
CS 181 LECTURE 2/22/24

Scribe Notes

Contents

1	Admin	3
2	Today: Max Margin	3
2.1	History of SVMs	4
3	Binary Classification Example	4
3.1	Degree of Freedom	6
3.2	Slack	7
3.3	Hard Margin SVM	7
3.4	Soft Margin SVM	7
4	Review of Losses for Classification	9
5	Concept Check	9
6	Concept Check Solutions	10

1 Admin

Today's lecture correlates with textbook chapters 5.1-5.3.

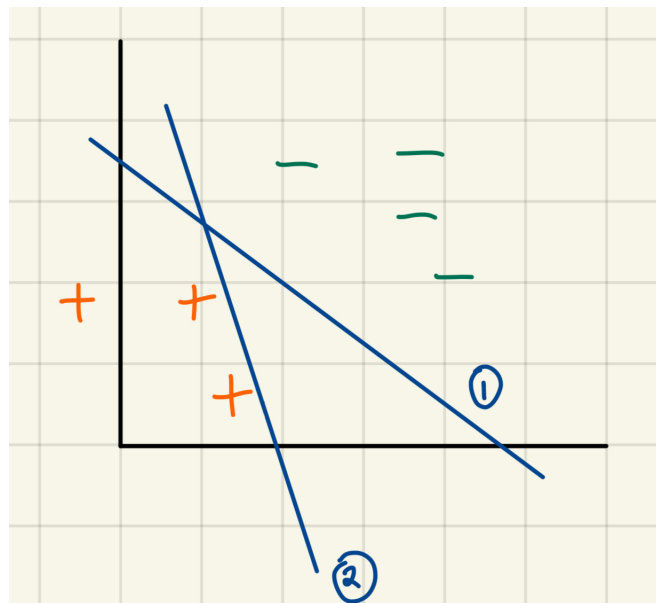
Next week we will talk about clustering but that will not be covered on Midterm I. We will release practice questions for Midterm I next week.

	non-prob	prob
disc.		
cont.		

2 Today: Max Margin

Max Margin is another loss.

Motivation: Say we have two classes to classify and two possible classifiers.



We would prefer line 1) because there is more space between the classifier and the datapoints. There are orange points that are really close to line 2), which intuitively make it a worse classifier.

Let us call the space between the closest point and the classifier the **margin**.

2.1 History of SVMs

SVMs were developed in the early 90's at the Bell Labs. SVMs were considered the best for around 20 years and even today SVMs are used a lot when there is limited data or you are worried about overfitting. Today we will be looking at SVMs in a linear classifier setting, but SVMs have a lot of real-world applications. There are a lot of applications to biomedical field.

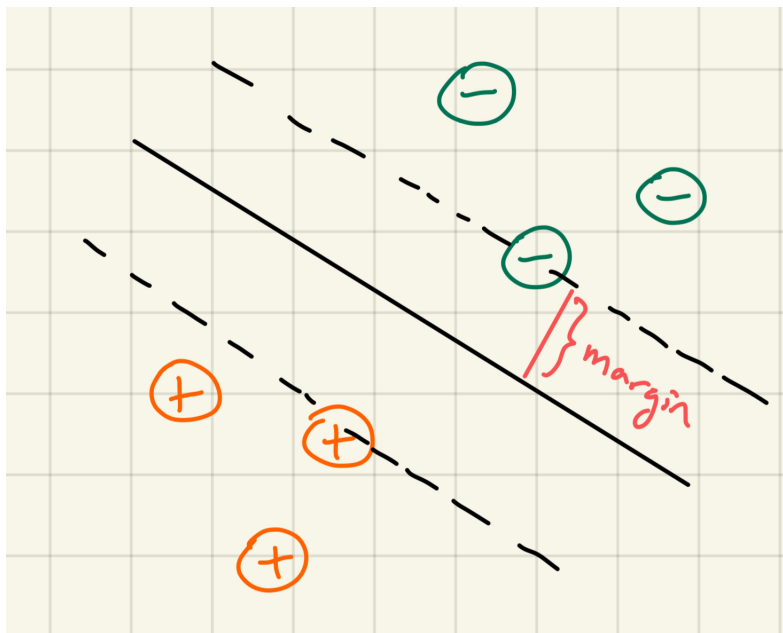
For example, SVMs were used for drug discoveries. When trying to predict if certain molecules have certain properties, researchers used an SVM. [Zernov '03]

