Learning in Big Data Analytics
Support Vector Machines

Alexander Schönhuth

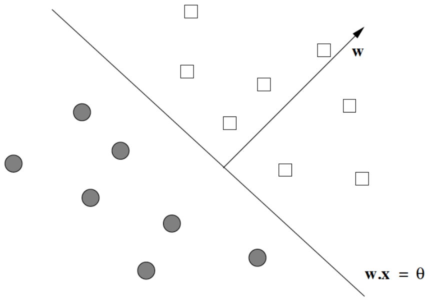
UNIVERSITÄT BIELEFELD
Faculty of Technology

Bielefeld University

November 17, 2021

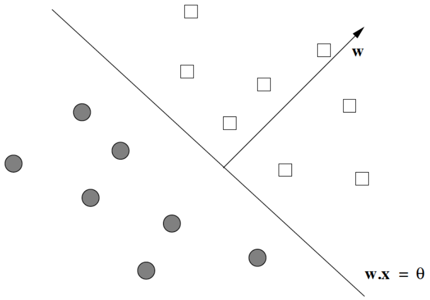
Perceptrons Revisited

Perceptron Revisited



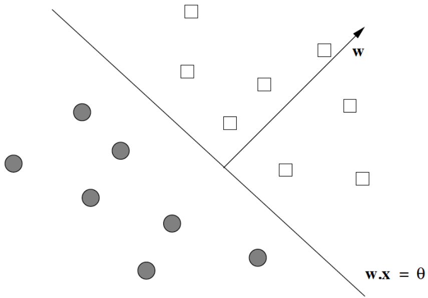
- A perceptron divides the space into two half spaces
- Half spaces capture the two different classes
- Normal vector alternative description of half space

Perceptron Revisited



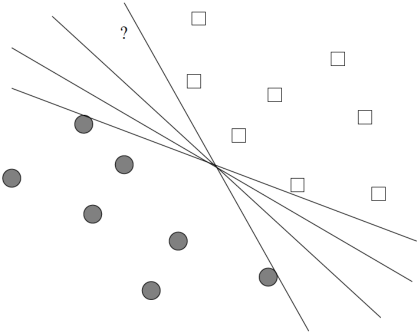
- A perceptron divides the space into two half spaces
- Half spaces capture the two different classes
- Normal vector alternative description of half space

Perceptron Revisited



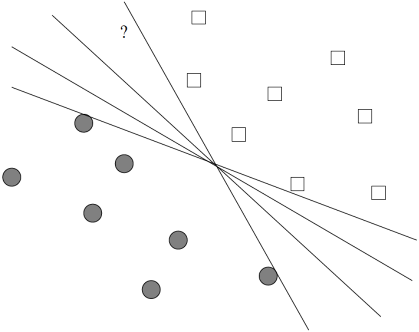
- A perceptron divides the space into two half spaces
- Half spaces capture the two different classes
- Normal vector alternative description of half space

Perceptron Revisited



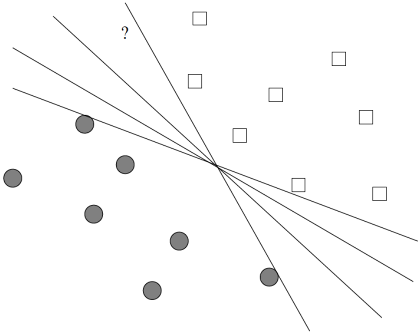
- Several half spaces (normal vectors) divide training data
- *Question:* any half space optimal, in a sensibly defined way?
- What to do if data cannot be separated (is *non-separable*)?

Perceptron Revisited



- Several half spaces (normal vectors) divide training data
- *Question:* any half space optimal, in a sensibly defined way?
- What to do if data cannot be separated (is *non-separable*)?

Perceptron Revisited



- Several half spaces (normal vectors) divide training data
- *Question:* any half space optimal, in a sensibly defined way?
- What to do if data cannot be separated (is *non-separable*)?

