ECE595 / STAT598: Machine Learning I
Lecture 17.2: Perceptron 2 - Optimality

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

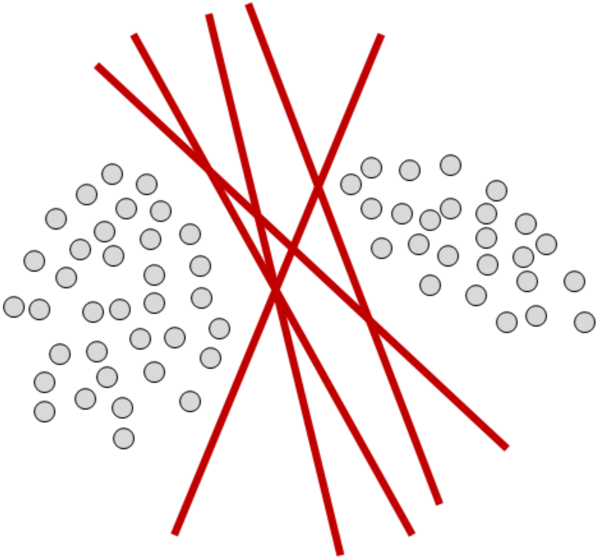
Discriminative Approaches

- Lecture 16 Perceptron 1: Definition and Basic Concepts
- Lecture 17 Perceptron 2: Algorithm and Property
- Lecture 18 Multi-Layer Perceptron: Back Propagation

This lecture: Perceptron 2

- Perceptron Algorithm
  - Loss Function
  - Algorithm
- Optimality
  - Uniqueness
  - Batch and Online Mode
- Convergence
  - Main Results
  - Implication

Non-uniqueness of Global Minimizer



Optimality of Perceptron Algorithm

- Let **perceptron algorithm** output

$$\boldsymbol{\theta}^*_{\text{perceptron}} = \text{Perceptron Algorithm}(\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}).$$

- Let **ideal** solution

$$\boldsymbol{\theta}^*_{\text{hard}} = \underset{\boldsymbol{\theta}}{\text{argmin}} \; J_{\text{hard}}(\boldsymbol{\theta}).$$

  That means

$$J_{\text{hard}}(\boldsymbol{\theta}^*_{\text{hard}}) \leq J_{\text{hard}}(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta}.$$

- If the two classes are linearly separable, then $\boldsymbol{\theta}^*_{\text{perceptron}}$ is a global minimizer:

$$J_{\text{hard}}(\boldsymbol{\theta}^*_{\text{perceptron}}) \leq J_{\text{hard}}(\boldsymbol{\theta}), \quad \forall \boldsymbol{\theta}.$$

  and

$$J_{\text{hard}}(\boldsymbol{\theta}^*_{\text{perceptron}}) = J_{\text{hard}}(\boldsymbol{\theta}^*_{\text{hard}}) = 0.$$

Uniqueness of Perceptron Solution

- If $\boldsymbol{\theta}^*$ minimizes $J_{\text{hard}}(\boldsymbol{\theta}^*)$, then $\alpha\boldsymbol{\theta}^*$ for some constant $\alpha > 0$ also minimizes $J_{\text{hard}}(\boldsymbol{\theta}^*)$.
- This is because

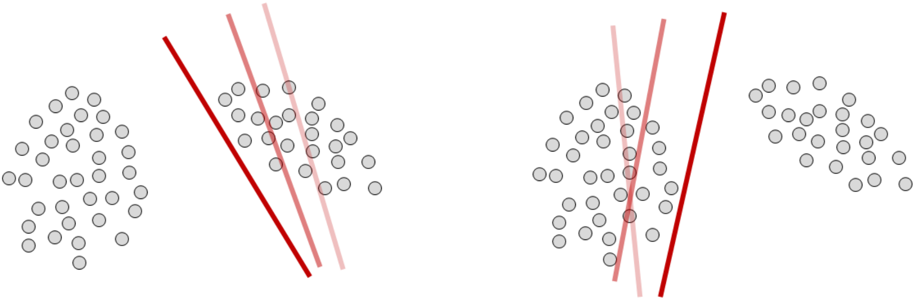
$$\begin{aligned} g_{\alpha\boldsymbol{\theta}}(\boldsymbol{x}) &= (\alpha\boldsymbol{w})^T\boldsymbol{x} + (\alpha w_0) \\ &= \alpha(\boldsymbol{w}^T\boldsymbol{x} + w_0). \end{aligned}$$

- If $g_{\boldsymbol{\theta}}(\boldsymbol{x}) > 0$, then $g_{\alpha\boldsymbol{\theta}}(\boldsymbol{x}) > 0$. So if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = +1$, then $h_{\alpha\boldsymbol{\theta}}(\boldsymbol{x}) = +1$.
- If $g_{\boldsymbol{\theta}}(\boldsymbol{x}) < 0$, then $g_{\alpha\boldsymbol{\theta}}(\boldsymbol{x}) < 0$. So if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) = -1$, then $h_{\alpha\boldsymbol{\theta}}(\boldsymbol{x}) = -1$.
- The sign of $\boldsymbol{w}^T\boldsymbol{x} + w_0$ is unchanged as long as $\alpha > 0$.

$$\begin{aligned} J_{\text{hard}}(\boldsymbol{\theta}^*) &= \sum_{j=1}^{N} \max\left\{ -y_j h_{\boldsymbol{\theta}^*}(\boldsymbol{x}_j), 0 \right\} \\ &= \sum_{j=1}^{N} \max\left\{ -y_j h_{\alpha\boldsymbol{\theta}^*}(\boldsymbol{x}_j), 0 \right\} = J_{\text{hard}}(\alpha\boldsymbol{\theta}^*) \end{aligned}$$

