Support Vector Machine (SVM)

Jinlong Wu

Department of Scientific and Engineering Computing,
School of Mathematical Sciences, PKU

2007. 6.1
Outline

1 Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2 Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3 Some Advanced Techniques
Outline

1. Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2. Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3. Some Advanced Techniques
Outline

1 Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2 Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3 Some Advanced Techniques
Support Vector Machine (SVM)

Outline

Some Classification Problems From The Real World

- Microarray Classification
- Heart Attack Prediction
- Phoneme Classification
- Handwritten Digit Recognition
- Ball Detection in Static Images
- Many Other Real-world Problems

Support Vector Classifier (SVC)

- SVC For Linearly Separable Problems
- SVC For Linearly Inseparable Data Sets
- Lagrangian (Wolfe) Dual Problem

Some Advanced Techniques
Outline

1 Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2 Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3 Some Advanced Techniques
Prediction according to demographic, diet and clinical measurements

Outline

Some Classification Problems From The Real World

Microarray Classification

Heart Attack Prediction

Phoneme Classification

Handwritten Digit Recognition

Ball Detection in Static Images

Many Other Real-world Problems

Support Vector Classifier (SVC)

SVC For Linearly Separable Problems

SVC For Linearly Inseparable Data Sets

Lagrangian (Wolfe) Dual Problem

Some Advanced Techniques
Outline

1. Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2. Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3. Some Advanced Techniques
Classify a recorded phoneme on the basis of a log-periodogram.
Outline

1. Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2. Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3. Some Advanced Techniques
Identify the numbers in a handwritten zip code, from a digitized image
Outline

1 Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2 Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3 Some Advanced Techniques
Detecting the occurrence of a goal during a soccer match
Outline

1. Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2. Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3. Some Advanced Techniques
Many other real-world problems

- Email spam detection
- Textual classification
- Face recognition
Many other real-world problems

- Email spam detection
- Textual classification
- Face recognition
Many other real-world problems

- Email spam detection
- Textual classification
- Face recognition
Outline

1. Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2. Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3. Some Advanced Techniques
Outline

1 Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2 Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3 Some Advanced Techniques
A trivial linearly separable example

The numbers of observations and variables are $n = 9$, $p = 2$ respectively.

**Figure:** Red: $+1$, Green: $-1$
Too many perfect hyperplanes

Which hyperplane should we choose since the three ones separate the toy example perfectly?

Figure: Red: +1, Green: −1
Signed distance of a point $x$ to a hyperplane $L$

$L = \{ x : \beta_0 + \beta^T x = 0 \}$

Therefore, the absolute distance from a point $x$ to the hyperplane $L$ is

$$\frac{1}{\| \beta \|} |\beta^T x + \beta_0|$$
Signed distance of a point $x$ to a hyperplane $L$

$L = \{ x : \beta_0 + \beta^T x = 0 \}$

\[
\frac{1}{\| \beta \|} \beta^T (x - x_0) = \frac{1}{\| \beta \|} \left( \beta^T x - \beta^T x_0 \right) = \frac{1}{\| \beta \|} (\beta^T x + \beta_0)
\]

Therefore, the absolute distance from a point $x$ to the hyperplane $L$ is

\[
\frac{1}{\| \beta \|} |\beta^T x + \beta_0|
\]
The optimal hyperplane proposed by Vapnik and Lerner (1963)

For an optional hyperplane $L$, we hope

1. Given some positive $C$,

$$\frac{1}{\|\beta\|} |\beta^T x_i + \beta_0| \geq C, \quad \forall i.$$

2. The larger $C$, the better.
The optimal hyperplane proposed by Vapnik and Lerner (1963)

For an optional hyperplane $L$, we hope

1. Given some positive $C$, 

$$\frac{1}{\| \beta \|} | \beta^T x_i + \beta_0 | \geq C, \quad \forall i.$$ 

2. The larger $C$, the better.
The optimal hyperplane proposed by Vapnik and Lerner (1963)

For an optional hyperplane \( L \), we hope

1. Given some positive \( C \),

\[
\frac{1}{\| \beta \|} |\beta^T x_i + \beta_0| \geq C, \quad \forall i.
\]

2. The larger \( C \), the better.
The optimal hyperplane proposed by Vapnik and Lerner (1963)

For an optional hyperplane $L$, we hope

1. Given some positive $C$,

\[
\frac{1}{\|\beta\|} |\beta^T x_i + \beta_0| \geq C, \quad \forall i.
\]

2. The larger $C$, the better.

Our goal $\Rightarrow$ find a $L$ which has the biggest $C$
The optimal hyperplane proposed by Vapnik and Lerner (1963)