Support Vector Machines: Motivation

- Support vector machines (SVM's) address to choose most reasonable half space
- SVM's choose half space that maximizes the *margin*, i.e. the distance between data points and half space
- If separable, maximize distance between hyperplane and closest data points
- If not separable, minimize *loss function* that
  - penalizes misclassified points
  - penalizes points correctly classified but too close to hyperplane (to a lesser extent)

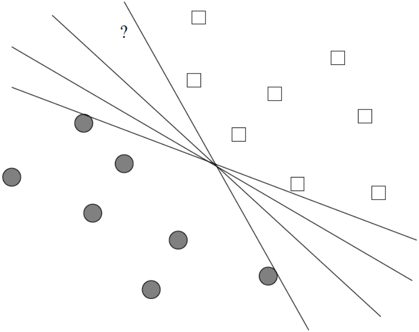
Support Vector Machines: Motivation

- Support vector machines (SVM's) address to choose most reasonable half space
- SVM's choose half space that maximizes the *margin*, i.e. the distance between data points and half space
- If separable, maximize distance between hyperplane and closest data points
- If not separable, minimize *loss function* that
  - penalizes misclassified points
  - penalizes points correctly classified but too close to hyperplane (to a lesser extent)

SUPPORT VECTOR MACHINES: MOTIVATION

- Support vector machines (SVM's) address to choose most reasonable half space
- SVM's choose half space that maximizes the *margin*, i.e. the distance between data points and half space
- If separable, maximize distance between hyperplane and closest data points
- If not separable, minimize *loss function* that
  - penalizes misclassified points
  - penalizes points correctly classified but too close to hyperplane (to a lesser extent)

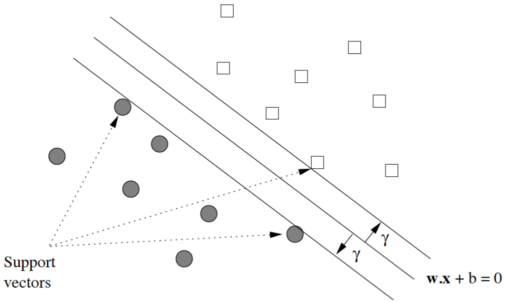
Perceptron Revisited



- Outer hyperplanes come very close to data points
- So, inner hyperplanes are likely the better choice
- ☞ Try to make explicit!

Separable Data

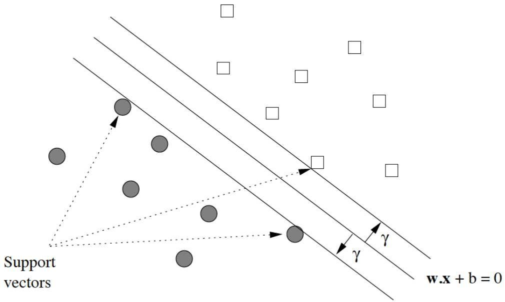
Separable Data



- *Goal:* Select hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ that maximizes distance γ
- *Intuition:* The further away data from hyperplane, the more certain their classification
- Increases chances to correctly classify unseen data (to generalize)

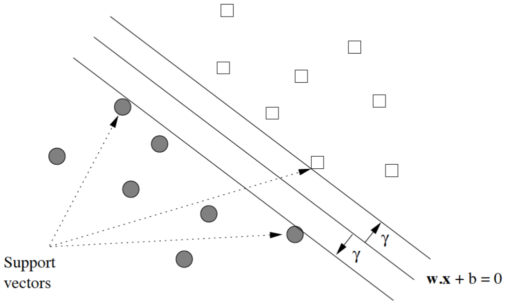
UNIVERSITÄT BIELEFELD

Separable Data



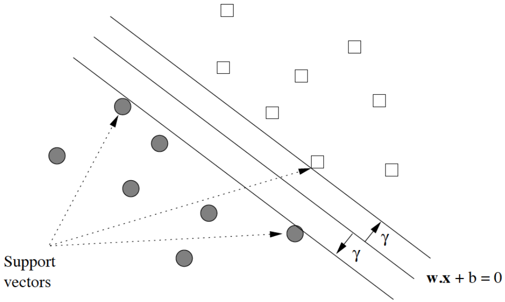
- *Goal:* Select hyperplane $\mathbf{w} \cdot \mathbf{x} + b = 0$ that maximizes distance γ
- *Intuition:* The further away data from hyperplane, the more certain their classification
- Increases chances to correctly classify unseen data (to generalize)

Support Vectors



- Two parallel hyperplanes at distance γ touch one or more of *support vectors*
- In most cases, d-dimensional data set has $d + 1$ support vectors (but there can be more)

Support Vectors



- Two parallel hyperplanes at distance γ touch one or more of *support vectors*
- In most cases, d-dimensional data set has $d + 1$ support vectors (but there can be more)

Problem Formulation: First Try

Let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ be a training data set, where $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}, i = 1, ..., n$.