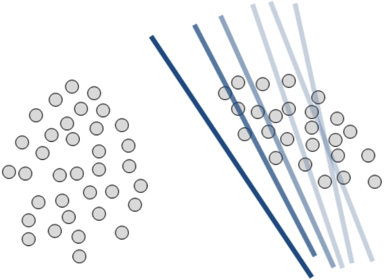
Factors for Uniqueness

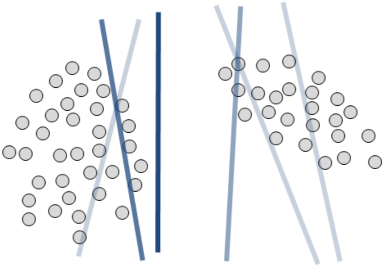
- **Initialization**
- Start at a different location, end on a different location
- You still converge, but no longer unique solution
- $\mathcal{M}_k$ changes



Factors for Uniqueness

- **Step Size**
- Too large step: oscillate
- Too small step: slow movement
- Terminates as long as no misclassification





Batch vs Online Mode

- **Batch mode**

$$\begin{bmatrix} \boldsymbol{w}^{(k+1)} \\ w_0^{(k+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}^{(k)} \\ w_0^{(k)} \end{bmatrix} + \alpha_k \sum_{j \in \mathcal{M}_k} \begin{bmatrix} y_j \boldsymbol{x}_j \\ y_j \end{bmatrix}.$$

  Update via the average of misclassified samples

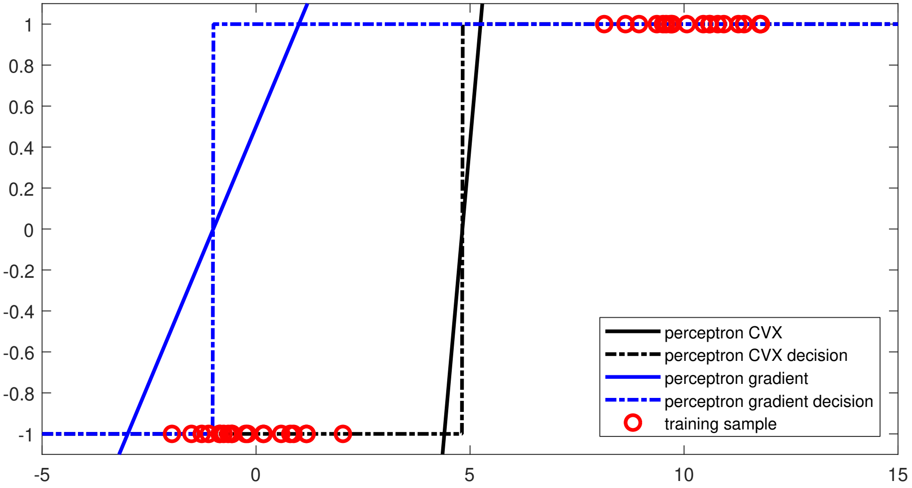
- **Online mode**

$$\begin{bmatrix} \boldsymbol{w}^{(k+1)} \\ w_0^{(k+1)} \end{bmatrix} = \begin{bmatrix} \boldsymbol{w}^{(k)} \\ w_0^{(k)} \end{bmatrix} + \alpha_k \begin{bmatrix} y_j \boldsymbol{x}_j \\ y_j \end{bmatrix},$$

  Update via a single misclassified sample

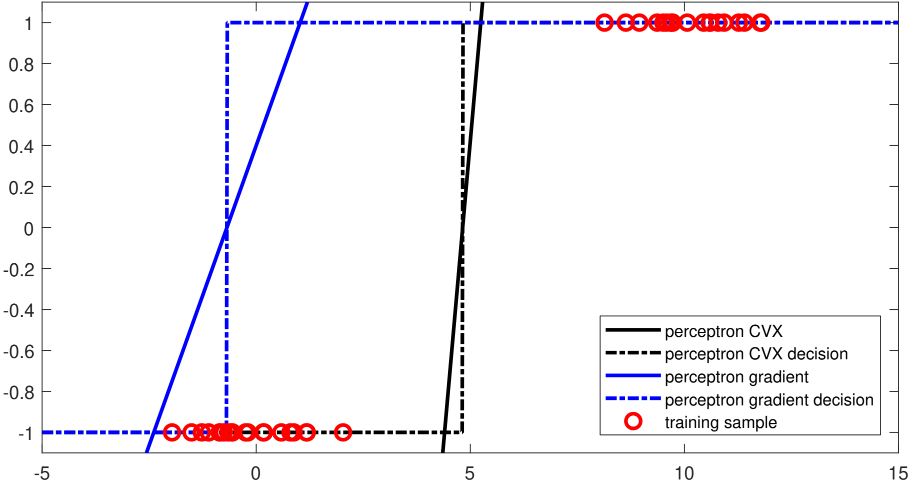
- j is a sample randomly picked from $\mathcal{M}_k$.
- Stochastic gradient descent.

