

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 7 | Part 1

Maximum Margin Classifiers

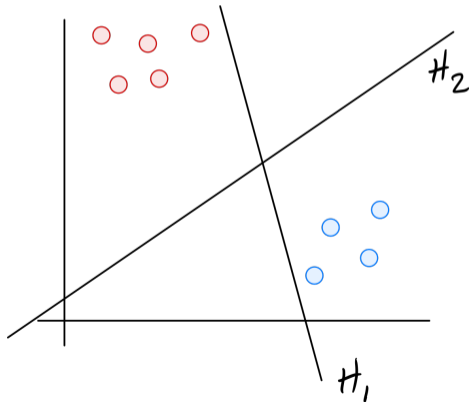
Recall: Perceptrons

- ▶ Linear classifier fit using loss function:

$$\ell_{\text{tron}}(H(\vec{x}), y) = \begin{cases} 0, & \text{sign}(H(\vec{x})) = y \\ |H(\vec{x})|, & \text{sign}(H(\vec{x})) \neq y \end{cases}$$

Exercise

What is the empirical risk with respect to the perceptron loss of H_1 ? What about H_2 ?

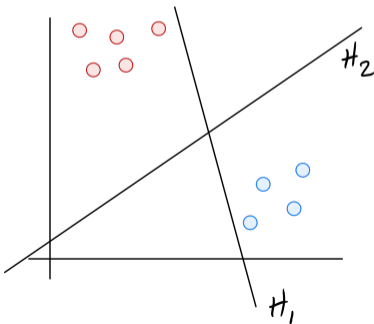


A **Problem** with the Perceptron

- ▶ **Recall:** the perceptron loss assigns no penalty to points that are correctly classified.
- ▶ No matter how close the point is to the boundary.
- ▶ **Problem:** we might learn decision boundary that is very close to the data (overfitting).

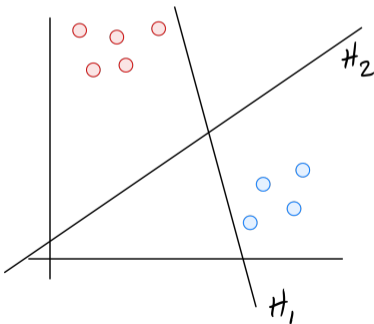
Linear Separability

- ▶ Data are **linearly separable** if there exists a linear classifier which perfectly classifies the data.



Margin

- ▶ The **margin** is the smallest distance between the decision boundary and a training point.



Maximum Margin Classifier

- ▶ If training data are linearly separable, there are **many** classifiers with zero error.
- ▶ We prefer classifiers with larger margins.
 - ▶ Better generalization performance.
- ▶ Can we find the **maximum margin classifier**?
 - ▶ I.e., the classifier with the largest possible margin?

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 7 | Part 2

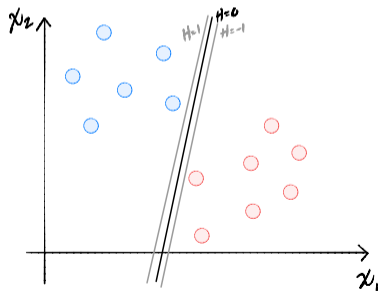
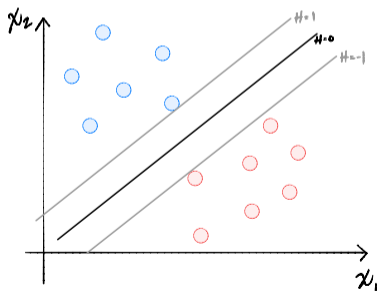
Hard SVM Optimization Problem

Goal

- ▶ Write down an optimization problem that, assuming linear separability, ensures:
 1. All points are classified correctly.
 2. The margin is as large as possible.

The Exclusion Zone

- ▶ Define the **exclusion zone** as the region where $|H(\vec{x})| < 1$.
 - ▶ That is, between $H(\vec{x}) = 1$ and $H(\vec{x}) = -1$.