Problem: By varying $\mathbf{w}, b$, maximize γ such that

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq \gamma \quad \text{for all } i = 1, ..., n \tag{1}$$

Problem Formulation: First Try

Let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ be a training data set, where $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}, i = 1, ..., n$.

Problem: By varying $\mathbf{w}, b$, maximize γ such that

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq \gamma \quad \text{for all } i = 1, ..., n \tag{1}$$

Issue

- Replacing $\mathbf{w}$ and b by $2\mathbf{w}$ and $2b$ yields $y_i(2\mathbf{w}\mathbf{x}_i + 2b) \geq 2\gamma$
- There is no optimal γ

Problem Formulation: First Try

Let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ be a training data set, where $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}, i = 1, ..., n$.

Problem: By varying $\mathbf{w}, b$, maximize γ such that

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq \gamma \quad \text{for all } i = 1, ..., n \tag{1}$$

Issue

- Replacing $\mathbf{w}$ and b by $2\mathbf{w}$ and $2b$ yields $y_i(2\mathbf{w}\mathbf{x}_i + 2b) \geq 2\gamma$
- There is no optimal γ

Problem badly formulated

PROBLEM FORMULATION: FIRST TRY

Let $(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)$ be a training data set, where $\mathbf{x}_i \in \mathbb{R}^d, y_i \in \{-1, +1\}, i = 1, ..., n$.

PROBLEM: By varying $\mathbf{w}, b$, maximize γ such that

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq \gamma \quad \text{for all } i = 1, ..., n \tag{1}$$

Issue

- Replacing $\mathbf{w}$ and b by $2\mathbf{w}$ and $2b$ yields $y_i(2\mathbf{w}\mathbf{x}_i + 2b) \geq 2\gamma$
- There is no optimal γ

Problem badly formulated

Try again!

Problem Formulation: Solution

- Data set $(\mathbf{x}_i, y_i), i = 1, ..., n$ as before; let $H := \{\mathbf{x} \mid \mathbf{wx} + b = 0\}$ be the hyperplane given by $\mathbf{w}$ and b.
- Let
$$d(\mathbf{x}_i, H) := \min_x \{d(\mathbf{x}_i, \mathbf{x}) \mid \mathbf{wx} + b = 0\} \tag{2}$$
be the distance between $\mathbf{x}_i$ and H.
- *Solution:* Impose additional constraint: consider only combinations $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$ such that for support vectors $\mathbf{x}$
$$y_i(\mathbf{wx} + b) \in \{-1, +1\} \tag{3}$$
- *Good Formulation:* By varying $\mathbf{w}, b,$ maximize γ such that
$$d(\mathbf{x}_i, H) \geq \gamma \quad \text{for all } i = 1, ..., n \tag{4}$$
and (3) applies

Problem Formulation: Solution

- Data set $(\mathbf{x}_i, y_i), i = 1, ..., n$ as before; let $H := \{\mathbf{x} \mid \mathbf{wx} + b = 0\}$ be the hyperplane given by $\mathbf{w}$ and b.
- Let

$$d(\mathbf{x}_i, H) := \min_x \{d(\mathbf{x}_i, \mathbf{x}) \mid \mathbf{wx} + b = 0\} \tag{2}$$

be the distance between $\mathbf{x}_i$ and H.

- *Solution:* Impose additional constraint: consider only combinations $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$ such that for support vectors $\mathbf{x}$