Online Mode



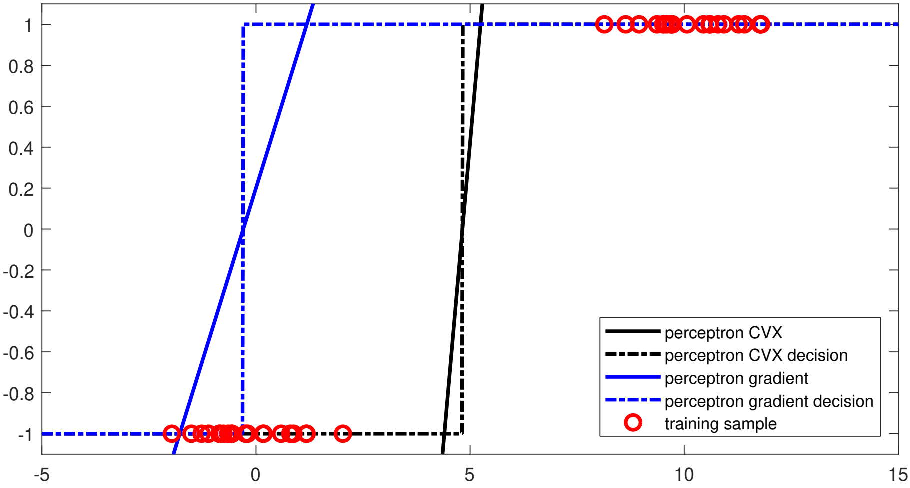
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

Online Mode



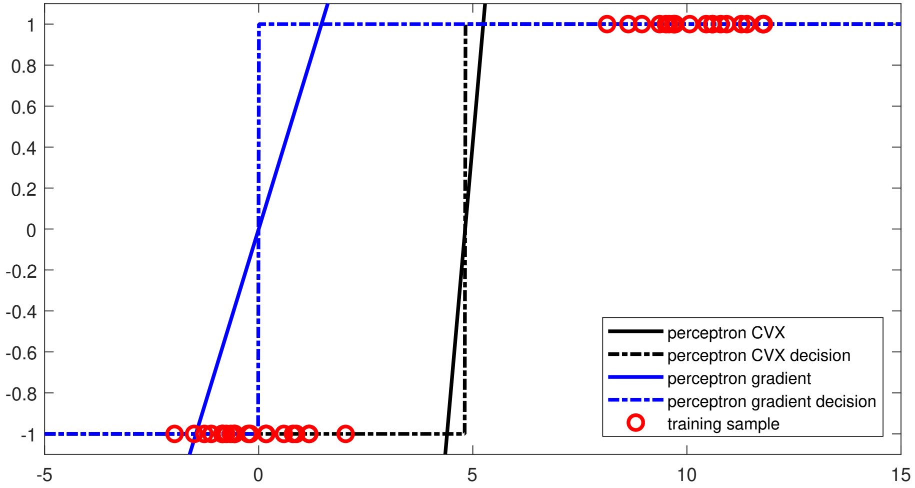
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

Online Mode



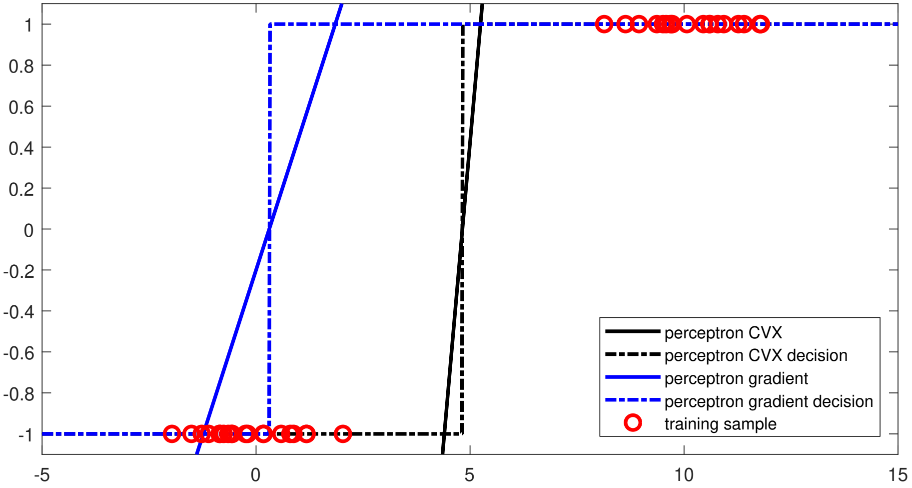
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