Maximizing the Margin

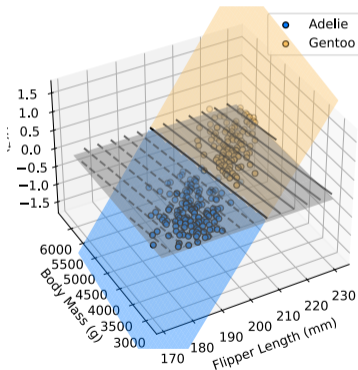
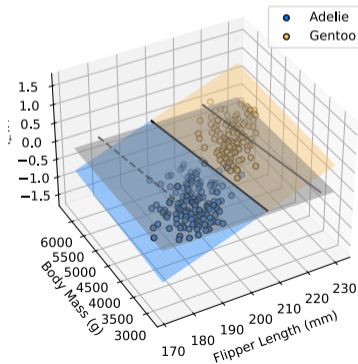
- ▶ The margin and the exclusion zone are related.
 - ▶ Width of exclusion zone \leq the margin.
- ▶ Maximizing the margin is **equivalent** to maximizing the width of the exclusion zone.

Goal

- ▶ We want to find a linear predictor H that:
 1. Classifies all points correctly.
 2. Has no training points in the exclusion zone.
 3. Has the widest exclusion zone possible.

Claim

- ▶ The width of the exc. zone is controlled by $\|\vec{w}\|$.
 - ▶ Larger $\|\vec{w}\|$ \Rightarrow steeper H \Rightarrow **narrower** exclusion zone.
 - ▶ Smaller $\|\vec{w}\|$ \Rightarrow shallower H \Rightarrow **wider** exclusion zone.



a

^aSee: [http:](http://)

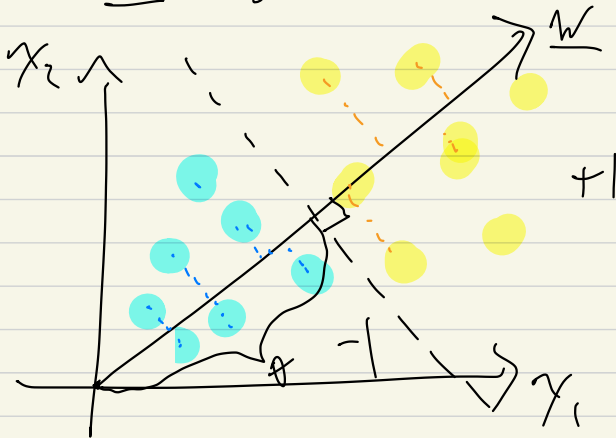
[//dsc140a.com/static/vic/cym/](http://dsc140a.com/static/vic/cym/)

Two Different Spaces

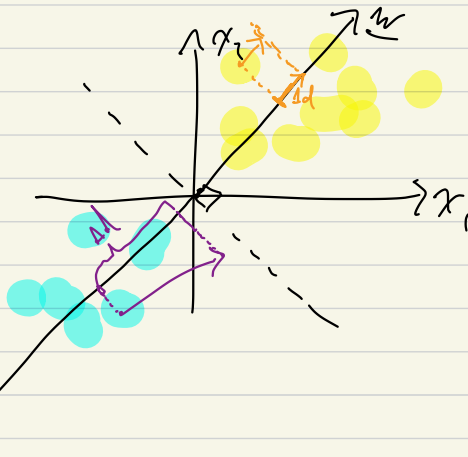
- ▶ There are **two** spaces we need to keep straight:
 - ▶ **Data space:** \mathbb{R}^d , where the training points \vec{x}_i live.
 - ▶ **Function space:** the scalar output of $H(\vec{x}) = \vec{w} \cdot \vec{x} + b$.
- ▶ The map H sends a point in data space to a scalar in function space.

linear classifier

$$\underline{x} \cdot \underline{w} \geq \theta$$



$$A(\underline{x}) = \underline{w} \cdot \text{Aug}(\underline{x}) \\ = \underbrace{\sum_{i=1}^d w_i x_i}_{\text{inner product}} + \underbrace{w_0}_{\text{Threshold}}$$



Where Do the Margins Live?