$$y_i(\mathbf{wx} + b) \in \{-1, +1\} \tag{3}$$

- *Good Formulation:* By varying $\mathbf{w}, b,$ maximize γ such that

$$d(\mathbf{x}_i, H) \geq \gamma \quad \text{for all } i = 1, ..., n \tag{4}$$

and (3) applies

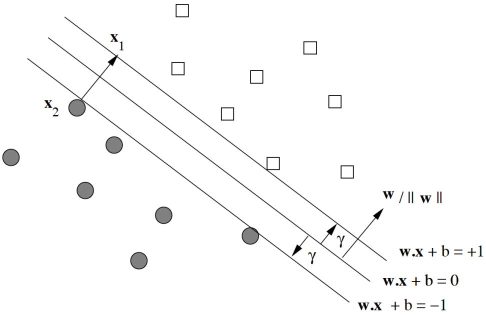
Problem Formulation: Solution

- Data set $(\mathbf{x}_i, y_i), i = 1, ..., n$ as before; let $H := \{\mathbf{x} \mid \mathbf{w}\mathbf{x} + b = 0\}$ be the hyperplane given by $\mathbf{w}$ and b.
- Let
$$d(\mathbf{x}_i, H) := \min_x \{d(\mathbf{x}_i, \mathbf{x}) \mid \mathbf{w}\mathbf{x} + b = 0\} \tag{2}$$
be the distance between $\mathbf{x}_i$ and H.
- *Solution:* Impose additional constraint: consider only combinations $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$ such that for support vectors $\mathbf{x}$
$$y_i(\mathbf{w}\mathbf{x} + b) \in \{-1, +1\} \tag{3}$$
- *Good Formulation:* By varying $\mathbf{w}, b,$ maximize γ such that
$$d(\mathbf{x}_i, H) \geq \gamma \quad \text{for all } i = 1, ..., n \tag{4}$$
and (3) applies

Problem Formulation: Solution

- Data set $(\mathbf{x}_i, y_i), i = 1, ..., n$ as before; let $H := \{\mathbf{x} \mid \mathbf{wx} + b = 0\}$ be the hyperplane given by $\mathbf{w}$ and b.
- Let
$$d(\mathbf{x}_i, H) := \min_x \{d(\mathbf{x}_i, \mathbf{x}) \mid \mathbf{wx} + b = 0\} \tag{2}$$
be the distance between $\mathbf{x}_i$ and H.
- *Solution:* Impose additional constraint: consider only combinations $\mathbf{w} \in \mathbb{R}^d, b \in \mathbb{R}$ such that for support vectors $\mathbf{x}$
$$y_i(\mathbf{wx} + b) \in \{-1, +1\} \tag{3}$$
- *Good Formulation:* By varying $\mathbf{w}, b$, maximize γ such that
$$d(\mathbf{x}_i, H) \geq \gamma \quad \text{for all } i = 1, ..., n \tag{4}$$
and (3) applies

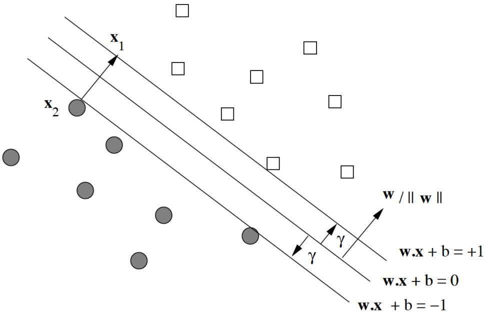
ALTERNATIVE PROBLEM FORMULATION I



- $\mathbf{w}, b, \gamma$ determined according to (3),(4)
- $\mathbf{x}_2$ is support vector on lower hyperplane, so by (3), $\mathbf{w}\mathbf{x}_2 + b = -1$
- Let $\mathbf{x}_1$ be the projection of $\mathbf{x}_2$ onto upper hyperplane:

$$\mathbf{x}_1 = \mathbf{x}_2 + 2\gamma \frac{\mathbf{w}}{||\mathbf{w}||} \tag{5}$$

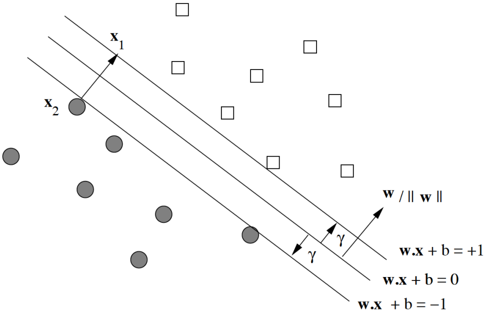
Alternative Problem Formulation I



- $\mathbf{w}, b, \gamma$ determined according to (3),(4)
- $\mathbf{x}_2$ is support vector on lower hyperplane, so by (3), $\mathbf{w}\mathbf{x}_2 + b = -1$
- Let $\mathbf{x}_1$ be the projection of $\mathbf{x}_2$ onto upper hyperplane:

$$\mathbf{x}_1 = \mathbf{x}_2 + 2\gamma \frac{\mathbf{w}}{||\mathbf{w}||} \tag{5}$$

