Recitation 4

Geometric Derivation of SVMs and Complementary Slackness

DS-GA 1003 Machine Learning

Spring 2021

February 24, 2021

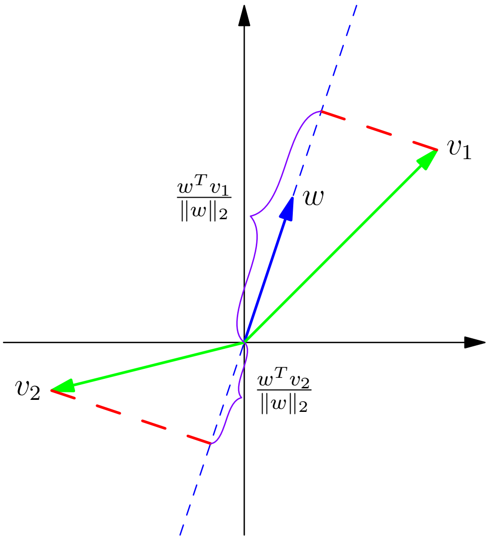
Intro Question

Question

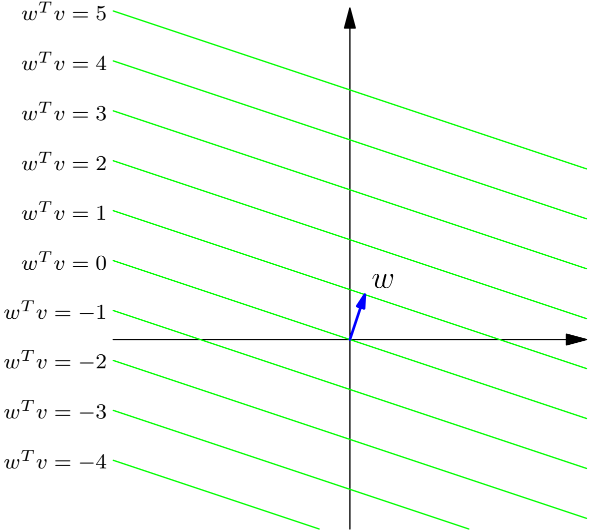
You have been given a data set (x_i, y_i) for $i = 1, \ldots, n$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Assume $w \in \mathbb{R}^d$ and $a \in \mathbb{R}$.

1. Suppose $y_i(w^T x_i + a) > 0$ for all i. Use a picture to explain what this means when $d = 2$.
2. Fix $M > 0$. Suppose $y_i(w^T x_i + a) \geq M$ for all i. Use a picture to explain what this means when $d = 2$.

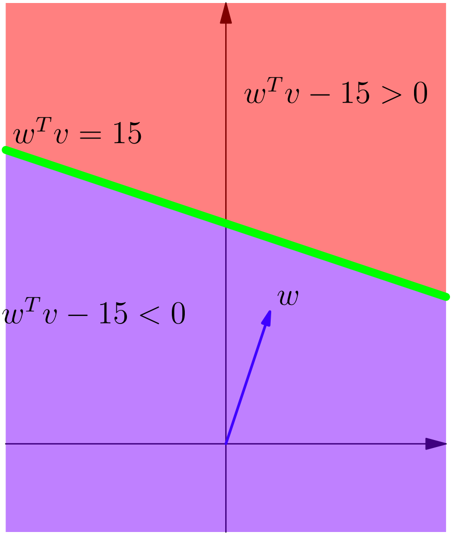
Component of v_1, v_2 in the direction w