- ▶ The margin planes are defined as $H(\vec{x}) = \pm 1$.
- ▶ The ± 1 is a value in **function space** — not a distance!
 - ▶ It is just a convention to pin down the scale of \vec{w} .
 - ▶ We could have picked ± 7 or ± 0.3 ; the geometry is unchanged.
- ▶ But we want to maximize the exclusion zone width in **data space**.
 - ▶ That is, the **actual Euclidean distance** between the margin planes.

The Key Question

- ▶ The margins are fixed at ± 1 in function space.
- ▶ How far apart are they in **data space**?
- ▶ Equivalently: how much physical distance in data space corresponds to a change of 1 unit in function space?

Intuition: H as a Stretch

- ▶ Think of H as **stretching** data space onto the real line.
- ▶ If $\|\vec{w}\|$ is **large**, a small step in data space causes a **large** jump in H .
 - ▶ The ± 1 level sets are geometrically **close together**.
- ▶ If $\|\vec{w}\|$ is **small**, a large step in data space causes only a small jump in H .
 - ▶ The ± 1 level sets are geometrically **far apart**.
- ▶ So $\|\vec{w}\|$ is the **stretch factor** between data space and function space.

Math: Step in Data Space

- ▶ Take a point \vec{x} and step a distance Δd in the direction of \vec{w} :

$$\vec{x} \longrightarrow \vec{x} + \Delta d \cdot \frac{\vec{w}}{\|\vec{w}\|}$$

Handwritten annotations:
- "Scalar" points to Δd .
- "vector" points to \vec{w} .
- "Scalar" points to $\|\vec{w}\|$.
- "The direction of \vec{w} " points to the fraction $\frac{\vec{w}}{\|\vec{w}\|}$.

- ▶ How much does H change?

$$\Delta f = H\left(\vec{x} + \Delta d \cdot \frac{\vec{w}}{\|\vec{w}\|}\right) - H(\vec{x}) = \vec{w} \cdot \left(\Delta d \cdot \frac{\vec{w}}{\|\vec{w}\|}\right) = \Delta d \cdot \frac{\|\vec{w}\|^2}{\|\vec{w}\|}$$

$$\Delta f = \|\vec{w}\| \cdot \Delta d$$

The Conversion Formula

- ▶ Rearranging:

$$\Delta d = \frac{\Delta f}{\|\vec{w}\|}$$

- ▶ A change of Δf in **function space** corresponds to a physical distance of $\Delta f / \|\vec{w}\|$ in **data space**.
- ▶ $\|\vec{w}\|$ is literally the **exchange rate** between the two spaces.

Width of the Exclusion Zone

- ▶ In function space, the margins are at ± 1 , so $\Delta f = 2$.
- ▶ In data space, the width of the exclusion zone is:

$$\text{margin width} = \frac{\Delta f}{\|\vec{w}\|} = \frac{2}{\|\vec{w}\|}$$

- ▶ So the exclusion zone width is **inversely proportional** to $\|\vec{w}\|$.

Maximize Margin \Leftrightarrow Minimize $\|\vec{w}\|$

- ▶ We want to **maximize** the margin width in data space:

$$\max_{\vec{w}, b} \frac{2}{\|\vec{w}\|} \quad \Leftrightarrow \quad \min_{\vec{w}, b} \|\vec{w}\| \quad \Leftrightarrow \quad \min_{\vec{w}, b} \frac{1}{2} \|\vec{w}\|^2$$

- ▶ The three problems have the **same solution**.
 - ▶ $2/x$ is monotonically decreasing for $x > 0$, so maximizing $2/\|\vec{w}\|$ is minimizing $\|\vec{w}\|$.
 - ▶ Squaring is monotonic on $[0, \infty)$, so minimizing $\|\vec{w}\|$ is minimizing $\|\vec{w}\|^2$.
 - ▶ The $\frac{1}{2}$ is for convenience (clean gradient: $\nabla \frac{1}{2} \|\vec{w}\|^2 = \vec{w}$).