ALTERNATIVE PROBLEM FORMULATION I



- $\mathbf{w}, b, \gamma$ determined according to (3),(4)
- $\mathbf{x}_2$ is support vector on lower hyperplane, so by (3), $\mathbf{w}\mathbf{x}_2 + b = -1$
- Let $\mathbf{x}_1$ be the projection of $\mathbf{x}_2$ onto upper hyperplane:

$$\mathbf{x}_1 = \mathbf{x}_2 + 2\gamma\frac{\mathbf{w}}{||\mathbf{w}||} \tag{5}$$

Alternative Problem Formulation II

That is, further, $\mathbf{x}_1$ is on the hyperplane defined by $\mathbf{wx} + b = 1$, meaning

$$\mathbf{wx}_1 + b = 1 \tag{6}$$

Alternative Problem Formulation II

That is, further, $\mathbf{x}_1$ is on the hyperplane defined by $\mathbf{w}\mathbf{x} + b = 1$, meaning

$$\mathbf{w}\mathbf{x}_1 + b = 1 \tag{6}$$

Substituting $\mathbf{x}_1 = \mathbf{x}_2 + 2\gamma\frac{\mathbf{w}}{||\mathbf{w}||}$ (5) into (6) yields

$$\mathbf{w} \cdot (\mathbf{x}_2 + 2\gamma\frac{\mathbf{w}}{||\mathbf{w}||}) + b = 1 \tag{7}$$

Alternative Problem Formulation II

That is, further, $\mathbf{x}_1$ is on the hyperplane defined by $\mathbf{wx} + b = 1$, meaning

$$\mathbf{wx}_1 + b = 1 \tag{6}$$

Substituting $\mathbf{x}_1 = \mathbf{x}_2 + 2\gamma\frac{\mathbf{w}}{||\mathbf{w}||}$ (5) into (6) yields

$$\mathbf{w} \cdot (\mathbf{x}_2 + 2\gamma\frac{\mathbf{w}}{||\mathbf{w}||}) + b = 1 \tag{7}$$

We obtain

$$\mathbf{wx}_2 + b + 2\gamma\frac{\mathbf{ww}}{||\mathbf{w}||} = 1 \tag{8}$$

Alternative Problem Formulation II

That is, further, $\mathbf{x}_1$ is on the hyperplane defined by $\mathbf{wx} + b = 1$, meaning

$$\mathbf{wx}_1 + b = 1 \tag{6}$$

Substituting $\mathbf{x}_1 = \mathbf{x}_2 + 2\gamma \frac{\mathbf{w}}{||\mathbf{w}||}$ (5) into (6) yields

$$\mathbf{w} \cdot (\mathbf{x}_2 + 2\gamma \frac{\mathbf{w}}{||\mathbf{w}||}) + b = 1 \tag{7}$$

We obtain

$$\mathbf{wx}_2 + b + 2\gamma \frac{\mathbf{ww}}{||\mathbf{w}||} = 1 \tag{8}$$

Because $\mathbf{ww} = ||\mathbf{w}||^2$, and by further regrouping, we conclude that

$$\gamma = \frac{1}{||\mathbf{w}||} \tag{9}$$

Alternative Problem Formulation III

Let dataset $(\mathbf{x}_i, y_i), i = 1, ..., n$ be as before.

Equivalent Problem Formulation:

By varying $\mathbf{w}, b$, minimize $||\mathbf{w}||$ subject to

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1 \quad \text{for all } i = 1, ..., n \tag{10}$$

Alternative Problem Formulation III

Let dataset $(\mathbf{x}_i, y_i), i = 1, ..., n$ be as before.

Equivalent Problem Formulation:

By varying $\mathbf{w}, b$, minimize $||\mathbf{w}||$ subject to

$$y_i(\mathbf{w}\mathbf{x}_i + b) \geq 1 \quad \text{for all } i = 1, ..., n \tag{10}$$

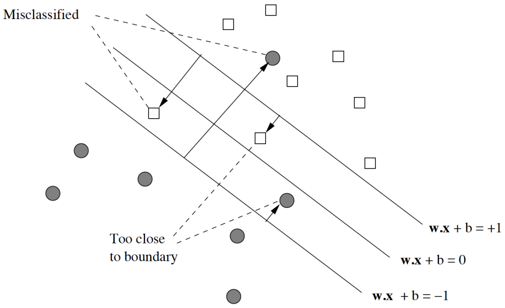
Optimizing under Constraints

- Topic is broadly covered
- Many packages can be used
- Target function $(||\mathbf{w}||)^2 = \sum_i w_i^2$ quadratic; well manageable