For an optional hyperplane $L$, we hope

1. Given some positive $C$,

$$\frac{1}{\|\beta\|} |\beta^T x_i + \beta_0| \geq C, \quad \forall i.$$ 

2. The larger $C$, the better.

To sum up, we get a constrained optimization problem (OP)

$$\max_{\beta, \beta_0} C$$

subject to

$$\frac{1}{\|\beta\|} y_i (\beta^T x_i + \beta_0) \geq C, \quad \forall i.$$
The black hyperplane is the optimum for this toy example.
Lagrange primal problem for linearly separable data set

Assume $\frac{1}{\|\beta\|} = C$, re-express the previous OP

$$
\begin{align*}
\max_{\beta, \beta_0} & \quad C \\
\text{subject to} & \quad \frac{1}{\|\beta\|} y_i (\beta^T x_i + \beta_0) \geq C, \quad \forall i
\end{align*}
$$

as

$$
\begin{align*}
\min_{\beta, \beta_0} & \quad \frac{1}{2} \|\beta\|^2 \\
\text{subject to} & \quad y_i (\beta^T x_i + \beta_0) \geq 1, \quad \forall i.
\end{align*}
$$
Outline

1. Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2. Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3. Some Advanced Techniques
A linearly inseparable data set

Hastie et al. (2001) gives a synthetic mixture example.

\[ n = 200, \ p = 2; \ \text{Red: } +1, \ \text{Green: } -1. \]
There does not exist any hyperplane, such that

\[ y_i (\beta^T x_i + \beta_0) \geq 1, \quad \forall i. \]
Two extensions (I)

Replacing a hyperplane with a hypersurface

\[ f(x) = \beta^T x_i + \beta_0 \implies f(x) = \beta^T \Phi(x_i) + \beta_0 \]
Hopefully, we can find some hypersurface, which satisfies

\[ y_i (\beta^T \Phi(x_i) + \beta_0) \geq 1, \quad \forall i. \]
The previous extension is useless for the mixture example.

Adding some new techniques:

1. Relaxing the constraints

$$y_i (\beta^T \Phi(x_i) + \beta_0) \geq 1 \Rightarrow y_i (\beta^T \Phi(x_i) + \beta_0) \geq 1 - \xi_i,$$

where $\xi_i \geq 0, \forall i$, and $\xi = (\xi_1, \cdots, \xi_n)$ is called slack vector.

2. Penalizing to $\xi_i$'s to avoid that $\xi_i$'s approach $+\infty$. 

Two extentions (II)

The previous extention is useless for the mixture example.

Adding some new techniques:

1. Relaxing the constraints

\[ y_i (\beta^T \Phi(x_i) + \beta_0) \geq 1 \Rightarrow y_i (\beta^T \Phi(x_i) + \beta_0) \geq 1 - \xi_i, \]

where \( \xi_i \geq 0, \forall i \), and \( \xi = (\xi_1, \cdots, \xi_n) \) is called slack vector.

2. Penalizing to \( \xi_i \)'s to avoid that \( \xi_i \)'s approach +\( \infty \).
Two extensions (II)

The previous extension is useless for the mixture example.

Adding some new techniques:

1. Relaxing the constraints

\[ y_i (\beta^T \Phi(x_i) + \beta_0) \geq 1 \Rightarrow y_i (\beta^T \Phi(x_i) + \beta_0) \geq 1 - \xi_i, \]

where \( \xi_i \geq 0, \forall i \), and \( \xi = (\xi_1, \cdots, \xi_n) \) is called slack vector.

2. Penalizing to \( \xi_i \)'s to avoid that \( \xi_i \)'s approach \(+\infty\).
The previous extent is useless for the mixture example.

Adding some new techniques:

1. Relaxing the constraints

\[ y_i(\beta^T \Phi(x_i) + \beta_0) \geq 1 \Rightarrow y_i(\beta^T \Phi(x_i) + \beta_0) \geq 1 - \xi_i, \]

where \( \xi_i \geq 0, \forall i \), and \( \xi = (\xi_1, \cdots, \xi_n) \) is called the slack vector.

2. Penalizing to \( \xi_i \)'s to avoid that \( \xi_i \)'s approach \( +\infty \).
A new OP for the inseparable cases:

\[
\min_{\beta, \beta_0} \quad \frac{1}{2} \| \beta \|^2 + C \sum_{i=1}^{n} \xi_i
\]

subject to \( y_i(\Phi(x_i)^T \beta + \beta_0) \geq 1 - \xi_i, \forall i, \)

\( \xi_i \geq 0. \)

where \( C \) is the penalization constant.
Outline

1. Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2. Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3. Some Advanced Techniques
Difficult to solve Lagrange primal problem directly.