Summary

- ▶ The ± 1 margins live in **function space** (a convention).
- ▶ The exclusion zone we care about lives in **data space**.
- ▶ $\|\vec{w}\|$ is the stretch factor converting between the two:

$$\Delta f = \|\vec{w}\| \cdot \Delta d$$

- ▶ Width of exclusion zone in data space is $2 / \|\vec{w}\|$.

Goal

- ▶ We want to find a linear predictor H that:
 1. Classifies all points correctly.
 2. Has no training points in the exclusion zone.
 3. Has the widest exclusion zone possible.
 - ▶ That is, has the smallest possible $\|\vec{w}\|$.

Goal

- ▶ We want to find a linear predictor H that:
 1. Classifies all points correctly.
 2. Has no training points in the exclusion zone.
 3. Has the widest exclusion zone possible.
 - ▶ That is, has the smallest possible $\|\vec{w}\|$.
- ▶ Now let's write down a formal optimization problem.

Goal

▶ We want to find a $H(\vec{x}) = \vec{w} \cdot \text{Aug}(\vec{x})$ that:

1. Classifies all points correctly.

▶ That is, $\text{sign}(H(\vec{x}^{(i)})) = y_i$ for all i .

$y_i \in \{-1, 1\} \forall i$

Goal

- ▶ We want to find a $H(\vec{x}) = \vec{w} \cdot \text{Aug}(\vec{x})$ that:
 1. Classifies all points correctly.
 - ▶ That is, $\text{sign}(H(\vec{x}^{(i)})) = y_i$ for all i .
 2. Has no training points in the exclusion zone.
 - ▶ That is, $|H(\vec{x}^{(i)})| \geq 1$ for all i .

Goal

- ▶ We want to find a $H(\vec{x}) = \vec{w} \cdot \text{Aug}(\vec{x})$ that:
 1. Classifies all points correctly.
 - ▶ That is, $\text{sign}(H(\vec{x}^{(i)})) = y_i$ for all i .
 2. Has no training points in the exclusion zone.
 - ▶ That is, $|H(\vec{x}^{(i)})| \geq 1$ for all i .
 3. Has the widest exclusion zone possible.
 - ▶ That is, has the smallest possible $\|\vec{w}\|$.

Goal

- ▶ Out of all \vec{w} satisfying:
 1. $\text{sign}(H(\vec{x}^{(i)})) = y_i$ for all i . (all predictions correct)
 2. $|H(\vec{x}^{(i)})| \geq 1$ for all i . (no points in exclusion zone)
- ▶ Find one with smallest $\|\vec{w}\|$
 - ▶ (Largest exclusion zone)

Hard-SVM Optimization Problem

- ▶ The **Hard Support Vector Machine** optimization problem is:

$$\vec{w}^* = \arg \min_{\vec{w}} \|\vec{w}\|$$

subject to:

$$\text{sign}(\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})) = y_i;$$

$$|\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})| \geq 1$$

for all i .

constraint optimization

Modifications

- ▶ The Hard-SVM problem is often stated slightly differently.

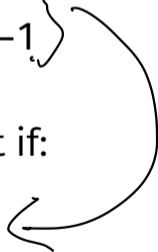
Observation #1

- ▶ A point is classified correctly when:

$$\left. \begin{array}{l} \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) > 0, \quad \text{if } y_i = 1 \\ \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) < 0, \quad \text{if } y_i = -1 \end{array} \right\}$$

- ▶ Equivalently, classification is correct if:

$$y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) > 0$$



Hard-SVM Optimization Problem

- ▶ The **Hard Support Vector Machine** optimization problem is:

$$\vec{w}^* = \arg \min_{\vec{w}} \|\vec{w}\|$$

subject to:

$$y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) > 0 \quad \leftarrow \text{correctness}$$

$$|\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})| \geq 1 \quad \leftarrow \text{exclusion}$$

for all i .