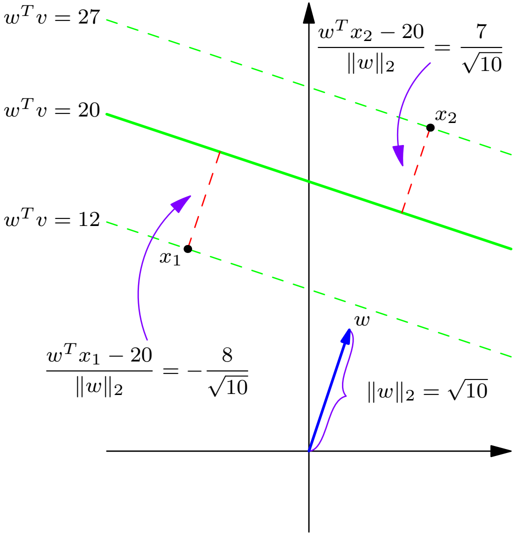
Level Surfaces of $f(v) = w^T v$ with $\|w\|_2 = 1$



Sides of the Hyperplane $w^T v = 15$



Signed Distance from x_1, x_2 to Hyperplane $w^T v = 20$

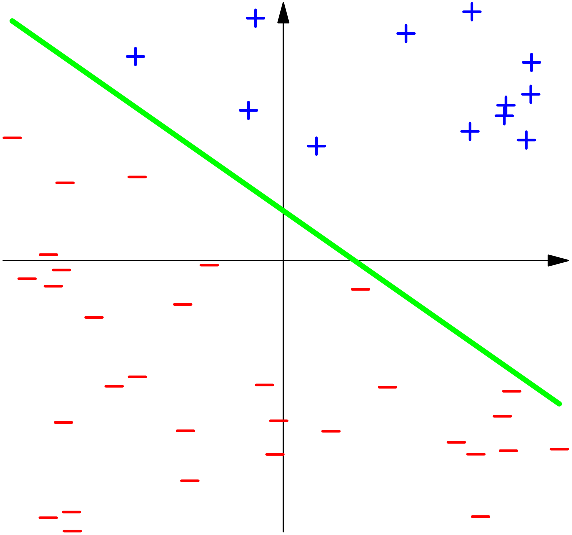


Linearly Separable

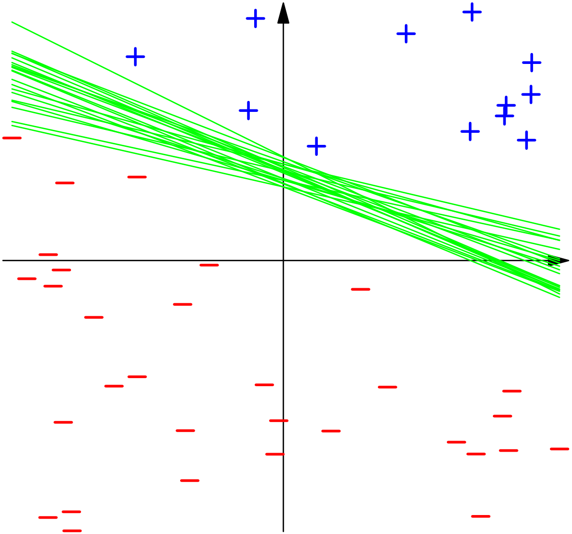
Definition

We say (x_i, y_i) for $i = 1, \ldots, n$ are *linearly separable* if there is a $w \in \mathbb{R}^d$ and $a \in \mathbb{R}$ such that $y_i(w^T x_i + a) > 0$ for all i. The set $\{v \in \mathbb{R}^d \mid w^T v + a = 0\}$ is called a *separating hyperplane*.