⇒ to solve Lagrangian dual problem

\[
\min_{\alpha} \ g(\alpha) = \frac{1}{2} \alpha^T Q \alpha - e^T \alpha
\]

subject to \[
\begin{align*}
y^T \alpha &= 0 \\
0 &\leq \alpha_i \leq C, \ \forall i
\end{align*}
\]

where \( \alpha = (\alpha_1, \cdots, \alpha_n) \) is the Lagrange multiplier vector, \( y = (y_1, \cdots, y_n) \), \( e = (1, \cdots, 1) \), \( C \) is the penalization constant, \( Q \) is a \( n \) by \( n \) matrix and \( Q_{ij} = y_i y_j \Phi(x_i)^T \Phi(x_j) \).
The discriminant function

Solving the Lagrangian dual problem and getting $\hat{\alpha}$,

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \Phi(x_i),$$

and $\hat{\beta}_0$ can be computed easily.

Finally, the discriminant function is

$$\hat{f}(x) = \Phi(x)^T \hat{\beta} + \hat{\beta}_0 = \sum_{i=1}^{n} \hat{\alpha}_i y_i \Phi(x)^T \Phi(x_i) + \hat{\beta}_0$$

If $\hat{f}(x) \geq 0$, we classify it to class $+1$, otherwise to class $-1$. 
The discriminant function

Solving the Lagrangian dual problem and getting $\hat{\alpha}$,

$$\hat{\beta} = \sum_{i=1}^{n} \hat{\alpha}_i y_i \Phi(x_i),$$

and $\hat{\beta}_0$ can be computed easily.

Finally, the discriminant function is

$$\hat{f}(x) = \Phi(x)^T \hat{\beta} + \hat{\beta}_0 = \sum_{i=1}^{n} \hat{\alpha}_i y_i \Phi(x)^T \Phi(x_i) + \hat{\beta}_0$$

If $\hat{f}(x) \geq 0$, we classify it to class $+1$, otherwise to class $-1$. 
Outline

1. Some Classification Problems From The Real World
   - Microarray Classification
   - Heart Attack Prediction
   - Phoneme Classification
   - Handwritten Digit Recognition
   - Ball Detection in Static Images
   - Many Other Real-world Problems

2. Support Vector Classifier (SVC)
   - SVC For Linearly Separable Problems
   - SVC For Linearly Inseparable Data Sets
   - Lagrangian (Wolfe) Dual Problem

3. Some Advanced Techniques
Some advanced techniques

1. Some particular methods of solving Lagrangian dual problems
   - Chunking
   - Decomposition
   - Sequential Minimal Optimization

2. Shrinking

3. Caching
Some advanced techniques

1. Some particular methods of solving Lagrangian dual problems
   - Chunking
   - Decomposition
   - Sequential Minimal Optimization

2. Shrinking

3. Caching
Some advanced techniques

1. Some particular methods of solving Lagrangian dual problems
   - Chunking
   - Decomposition
   - Sequential Minimal Optimization

2. Shrinking

3. Caching
Some advanced techniques

1. Some particular methods of solving Lagrangian dual problems
   - Chunking
   - Decomposition
   - Sequential Minimal Optimization

2. Shrinking

3. Caching
Some advanced techniques

1. Some particular methods of solving Lagrangian dual problems
   - Chunking
   - Decomposition
   - Sequential Minimal Optimization

2. Shrinking

3. Caching
Some advanced techniques

1. Some particular methods of solving Lagrangian dual problems
   - Chunking
   - Decomposition
   - Sequential Minimal Optimization

2. Shrinking

3. Caching
Some advanced techniques

1. Some particular methods of solving Lagrangian dual problems
   - Chunking
   - Decomposition
   - Sequential Minimal Optimization

2. Shrinking

3. Caching

All of these techniques have been utilized in a new SVM software package—PKSVM.
Outline

Some Classification Problems From The Real World
- Microarray Classification
- Heart Attack Prediction
- Phoneme Classification
- Handwritten Digit Recognition
- Ball Detection in Static Images
- Many Other Real-world Problems

Support Vector Classifier (SVC)
- SVC For Linearly Separable Problems
- SVC For Linearly Inseparable Data Sets
- Lagrangian (Wolfe) Dual Problem

Some Advanced Techniques

Reference I


Support Vector Machine (SVM)

Jinlong Wu

Outline

Some Classification Problems From The Real World
- Microarray Classification
- Heart Attack Prediction
- Phoneme Classification
- Handwritten Digit Recognition
- Ball Detection in Static Images
- Many Other Real-world Problems

Support Vector Classifier (SVC)

SVC For Linearly Separable Problems
SVC For Linearly Inseparable Data Sets
Lagrangian (Wolfe) Dual Problem

Some Advanced Techniques

Have a good day