Observation #2

- ▶ Since $y_i = \pm 1$, we can simplify the constraints:

$$y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) > 0 \quad \Leftarrow$$

$$|\vec{w} \cdot \text{Aug}(\vec{x}^{(i)})| \geq 1 \quad \Leftarrow$$

becomes simply:

$$y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) \geq 1$$

Simpler Constraints

- ▶ That is, $\vec{x}^{(i)}$ is 1) correctly classified and 2) outside of the exclusion zone if and only if:

$$y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) \geq 1$$

Observation #3

- ▶ Minimizing $\|\vec{w}\|^2$ or $\|\vec{w}\|$ gives same solution.
 - ▶ allows us to use **quadratic programming** solvers.

Hard-SVM Optimization Problem

- ▶ The Hard-SVM optimization problem is:

$$\vec{w}^* = \arg \min_{\vec{w}} \|\vec{w}\|^2$$

subject to:

$$y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) \geq 1$$

for all i .

Hard-SVM

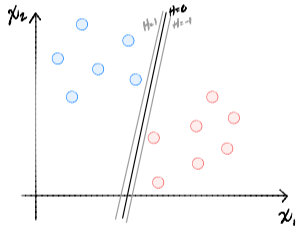
- ▶ This optimization problem is called the **Hard Support Vector Machine** classifier problem.
- ▶ Only makes sense if data are linearly separable.
- ▶ In a moment, we'll see the Soft-SVM.

How?

- ▶ Turn it into a **convex quadratic** optimization problem:
 - ▶ Minimize $\|\vec{w}\|^2$ subject to $y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) \geq 1$ for all i .
- ▶ Can be solved efficiently with **quadratic programming**.
 - ▶ But there is no exact general formula for the solution

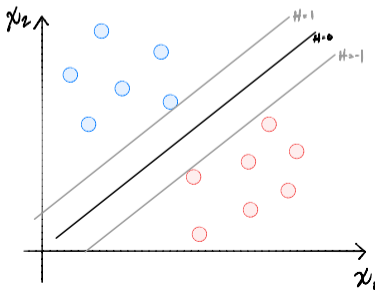
Exercise

Can the below predictor be a solution of the Hard-SVM?



SVMs are Maximum Margin Classifiers

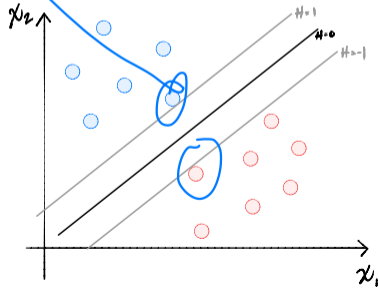
- ▶ Intuition says solutions of Hard-SVM will have large margins.
- ▶ Fact: they maximize the margin.



Support Vectors

- A **support vector** is a training point $\vec{x}^{(i)}$ such that

$$y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) = 1$$



Support Vectors

- ▶ **Fact:** the solution to Hard-SVM is always a linear combination of the support vectors.
- ▶ That is, let S be the set of support vectors. Then

$$\vec{w}^* = \sum_{i \in S} y_i \alpha_i \text{Aug}(\vec{x}^{(i)})$$

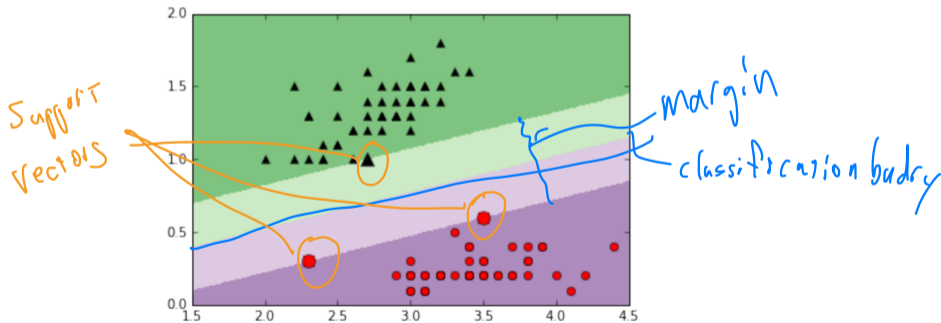
Example: Irises



- ▶ 3 classes: *iris setosa*, *iris versicolor*, *iris virginica*
- ▶ 4 measurements: petal width/height, sepal width/height