Another application is cancer detection. Researchers would take a cell and extract gene expression data to run an SVM for predictions. [Furey et al '00]

3 Binary Classification Example

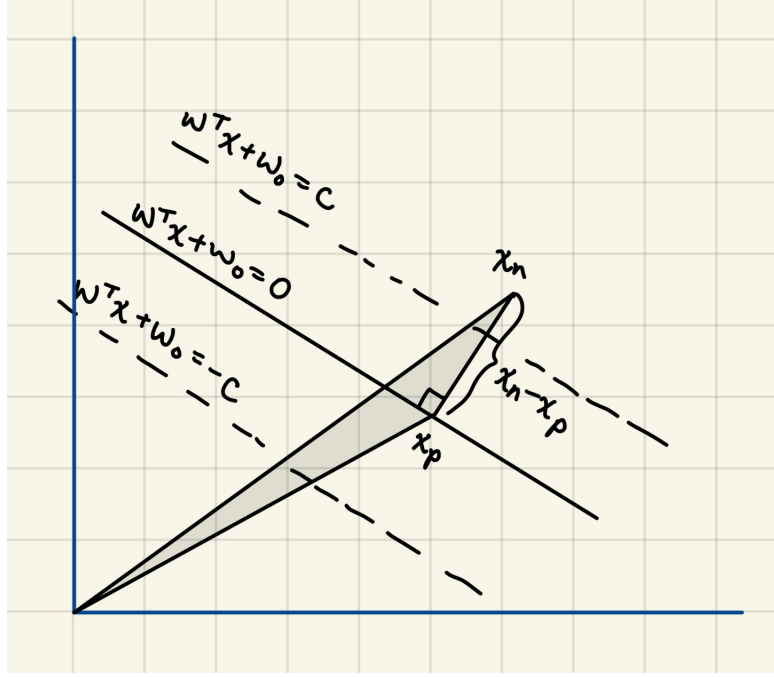
So there is the idea of having “space” so we get a classifier that generalizes better. The goal is to maximize the separation between the boundary and the closest points.

Let's assume we are doing binary classification, classifying points to the groups $\{-1, 1\}$.



Say we have a function $f(x) = w^\top x + w_0$.

Decompose the point x_n :



$$x_n = x_p + (x_n - x_p) = x_p + r \frac{w}{\|w\|_2}$$

We are looking for r , because that is our margin!

$$w^\top x_n = w^\top x_p + r \frac{w^\top w}{\|w\|_2}$$

Because x_p is on the boundary, then we know $w^\top x_p + w_0 = 0$.

$$\begin{aligned} w^\top x_n &= -w_0 + r \frac{w^\top w}{\|w\|_2} \\ r &= \frac{w^\top x_n + w_0}{\|w\|_2} \end{aligned}$$

This is the distance between a point and a hyperplane. Note also that this is a signed distance, meaning if the point is on the positive side, r will be positive, and if it is negative, then r will be negative.

To recognize whether the point was classified correctly or not, we will multiply by y_n .

$$\text{margin}(x_n; w, w_0) = \frac{y_n(w^\top x_n + w_0)}{\|w\|_2}$$

So, this is the margin for a single point, capturing both correctness and distance. The value will be greater than or equal to zero when the point is classified correctly and it will be negative otherwise.