Online Mode

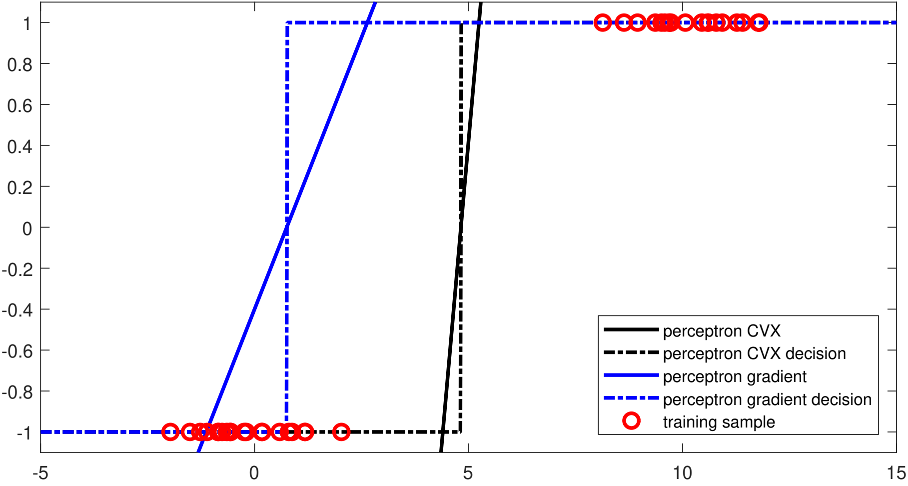


1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

Online Mode

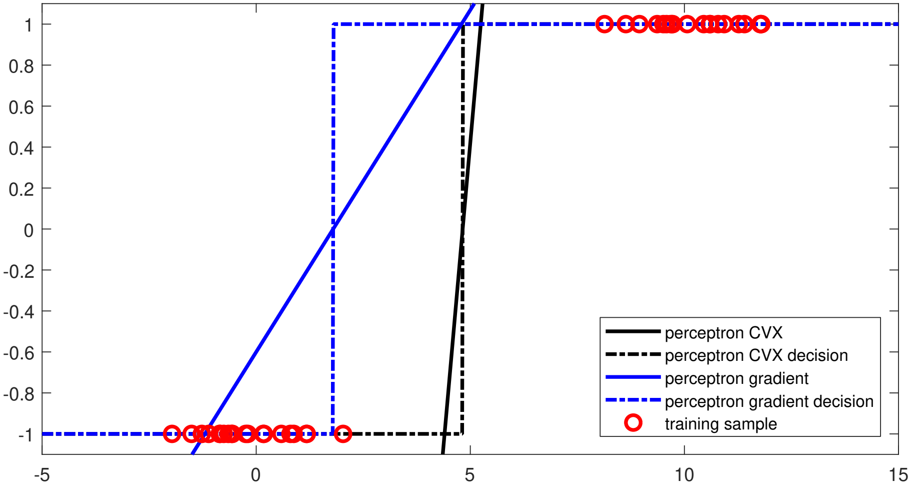

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

Online Mode



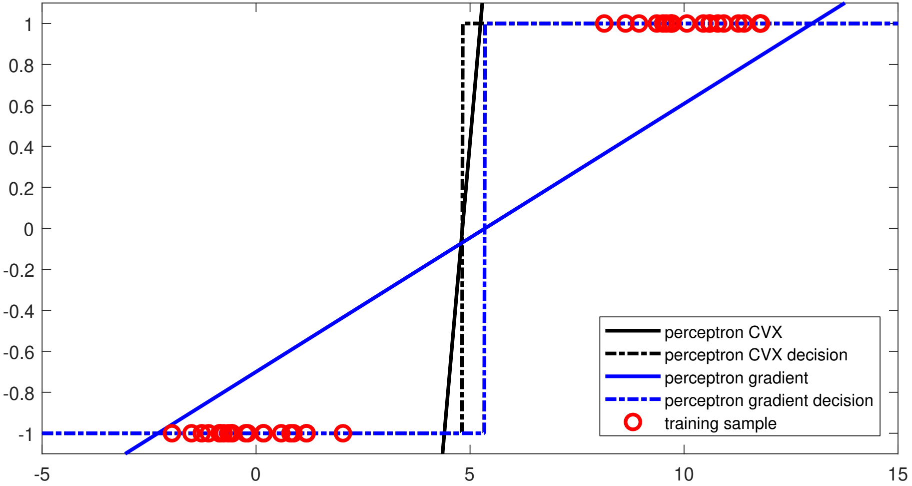
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

Online Mode



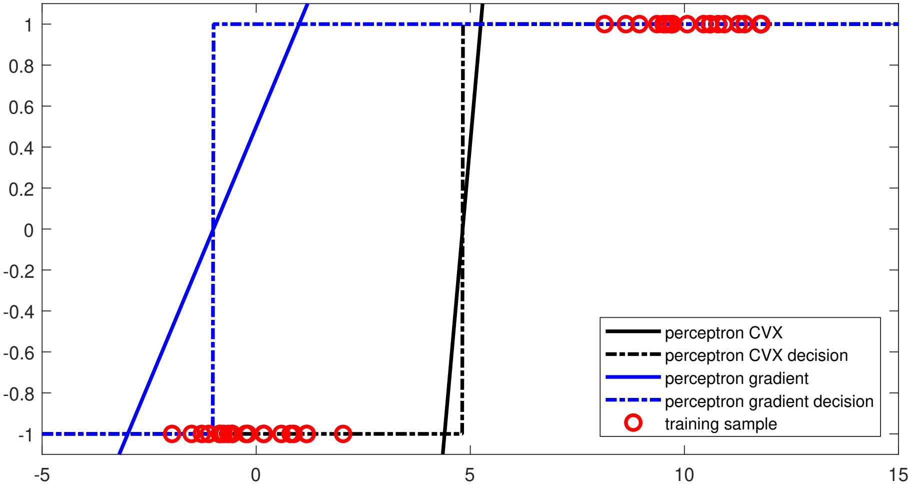
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

Online Mode



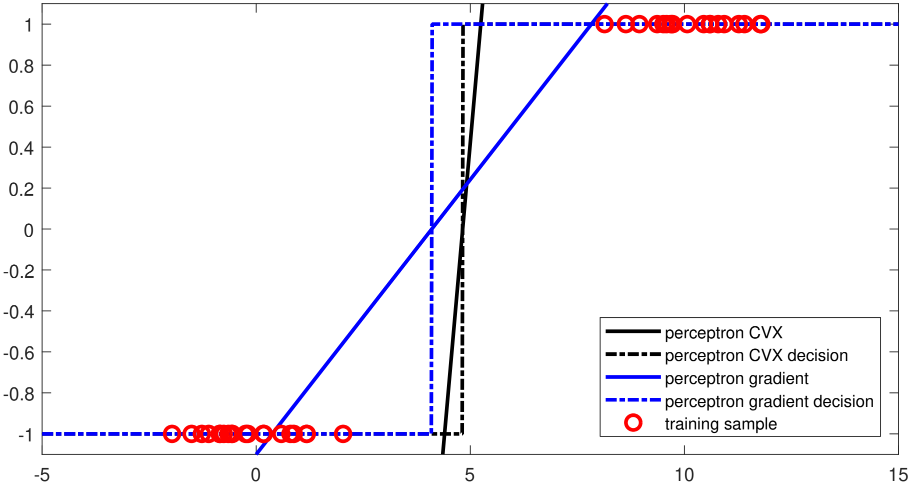
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