Example: Irises

- ▶ Using only sepal width/petal width
- ▶ Two classes: versicolor (black), setosa (red)



DSC 140A

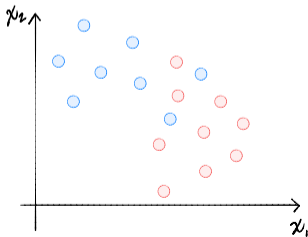
Probabilistic Modeling & Machine Learning

Lecture 7 | Part 3

Soft-Margin SVMs

Non-Separability

- ▶ So far we've assumed data is linearly separable.
- ▶ What if it isn't?

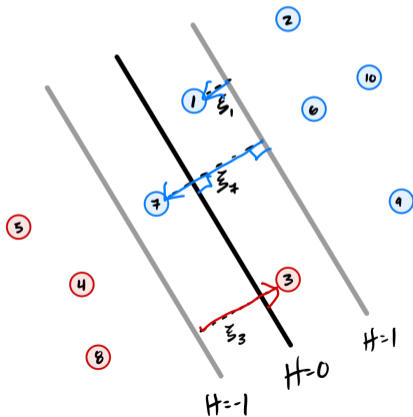


The Problem

- ▶ **Old Goal:** Minimize $\|\vec{w}\|^2$ subject to $y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) \geq 1$ for all i .
- ▶ This **no longer makes sense**.

Cut Some Slack

- ▶ **Idea:** allow some classifications to be ξ_i wrong, but not too wrong.



" x_i "
"slack variables"

Cut Some Slack

- ▶ **New problem.** Fix some number $C \geq 0$.

$$\min_{\vec{w} \in \mathbb{R}^{d+1}, \vec{\xi} \in \mathbb{R}^n} \|\vec{w}\|^2 + C \sum_{i=1}^n \xi_i$$

subject to $\underbrace{y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)})}_{\text{correctness}} \geq 1 - \underbrace{\xi_i}_{\text{error}}$ for all i , $\xi_i \geq 0$.

The Slack Parameter, C

- ▶ C controls how much slack is given.

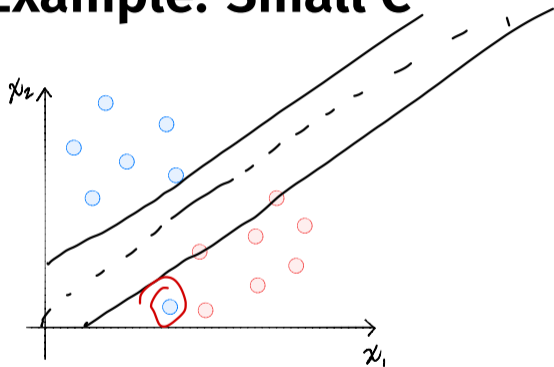
Trade off
between errors
+ margin

$$\min_{\vec{w} \in \mathbb{R}^{d+1}, \vec{\xi} \in \mathbb{R}^n} \|\vec{w}\|^2 + C \sum_{i=1}^n \xi_i$$

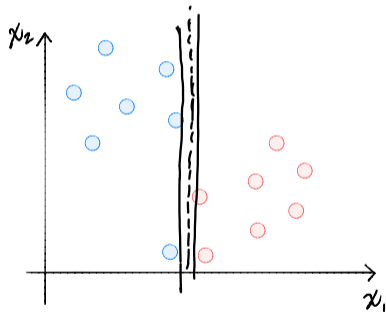
subject to $y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) \geq 1 - \xi_i$ for all i , $\xi_i \geq 0$.