Thus, the margin over the entire dataset is: $\text{margin}(X; w, w_0) = \min_n \{\text{margin}(x_n; w, w_0)\}$

Now, we want to maximize the margin.

$$\max_{w, w_0} \min_n \{\text{margin}(x_n; w, w_0)\}$$

3.1 Degree of Freedom

There is a degree of freedom that is easy to see when we multiply by a constant. The margin of a point $\text{margin}(x_n; \beta w, \beta w_0)$, assuming that $\beta > 0$, then we get the same margin because the constant gets canceled out.

$$\frac{y_n(\beta w^\top x_n + \beta w_0)}{\beta \|w\|_2} = \frac{y_n(w^\top x_n + w_0)}{\|w\|_2}$$

We will use that degree of freedom to our advantage.

Pick β such that the closest point has margin = 1.

$$\begin{aligned} & \max \min \frac{1}{\|w\|_2} y_n(w^\top x_n + w_0) \text{ s.t. } y_n(w^\top x_n + w_0) \geq 1 \forall n \\ &= \max \min \frac{1}{\|w\|_2} [1] \text{ s.t. } y_n(w^\top x_n + w_0) \geq 1 \forall n \\ &= \min \frac{1}{2} \|w\|^2 \text{ s.t. } y_n(w^\top x_n + w_0) \geq 1 \forall n \end{aligned}$$

Not that maximizing $1/\|w\|$ is the same as minimizing $\|w\|$.

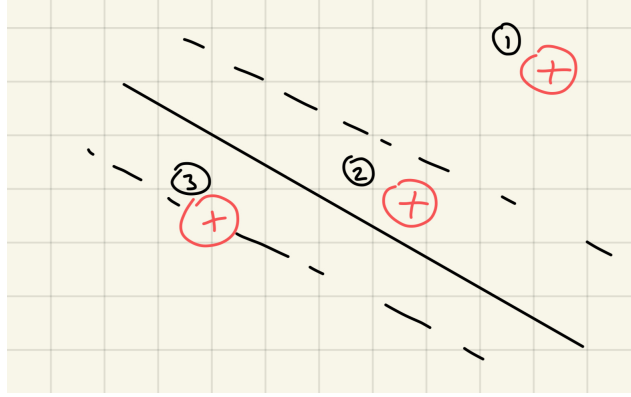
We have a quadratic formula and linear constraints, making this a Quadratic programming problem.

3.2 Slack

We can misclassify a few points to get a better margin for the rest of the points.

We have three types of points.

- 1) points that are classified correctly and very far from the boundary
- 2) points that are classified correctly but are too close to the boundary (violating a constraint)
- 3) points that are violating a constraint and misclassified



For point 1, we don't need any slack. For point 2), we need a slack between (0, 1). For point 3), we need a slack greater than 1.

The slack is used to quantify how much we are not satisfying the constraints.

3.3 Hard Margin SVM

In a hard margin SVM, the algorithm aims to find a hyperplane that perfectly separates the data points of different classes without any misclassifications. Mathematically, for a binary classification problem, we have:

Given training data points $\{(x_i, y_i)\}_{i=1}^N$, where x_i is the feature vector and y_i is the class label ($y_i \in \{-1, 1\}$), and N is the number of data points.

We aim to find a hyperplane represented by $w \cdot x + b = 0$, where w is the weight vector, b is the bias term, and $\cdot$ denotes the dot product.

The optimization problem can be formulated as:

$$\text{Minimize } \frac{1}{2} \|w\|^2$$

subject to:

$$y_i(w \cdot x_i + b) \geq 1, \quad \text{for all } i = 1, 2, \dots, N$$

This means that all data points must lie on the correct side of the decision boundary with a margin of at least 1. In hard margin SVM, even a single misclassification leads to infeasibility, indicating that the data is not linearly separable.

3.4 Soft Margin SVM

In real-world scenarios, it's rare to have perfectly separable data without noise or outliers. Soft margin SVM allows for some misclassifications by introducing a slack variable (ξ) to

the optimization problem. This slack variable allows some data points to fall within the margin or even on the wrong side of the decision boundary.