Example

Non Separable Data

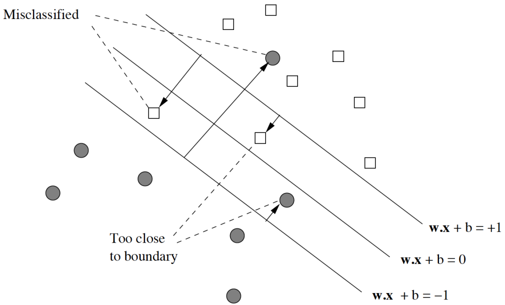
NON SEPARABLE DATA SETS



Situation:

- Some points misclassified, some too close to boundary
  ☞ *bad points*
- *Non separable data:* any choice of **w**, *b* yields bad points

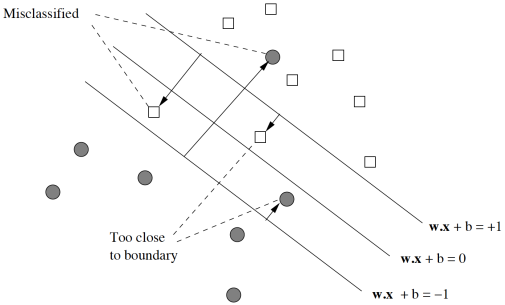
NON SEPARABLE DATA SETS



Situation:

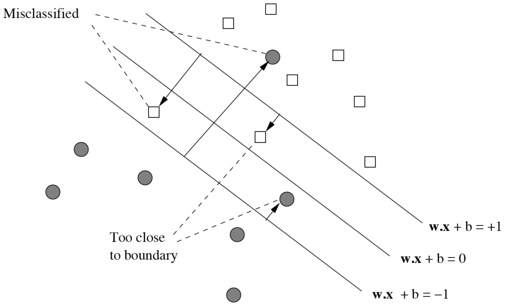
- Some points misclassified, some too close to boundary
  ☞ *bad points*
- *Non separable data:* any choice of **w**, *b* yields bad points

Non Separable Data: Motivation



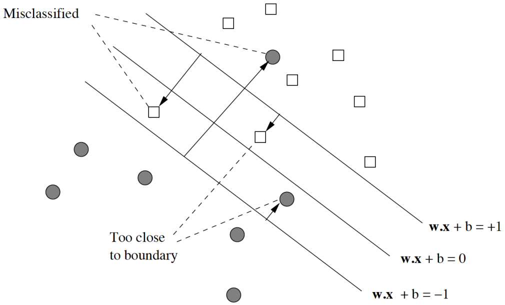
- *Situation:* No hyperplane can separate the data points correctly
- *Approach:*
  - Determine appropriate penalties for bad points
  - Solve original problem, by involving penalties

Non Separable Data: Motivation



- *Situation:* No hyperplane can separate the data points correctly
- *Approach:*
  - Determine appropriate penalties for bad points
  - Solve original problem, by involving penalties

NON SEPARABLE DATA: MOTIVATION



- *Situation:* No hyperplane can separate the data points correctly
- *Approach:*
    - Determine appropriate penalties for bad points
    - Solve original problem, by involving penalties

Non Separable Data: Motivation II

Let $(\mathbf{x}_i, y_i), i = 1, ...n$ be training data, where

- $\mathbf{x}_i = (x_{i1}, ..., x_{id})$,
- $y_i \in \{-1, +1\}$

and let $\mathbf{w} = (w_1, ..., w_d)$.

Non Separable Data: Motivation II

Let $(\mathbf{x}_i, y_i), i = 1, ...n$ be training data, where

- $\mathbf{x}_i = (x_{i1}, ..., x_{id})$,
- $y_i \in \{-1, +1\}$

and let $\mathbf{w} = (w_1, ..., w_d)$.