- ▶ Large C: don't give much slack. Avoid misclassifications.
- ▶ Small C: allow more slack at the cost of misclassifications.

Example: Small C



Example: Large C



Soft and Hard Margins

- ▶ Max-margin SVM from before has **hard margin**.
- ▶ Now: the **soft margin** SVM.
- ▶ As $C \rightarrow \infty$, the margin hardens. \Leftrightarrow hard margin is a special case

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 7 | Part 4

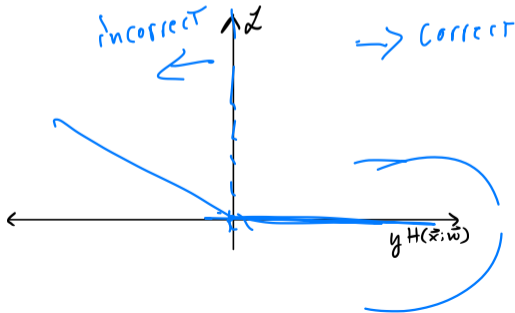
Hinge Loss

Loss Functions?

- ▶ So far, we've learned predictors by minimizing **expected loss** via ERM.
- ▶ But this isn't what we did with Hard-SVM and Soft-SVM.
- ▶ It turns out, we can frame Soft-SVM as an ERM problem.

Recall: Perceptron Loss

$$\ell_{\text{tron}}(H(\vec{x}), y) = \begin{cases} 0, & \text{sign}(H(\vec{x})) = y \\ |H(\vec{x})|, & \text{sign}(H(\vec{x})) \neq y \end{cases}$$

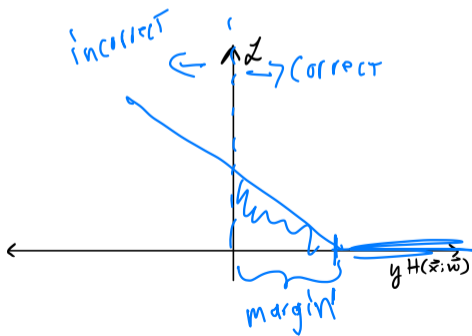


Perceptron Loss

- ▶ Perceptron loss did not penalize correct classifications.
- ▶ Even if they were very close to boundary.
- ▶ **Idea:** penalize predictions that are close to the boundary, too.

The Hinge Loss

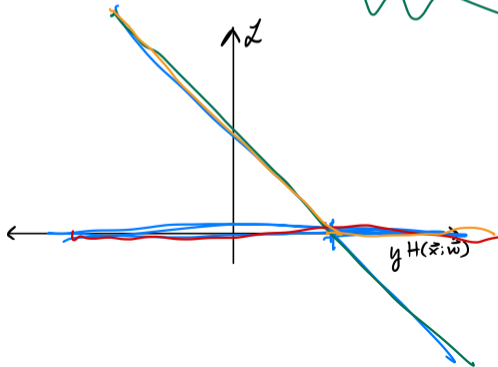
$$l_{\text{hinge}}(H(\vec{x}), y) = \begin{cases} 0, & yH(\vec{x}) \geq 1, \\ 1 - yH(\vec{x}), & yH(\vec{x}) < 1 \end{cases}$$



outside of margin
inside of margin

The Hinge Loss

$$\ell_{\text{hinge}}(H(\vec{x}), y) = \max\{0, 1 - yH(\vec{x})\}$$



Equivalence

- ▶ Recall the Soft-SVM problem:

$$\min_{\vec{w} \in \mathbb{R}^{d+1}, \vec{\xi} \in \mathbb{R}^n} \|\vec{w}\|^2 + C \sum_{i=1}^n \xi_i$$

subject to $y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)}) \geq 1 - \xi_i$ for all i , $\xi_i \geq 0$.