The optimization problem for soft margin SVM becomes:

$$\text{Minimize } \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi_i$$

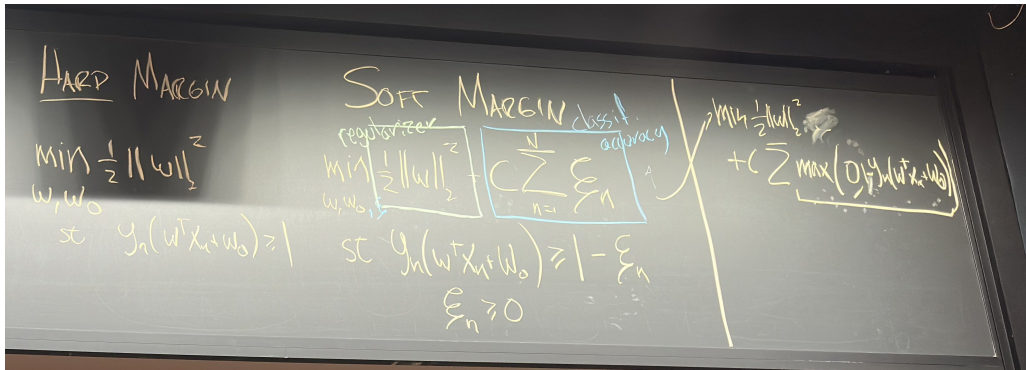
subject to:

$$y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad \text{for all } i = 1, 2, \dots, N$$

$$\xi_i \geq 0, \quad \text{for all } i = 1, 2, \dots, N$$

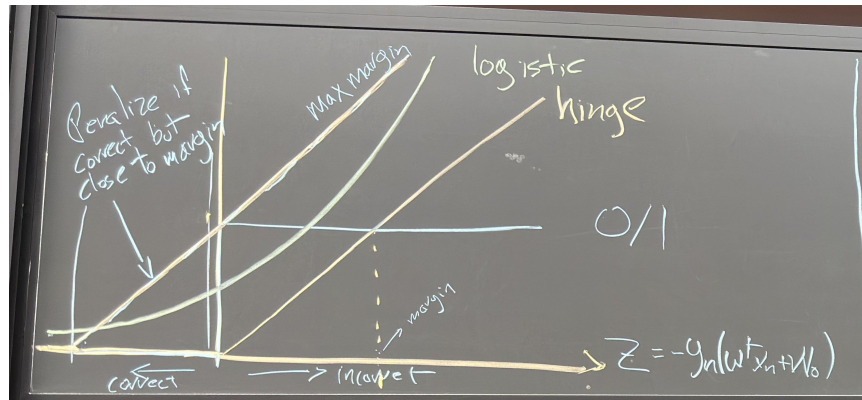
where C is a regularization parameter that controls the trade-off between maximizing the margin and minimizing the classification error. A larger C implies a smaller margin but fewer misclassifications, while a smaller C implies a larger margin with more misclassifications allowed.

By introducing the slack variables ξ , the soft margin SVM allows for a balance between maximizing the margin and minimizing the classification error, making it more robust to noisy data.



4 Review of Losses for Classification

We talked about the 0/1 loss, logistic loss, hinge loss (ReLU).



The max margin penalizes when we are correct but close to the margin. That is the difference between the max margin and the hinge loss.

5 Concept Check

1) Effect of C on Softmax objective

For each of these questions, answer with small c or big c .

1. Which would you choose if the goal is to maximize the margin?
2. Which would you choose if the goal is to minimize misclassifying the train?
3. Which would you choose if the goal is to overfit?
4. Which would you choose if the goal is to underfit?
5. Which would you choose if the goal is to have a less "flat" boundary?
6. Which would you choose if the goal is to have a more "flat" boundary?

6 Concept Check Solutions

1)

1. Which would you choose if the goal is to maximize the margin? Small c is less sensitive to outliers because small c is promoting generalization.
2. Which would you choose if the goal is to minimize misclassifying the train? Big c .
3. Which would you choose if the goal is to overfit? Big c , because we are emphasizing getting the training data right.
4. Which would you choose if the goal is to underfit? Small c .
5. Which would you choose if the goal is to have a less "flat" boundary? Small c , because there is more regularization.
6. Which would you choose if the goal is to have a more "flat" boundary? Big c .