Batch Mode



1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

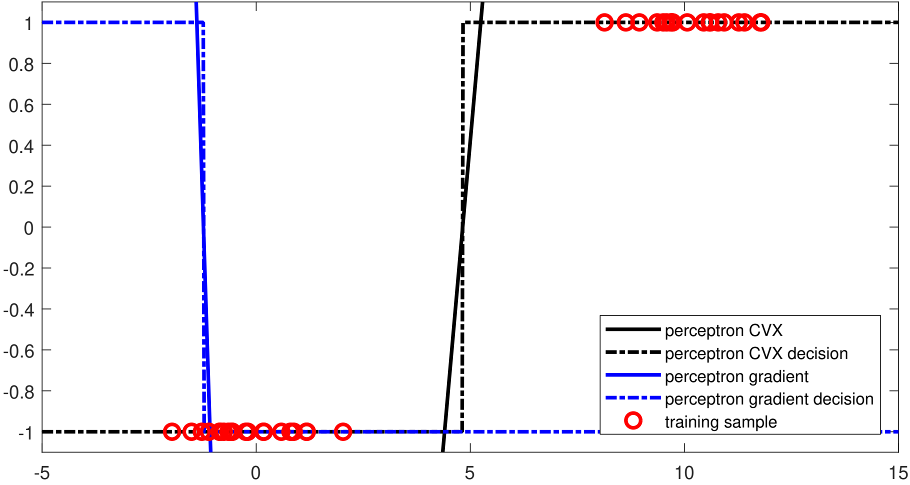
Batch Mode



1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1
-5
0
5
10
15
perceptron CVX
perceptron CVX decision
perceptron gradient
perceptron gradient decision
training sample

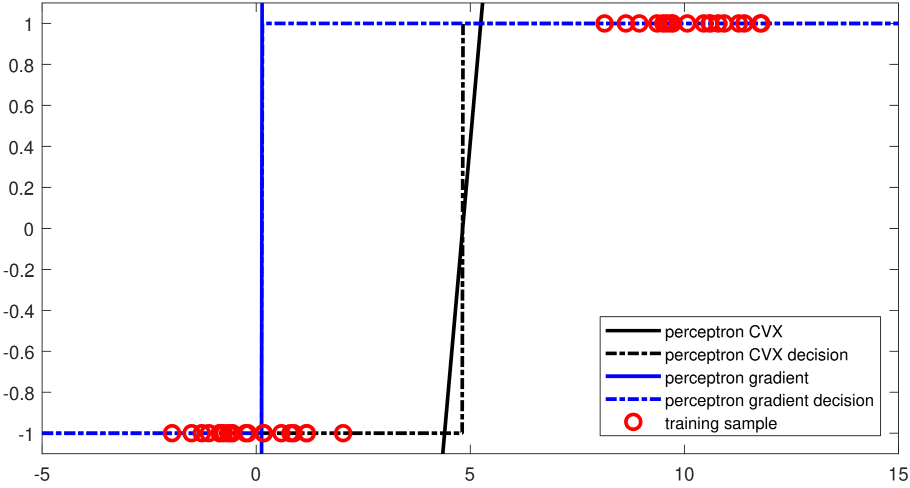
Step Size

Batch mode: Step size too large.



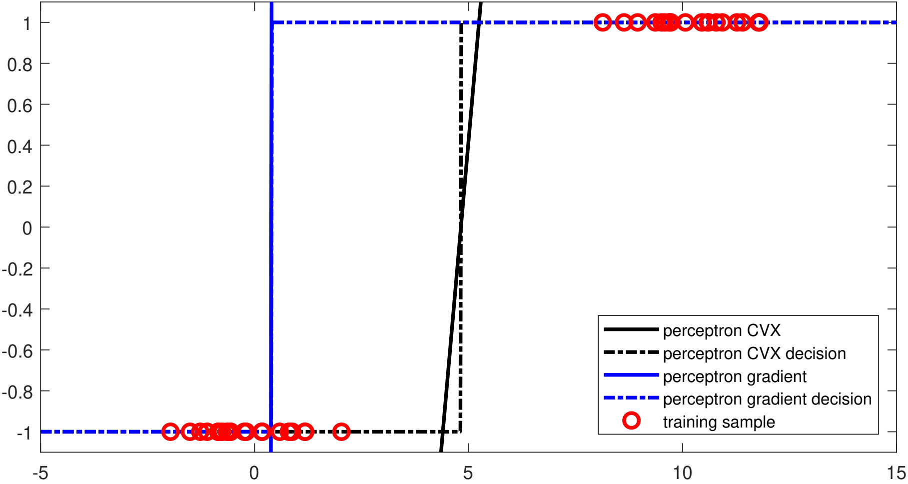
Step Size

Batch mode: Step size too large.