- ▶ **Note:** if $\vec{x}^{(i)}$ is misclassified, then

$$\xi_i = 1 - y_i \vec{w} \cdot \text{Aug}(\vec{x}^{(i)})$$

↳ good solution

Equivalence

↳ "correct solution"

max margin

- ▶ The Soft-SVM problem is equivalent to finding \vec{w} that minimizes:

$$R_{\text{svm}}(\vec{w}) = \|\vec{w}\|^2 + C \sum_{i=1}^n \max\{0, 1 - y_i \vec{w} \cdot \vec{x}^{(i)}\}$$

Min error

risk

- ▶ R_{svm} is the **regularized** risk.

⇒ no more conditions

- ▶ C is a parameter affecting "softness" of boundary; chosen by you.

Another Way to Optimize

- ▶ In practice, SGD is often used to train soft SVMs.

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 7 | Part 5

Demo: Sentiment Analysis

Why use linear predictors?

- ▶ Linear classifiers look to be very simple.
- ▶ That can be both **good** and **bad**.
 - ▶ **Good**: the math is tractable, less likely to overfit
 - ▶ **Bad**: may be too simple, underfit
- ▶ They can work surprisingly well.

Sentiment Analysis

- ▶ **Given:** a piece of text.
- ▶ **Determine:** if it is **positive** or **negative** in tone
- ▶ Example: “Needless to say, I wasted my money.”

The Data

- ▶ Sentences from reviews on Amazon, Yelp, IMDB.
- ▶ Each labeled (by a human) **positive** or **negative**.
- ▶ Examples:
 - ▶ **“Needless to say, I wasted my money.”**
 - ▶ **“I have to jiggle the plug to get it to line up right.”**
 - ▶ **“Will order from them again!”**
 - ▶ **“He was very impressed when going from the original battery to the extended battery.”**

The Plan

- ▶ We'll train a soft-margin SVM.
- ▶ **Problem:** SVMs take **fixed-length vectors** as inputs, not sentences.

Bags of Words

To turn a document into a fixed-length vector:

- ▶ First, choose a **dictionary** of words:
 - ▶ E.g.: ["wasted", "impressed", "great", "bad", "again"]
- ▶ Count number of occurrences of each dictionary word in document.
 - ▶ "It was bad. So bad that I was impressed at how bad it was." → $(0, 1, 0, 3, 0)^T$
- ▶ This is called a **bag of words** representation.

Choosing the Dictionary

- ▶ Many ways of choosing the dictionary.
- ▶ Easiest: take all of the words in the training set.
 - ▶ Perhaps throw out **stop words** like “the”, “a”, etc.
- ▶ Resulting dimensionality of feature vectors:
large.

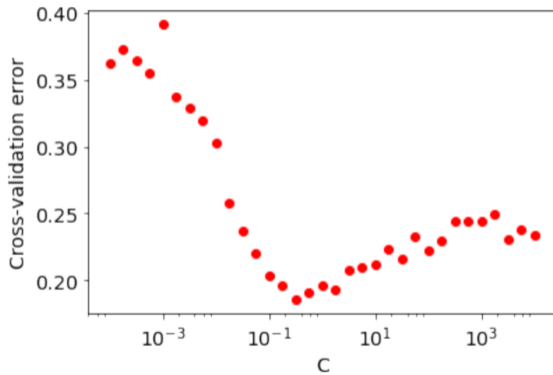
Experiment

- ▶ Bag of words features with 4500 word dictionary.
- ▶ 2500 training sentences, 500 test sentences.
- ▶ Train a soft margin SVM.

Choosing C

- ▶ We have to choose the slack parameter, C .
- ▶ Use **cross validation!**

Cross Validation



Results

- ▶ With $C = 0.32$, test error $\approx 15.6\%$.

C	training error (%)	test error (%)	# support vectors
0.01	23.72	28.4	2294
0.1	7.88	18.4	1766
1	1.12	16.8	1306
10	0.16	19.4	1105
100	0.08	19.4	1035
1000	0.08	19.4	950