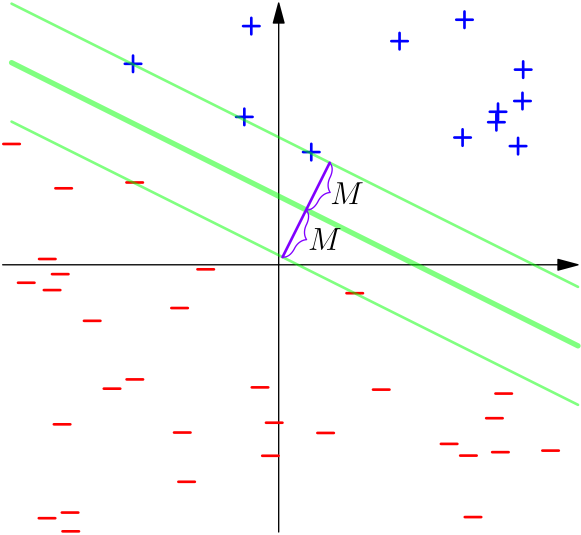
Linearly Separable Data



Many Separating Hyperplanes Exist



Maximum Margin Separating Hyperplane

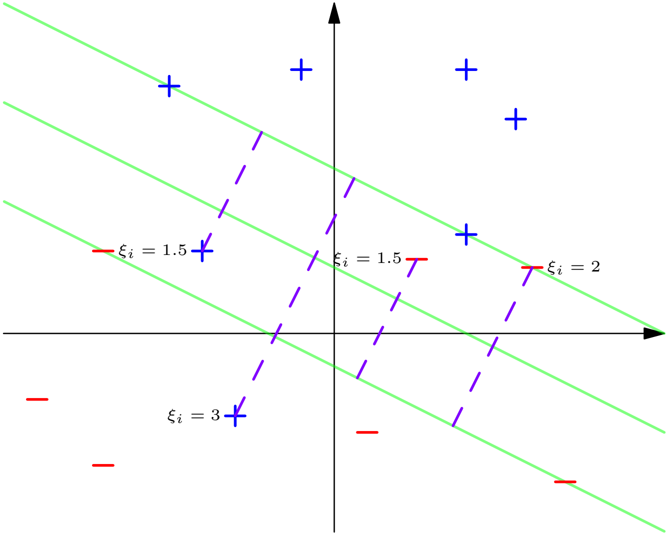


$$\frac{w^T v + a}{\|w\|_2} = M$$

$$\frac{w^T v + a}{\|w\|_2} = 0$$

$$\frac{w^T v + a}{\|w\|_2} = -M$$

Soft Margin SVM (unlabeled points have $\xi_i = 0$)



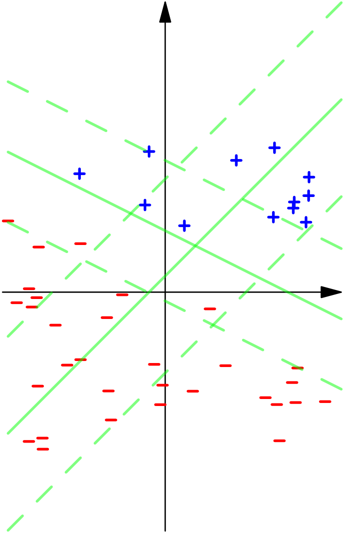
Questions

Questions

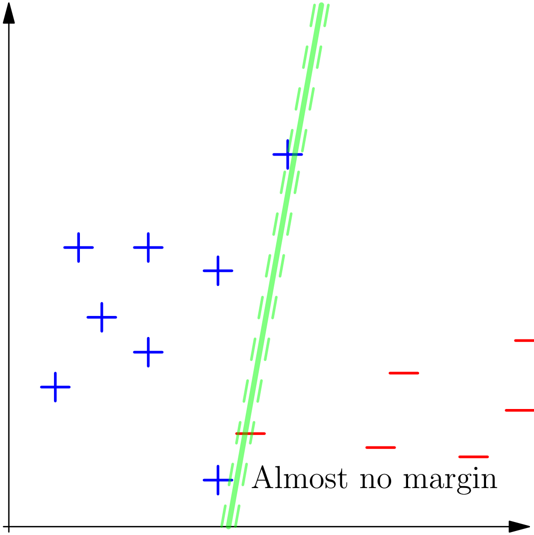
1. If your data is linearly separable, which SVM (hard margin or soft margin) would you use?
2. Explain geometrically what the following optimization problem computes:

$$\begin{array}{ll} \text{minimize}_{w,a,\xi} & \frac{1}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & y_i(w^T x_i + a) \geq 1 - \xi_i \quad \text{for all } i \\ & \|w\|_2^2 \leq r^2 \\ & \xi_i \geq 0 \quad \text{for all } i. \end{array}$$