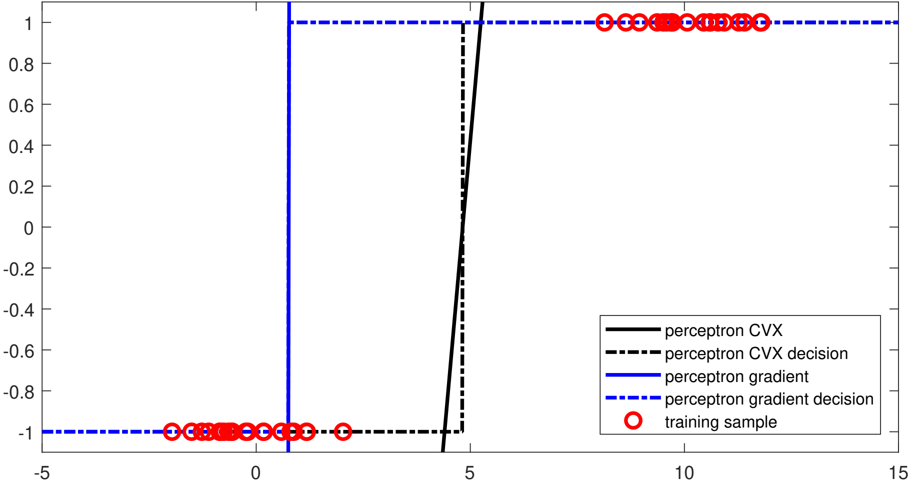
Step Size

Batch mode: Step size too large.



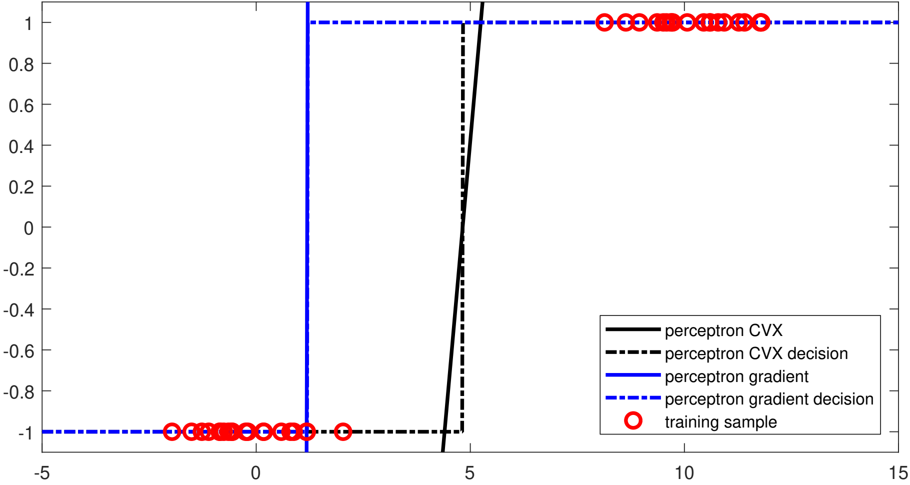
Step Size

Batch mode: Step size too large.

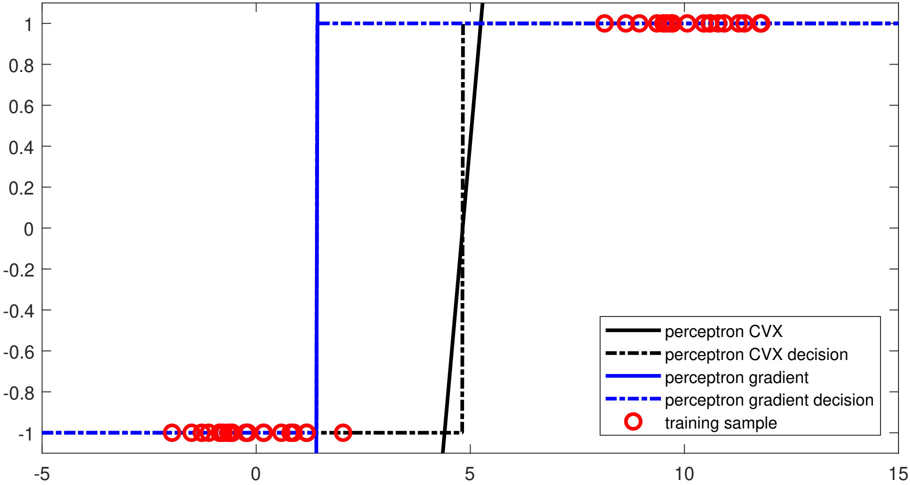


Step Size

Batch mode: Step size too large.

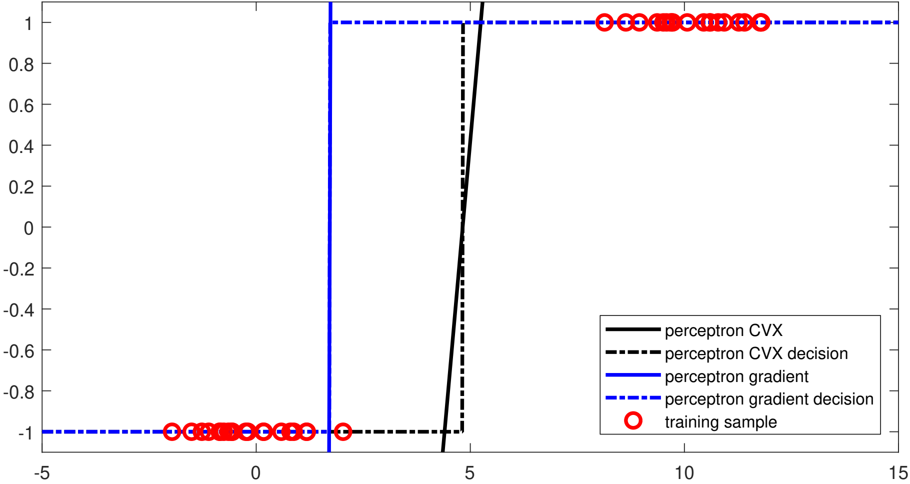


Batch mode: Step size too large.