Minimize the following function:

$$f(\mathbf{w}, b) = \frac{1}{2}\sum_{j=1}^{d} w_j^2 + C\sum_{i=1}^{n} \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\} \qquad (11)$$

Non Separable Data: Motivation II

$$f(\mathbf{w}, b) = \underbrace{\frac{1}{2}\sum_{j=1}^{d} w_j^2}_{\text{Seek minimal } ||\mathbf{w}||} + \underbrace{C\sum_{i=1}^{n} \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}}_{\text{Bad point penalty}}$$

- Minimizing $||\mathbf{w}||$ equivalent to minimizing monotone function of $||\mathbf{w}||$
  ☞ Minimizing f seeks minimal $||\mathbf{w}||$
- Vectors $\mathbf{w}$ and training data balanced in terms of basic units:

$$\frac{\partial(||\mathbf{w}||^2/2)}{\partial w_i} = w_i \quad \text{and} \quad \frac{\partial(\sum_{j=1}^{d} w_j x_{ij} + b)}{\partial w_i} = x_{ij}$$

- C is a regularization parameter
  - Large C: minimize misclassified points, but accept narrow margin
  - Small C: accept misclassified points, but widen margin

Non Separable Data: Motivation II

$$f(\mathbf{w}, b) = \underbrace{\frac{1}{2}\sum_{j=1}^{d} w_j^2}_{\text{Seek minimal } ||\mathbf{w}||} + \underbrace{C\sum_{i=1}^{n} \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}}_{\text{Bad point penalty}}$$

- Minimizing $||\mathbf{w}||$ equivalent to minimizing monotone function of $||\mathbf{w}||$
  ☞ Minimizing f seeks minimal $||\mathbf{w}||$
- Vectors $\mathbf{w}$ and training data balanced in terms of basic units:

$$\frac{\partial(||\mathbf{w}||^2/2)}{\partial w_i} = w_i \quad \text{and} \quad \frac{\partial(\sum_{j=1}^{d} w_j x_{ij} + b)}{\partial w_i} = x_{ij}$$

- C is a regularization parameter
  - Large C: minimize misclassified points, but accept narrow margin
  - Small C: accept misclassified points, but widen margin

Non Separable Data: Motivation II

$$f(\mathbf{w}, b) = \underbrace{\frac{1}{2}\sum_{j=1}^{d} w_j^2}_{\text{Seek minimal } ||\mathbf{w}||} + \underbrace{C\sum_{i=1}^{n} \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}}_{\text{Bad point penalty}}$$

- Minimizing $||\mathbf{w}||$ equivalent to minimizing monotone function of $||\mathbf{w}||$
  ☞ Minimizing f seeks minimal $||\mathbf{w}||$
- Vectors $\mathbf{w}$ and training data balanced in terms of basic units:

$$\frac{\partial(||\mathbf{w}||^2/2)}{\partial w_i} = w_i \quad \text{and} \quad \frac{\partial(\sum_{j=1}^{d} w_j x_{ij} + b)}{\partial w_i} = x_{ij}$$

- C is a regularization parameter
  - Large C: minimize misclassified points, but accept narrow margin
  - Small C: accept misclassified points, but widen margin

Non Separable Data: Motivation II

$$f(\mathbf{w}, b) = \underbrace{\frac{1}{2}\sum_{j=1}^{d} w_j^2}_{\text{Seek minimal } ||\mathbf{w}||} + \underbrace{C\sum_{i=1}^{n} \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}}_{\text{Bad point penalty}}$$

- Minimizing $||\mathbf{w}||$ equivalent to minimizing monotone function of $||\mathbf{w}||$
  ☞ Minimizing f seeks minimal $||\mathbf{w}||$
- Vectors $\mathbf{w}$ and training data balanced in terms of basic units:

$$\frac{\partial(||\mathbf{w}||^2/2)}{\partial w_i} = w_i \quad \text{and} \quad \frac{\partial(\sum_{j=1}^{d} w_j x_{ij} + b)}{\partial w_i} = x_{ij}$$

- C is a regularization parameter
  - Large C: minimize misclassified points, but accept narrow margin
  - Small C: accept misclassified points, but widen margin

Non Separable Data: Motivation II

$$f(\mathbf{w}, b) = \underbrace{\frac{1}{2}\sum_{j=1}^{d} w_j^2}_{\text{Seek minimal } ||\mathbf{w}||} \underbrace{+ C\sum_{i=1}^{n} \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}}_{\text{Bad point penalty}}$$

- Minimizing $||\mathbf{w}||$ equivalent to minimizing monotone function of $||\mathbf{w}||$
  ☞ Minimizing f seeks minimal $||\mathbf{w}||$
- Vectors $\mathbf{w}$ and training data balanced in terms of basic units:

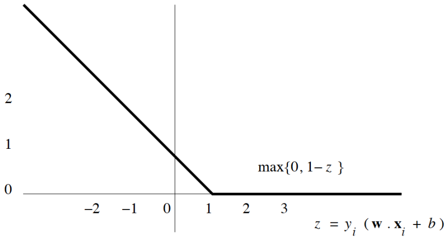
$$\frac{\partial(||\mathbf{w}||^2/2)}{\partial w_i} = w_i \quad \text{and} \quad \frac{\partial(\sum_{j=1}^{d} w_j x_{ij} + b)}{\partial w_i} = x_{ij}$$

- C is a regularization parameter
  - Large C: minimize misclassified points, but accept narrow margin
  - Small C: accept misclassified points, but widen margin

Non Separable Data: Hinge Function

Let the *hinge function* L be defined by

$$L(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\} \tag{12}$$

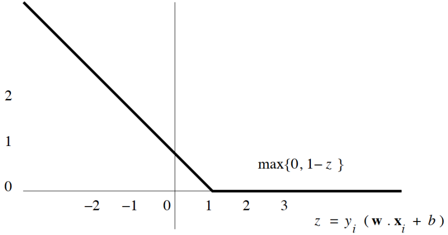


- $L(\mathbf{x}_i, y_i) = 0$ iff $\mathbf{x}_i$ on the correct side of hyperplane with sufficient margin
- The worse $\mathbf{x}_i$ is located the greater $L(\mathbf{x}_i, y_i)$

Non Separable Data: Hinge Function

Let the *hinge function* L be defined by

$$L(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\} \tag{12}$$

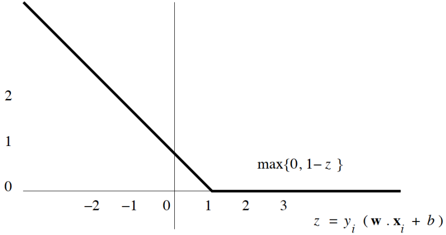


- $L(\mathbf{x}_i, y_i) = 0$ iff $\mathbf{x}_i$ on the correct side of hyperplane with sufficient margin
- The worse $\mathbf{x}_i$ is located the greater $L(\mathbf{x}_i, y_i)$

Non Separable Data: Hinge Function

Let the *hinge function* L be defined by

$$L(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\} \tag{12}$$



- $L(\mathbf{x}_i, y_i) = 0$ iff $\mathbf{x}_i$ on the correct side of hyperplane with sufficient margin
- The worse $\mathbf{x}_i$ is located the greater $L(\mathbf{x}_i, y_i)$

NON SEPARABLE DATA: HINGE FUNCTION

Let the *hinge function* L be defined by

$$L(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}$$

Non Separable Data: Hinge Function

Let the *hinge function* L be defined by

$$L(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}$$

Partial derivatives of hinge function:

$$\frac{\partial L}{\partial w_j} = \begin{cases} 0 & \text{if } y_i(\sum_{j=1}^{d} w_j x_{ij} + b) \geq 1 \\ -y_i x_{ij} & \text{otherwise} \end{cases} \tag{13}$$

Non Separable Data: Hinge Function

Let the *hinge function* L be defined by

$$L(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}$$

Partial derivatives of hinge function:

$$\frac{\partial L}{\partial w_j} = \begin{cases} 0 & \text{if } y_i(\sum_{j=1}^{d} w_j x_{ij} + b) \geq 1 \\ -y_i x_{ij} & \text{otherwise} \end{cases} \tag{13}$$

Reflecting:

- If $\mathbf{x}_i$ is on right side with sufficient margin: nothing to be done
- Otherwise adjust w_j to have $\mathbf{x}_i$ better placed

Non Separable Data: Hinge Function

Let the *hinge function* L be defined by

$$L(\mathbf{x}_i, y_i) = \max\{0, 1 - y_i(\sum_{j=1}^{d} w_j x_{ij} + b)\}$$

Partial derivatives of hinge function:

$$\frac{\partial L}{\partial w_j} = \begin{cases} 0 & \text{if } y_i(\sum_{j=1}^{d} w_j x_{ij} + b) \geq 1 \\ -y_i x_{ij} & \text{otherwise} \end{cases} \tag{13}$$

Reflecting:

- If $\mathbf{x}_i$ is on right side with sufficient margin: nothing to be done
- Otherwise adjust w_j to have $\mathbf{x}_i$ better placed

General / Further Reading

Literature

- Mining Massive Datasets , Chapter 12, Section 3: `http://infolab.stanford.edu/~ullman/mmds/ch12.pdf`

Thank you for listening!