estimates (hard classifiers)
Other Active Problems in SVM

- Variable/Feature Selection
- Proximal SVM – faster computation
- Robust SVM – get rid of outliers
- Choice of kernels
The $L_1$ SVM

- Replace the $L_2$ penalty by the $L_1$ penalty.
- The $L_1$ penalty tends to give sparse solutions.
- For $f(x) = h(x)^T \beta + \beta_0$, the $L_1$ SVM solves

$$\min_{\beta_0, \beta} \sum_{i=1}^{n} [1 - y_i f(x_i)]_+ + \lambda \sum_{j=1}^{d} |\beta_j|.$$  

(1)

- The solution will have at most $n$ nonzero coefficients $\beta_j$. 
$L_1$ Penalty versus $L_2$ Penalty

![Graph showing $L_1$ and $L_2$ penalties]
Robust Support Vector Machines

- Hinge loss is unbounded; sensitive to outliers (e.g. wrong labels etc)
- Support Vectors: \( y_i f(x_i) \leq 1 \).
- Truncated hinge loss: \( T_s(u) = H_1(u) - H_s(u) \), where \( H_s(u) = [s - u]^+ \).
- Remove some “bad” SVs (Wu and Liu, 2006).
Decomposition: Difference of Convex Functions

- $T_s(u) = H_1(u) - H_s(u)$. 

![Graphs showing Decomposition](image-url)
D.C. Algorithm: The Difference Convex Algorithm for minimizing
\[ J(\Theta) = J_{vex}(\Theta) + J_{cav}(\Theta) \]

1. Initialize \( \Theta_0 \).
2. Repeat
   \[ \Theta_{t+1} = \text{argmin}_\Theta \left( J_{vex}(\Theta) + \left\langle J'_{cav}(\Theta_t), \Theta - \Theta_t \right\rangle \right) \]
   until convergence of \( \Theta_t \).

- The algorithm converges in finite steps (Liu et al. (2005)).
- Choice of initial values: Use SVM’s solution.
- RSVM: The set of SVs is a only a SUBSET of the original one!
- Nonlinear learning can be achieved by the kernel trick.