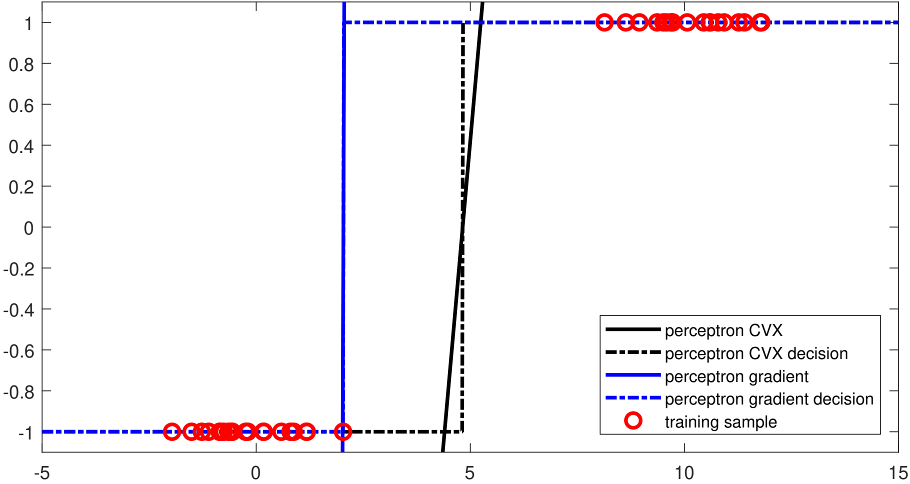
Step Size

Batch mode: Step size too large.

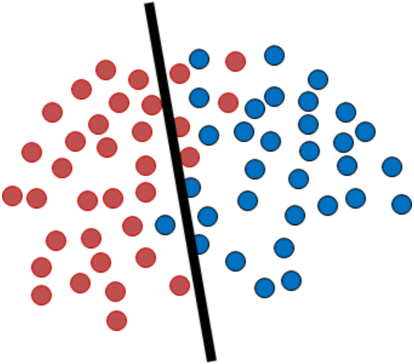


Step Size

Batch mode: Step size too large.



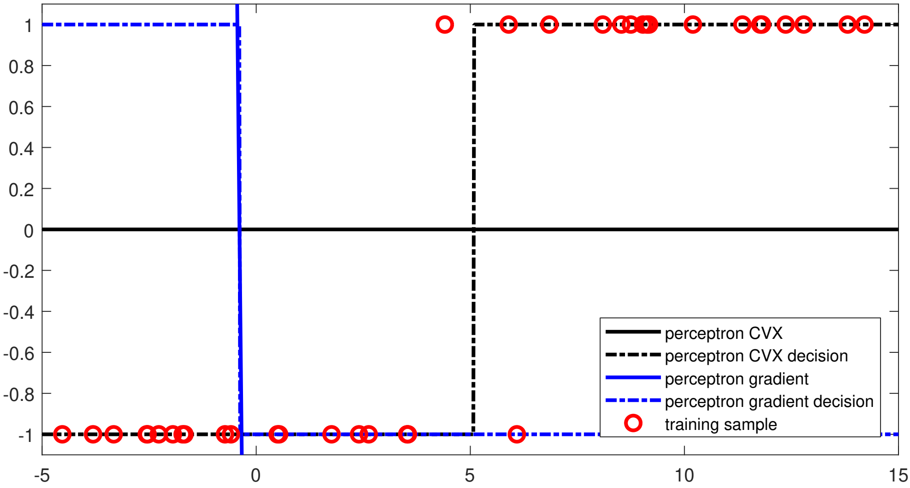
Linearly Not Separable



- No separating hyperplane
- CVX will still find you a solution
- But loss is no longer zero
- Perceptron algorithm will not converge