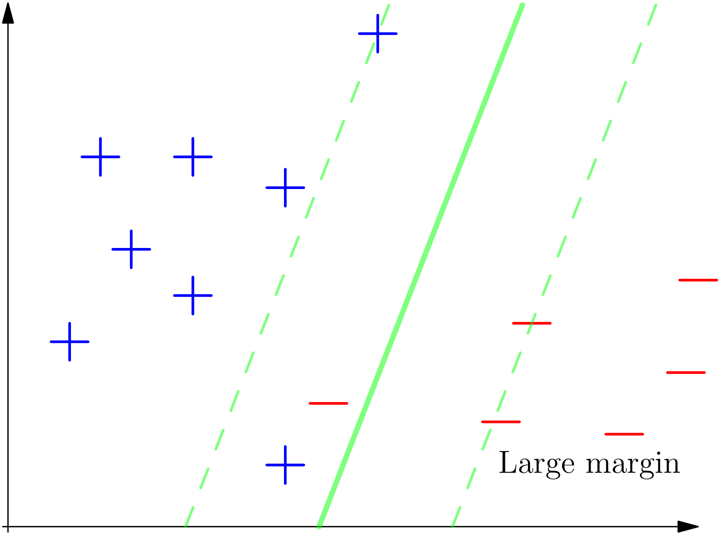
Optimize Over Cases Where Margin Is At Least $1/r$



Overfitting: Tight Margin With No Misclassifications



Training Error But Large Margin



SVM Review : Primal and Dual Formulations

The SVM Dual Problem

- We found the SVM dual problem can be written as::

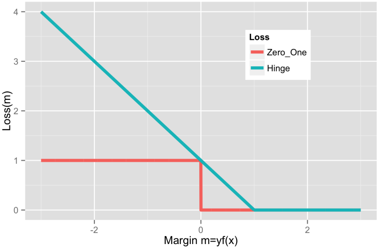
$$
\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \;\; i = 1, \ldots, n.
\end{aligned}
$$

- Given solution α^* to the dual problem, primal solution is $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$.
- Note $\alpha_i^* \in [0, \frac{c}{n}]$. So c controls max weight on each example (**Robustness!**).

Insights from Complementary Slackness: Margin and Support Vectors

The Margin and Some Terminology

- For notational convenience, define $f^*(x) = x^T w^* + b^*$.
- Margin $yf^*(x)$



- Incorrect classification: $yf^*(x) \leq 0$.
- Margin error: $yf^*(x) < 1$.
- "On the margin": $yf^*(x) = 1$.
- "Good side of the margin": $yf^*(x) > 1$.

Support Vectors and The Margin

- Recall "**slack variable**" $\xi^* = max(0, 1 - y_i f^*(x_i))$ is the hinge loss on (x_i, y_i).
- Suppose $\xi^* = 0$,
- Then $y_i(f^*(x_i)) \geq 1$
    - "on the margin" $(=1)$ or
    - "on the good side" (> 1)

Complementary Slackness Conditions

- Recall our primal constraints and Lagrange multipliers:

Lagrange Multiplier	Constraint
λ_i	$-\xi_i \leq 0$
α_i	$((1 - y_i f(x_i)) - \xi_i) \leq 0$

- Recall first order condition $\nabla_{\xi_i} L = 0$ gave us $\lambda_i^* = \frac{c}{n} - \alpha_i^*$
- By strong duality, we must have complementary slackness:

$$\alpha_i^*(1 - y_i f^*(x_i) - \xi_i^*) = 0$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