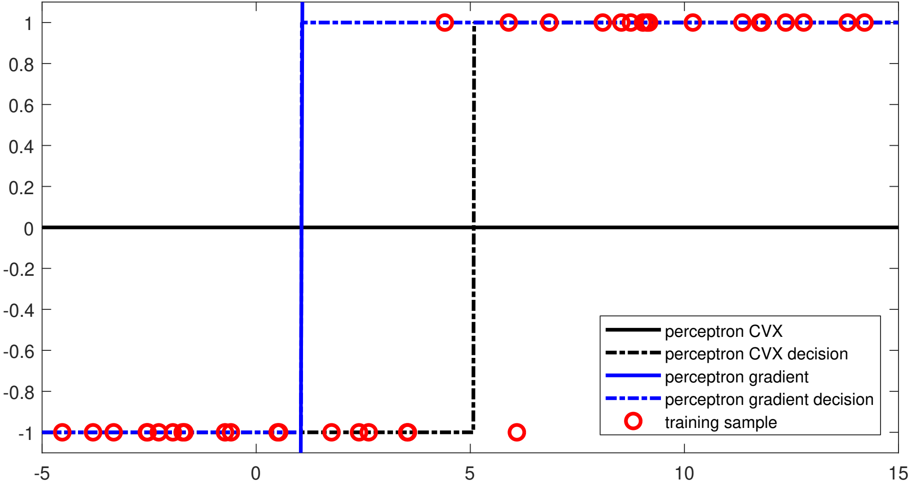
Linearly Not Separable

If the two classes are overlapping



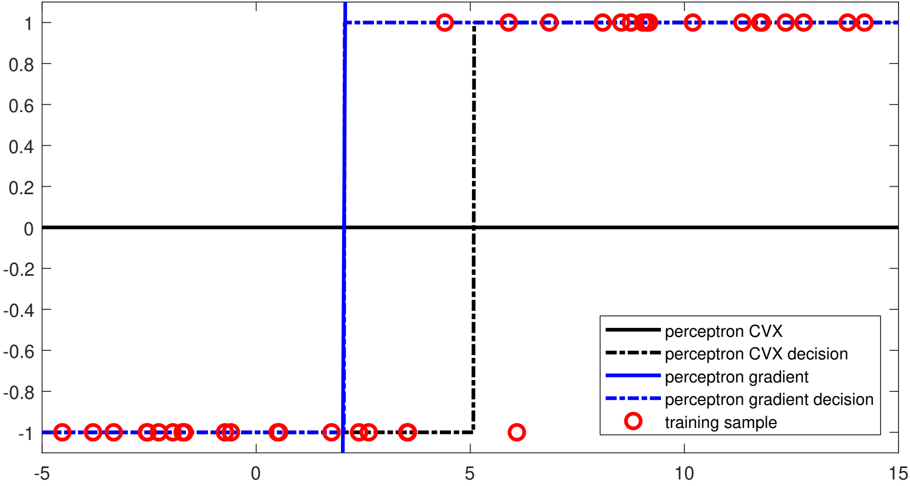
Linearly Not Separable

If the two classes are overlapping



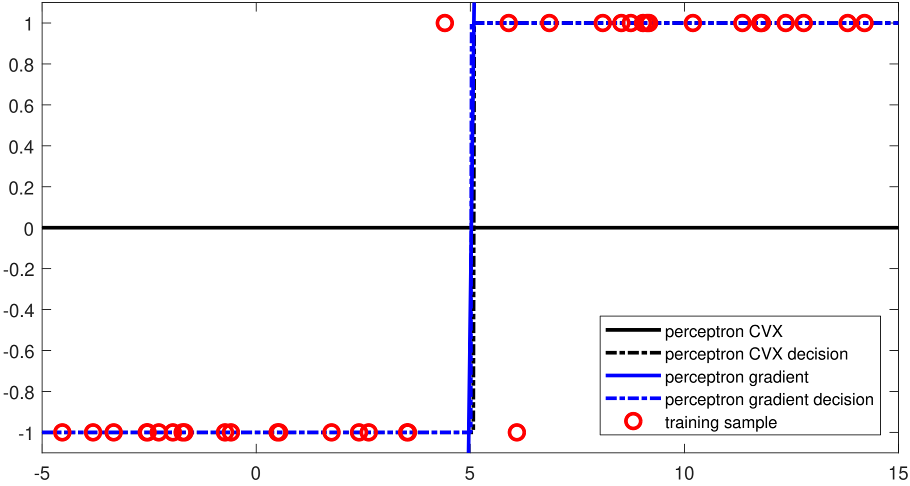
Linearly Not Separable

If the two classes are overlapping



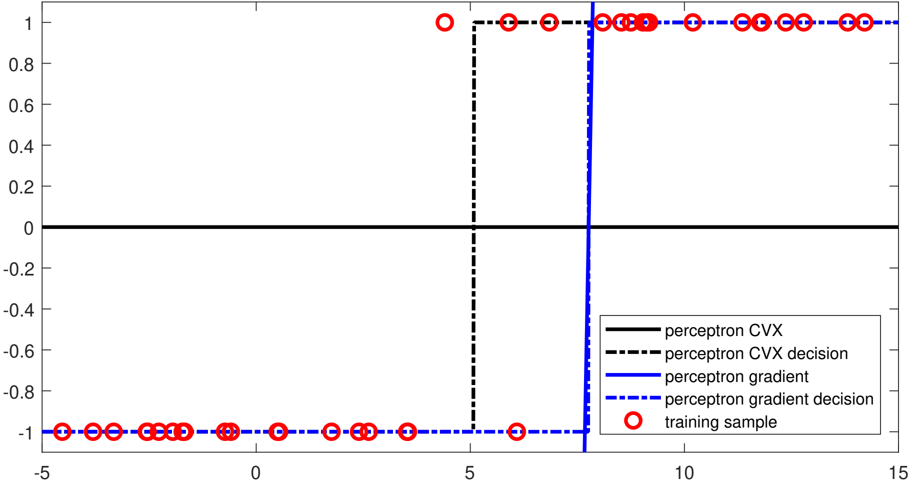
Linearly Not Separable

If the two classes are overlapping



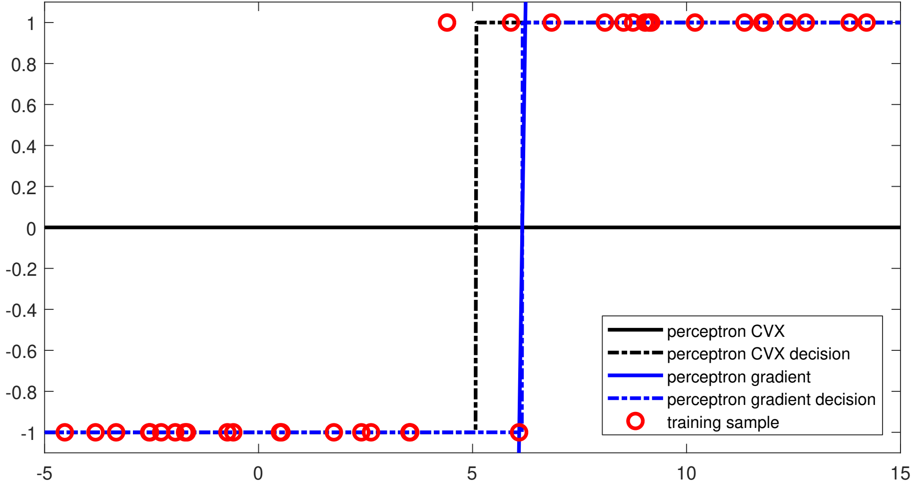
Linearly Not Separable

If the two classes are overlapping



Linearly Not Separable

If the two classes are overlapping



Linearly Not Separable

If the two classes are overlapping