Consequences of Complementary Slackness

- By strong duality, we must have complementary slackness:

$$\alpha_i^*(1 - y_i f^*(x_i) - \xi_i^*) = 0$$

$$\left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0$$

- if $y_i f^*(x_i) > 1$, then the margin loss $\xi_i^* = 0$ and we get $\alpha_i^* = 0$
- if $y_i f^*(x_i) < 1$, then the margin loss $\xi_i^* > 0$, so $\alpha_i^* = \frac{c}{n}$
- if $\alpha_i^* = 0$, then $\xi_i^* = 0$, which implies no loss, so $y_i f^*(x_i) \geq 1$
- if $\alpha_i^* \in \left(0, \frac{c}{n}\right)$, then $\xi_i^* = 0$, which implies $1 - y_i f^*(x_i) = 0$

Complementary Slackness Results: Summary

$$
\begin{aligned}
\alpha_i^* = 0 \quad &\implies \quad y_i f^*(x_i) \geq 1 \\
\alpha_i^* \in \left(0, \frac{c}{n}\right) \quad &\implies \quad y_i f^*(x_i) = 1 \\
\alpha_i^* = \frac{c}{n} \quad &\implies \quad y_i f^*(x_i) \leq 1
\end{aligned}
$$

$$
\begin{aligned}
y_i f^*(x_i) < 1 \quad &\implies \quad \alpha_i^* = \frac{c}{n} \\
y_i f^*(x_i) = 1 \quad &\implies \quad \alpha_i^* \in \left[0, \frac{c}{n}\right] \\
y_i f^*(x_i) > 1 \quad &\implies \quad \alpha_i^* = 0
\end{aligned}
$$

Support Vectors

- if α_i^* is a solution to the dual problem, then primal solution is

$$w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$$

  with $\alpha_i^* \in \left[0, \frac{c}{n}\right]$
- The x_i's corresponding to $\alpha_i^* > 0$ are called **support vectors.**
- Few margin errors or "on the margin" examples $\Longrightarrow$ **sparsity in input examples.**

Complementary Slackness to get b^*

The Bias Term: **b**

- For our SVM primal, the complementary slackness conditions are:

$$\alpha_i^*(1 - y_i[x_i^T w^* + b] - \xi_i^*) = 0 \tag{1}$$

$$\lambda_i^* \xi_i^* = \left(\frac{c}{n} - \alpha_i^*\right) \xi_i^* = 0 \tag{2}$$

- Suppose there's an i such that $\alpha_i^* \in \left(0, \frac{c}{n}\right)$
- (2) implies $\xi_i^* = 0$
- (1) implies

$$y_i[x_i^T w^* + b^*] = 1$$

$$\iff x_i^T w^* + b^* = y_i (use\, y_i \in \{-1, 1\})$$

$$\iff b^* = y_i - x_i^T w^*$$

The Bias Term: **b**

- The optimal b is,

$$b^* = y_i - x_i^T w^*$$

- We get the same b^* for any choice of i with $\alpha_i^* \in \left(0, \frac{c}{n}\right)$
    - **With exact calculations**
- With numerical error, more robust to average over all eligible i s:

$$b^* = mean\left\{y_i - x_i^T w^* | \alpha_i^* \in \left(0, \frac{c}{n}\right)\right\}$$

- If there are no $\alpha_i^* \in \left(0, \frac{c}{n}\right)$?
    - Then we have a **degenerate SVM training problem**[1] - $(w^* = 0)$

[1] See Rifkin et al.'s A Note on Support Vector Machine Degeneracy, an MIT AI Lab Technical Report

Teaser for Kernelization

Dual Problem: Dependence on x through inner products

- SVM Dual Problem:

$$\begin{aligned}
\sup_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_j^T x_i \\
\text{s.t.} \quad & \sum_{i=1}^{n} \alpha_i y_i = 0 \\
& \alpha_i \in \left[0, \frac{c}{n}\right] \ i = 1, \ldots, n.
\end{aligned}$$

- Note that all dependence on inputs x_i and x_j is through their inner product: $\langle x_j, x_i \rangle = x_j^T x_i$.
- We can replace $x_j^T x_i$ by any other inner product...
- This is a "kernelized" objective function.

References

- DS-GA 1003 Machine Learning Spring 2019