Recitation 4

Recap of SVMs and Complementary Slackness

Vishakh

CDS

February 16, 2022

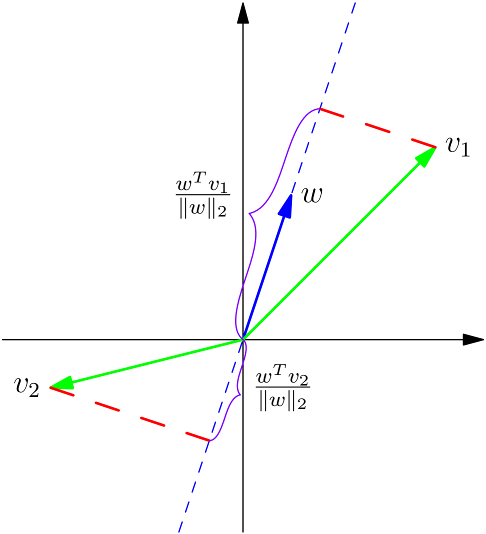
Announcement

- HW 2 is due tonight + HW 3 will be out
- Grading of HW 1 is done and scores will be out tonight
- Selected solutions (Brightspace) + Regrade requests (Gradescope)

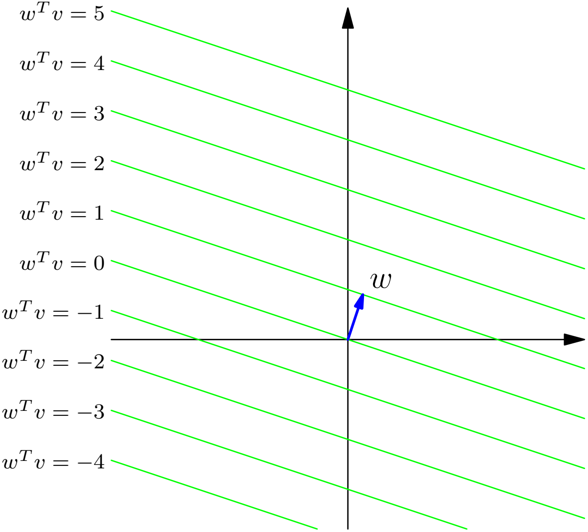
Agenda

- Recap: Hyperplanes to SVMs
- Hard-margin vs Soft-margin SVMs
- Preview to Complementary Slackness + Kernelization

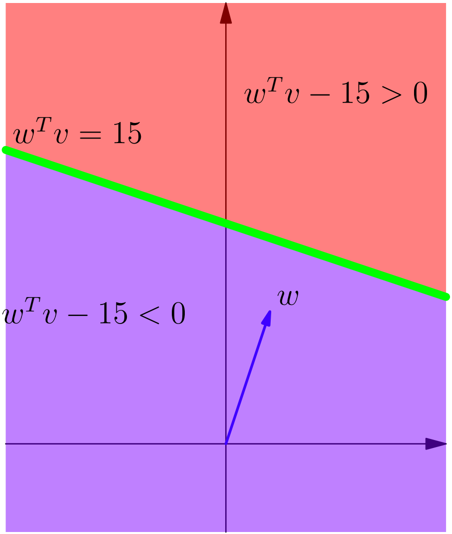
Component of v_1, v_2 in the direction w



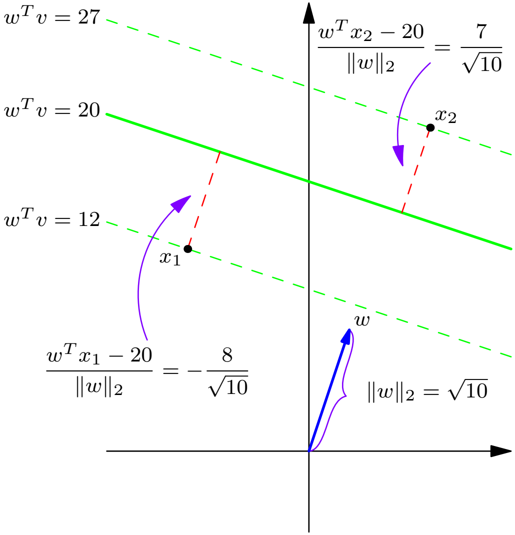
Level Surfaces of $f(v) = w^T v$ with $\|w\|_2 = 1$



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



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

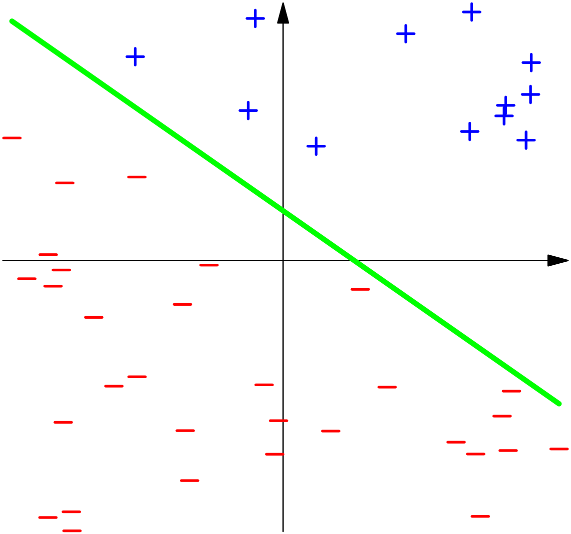


Linearly Separable

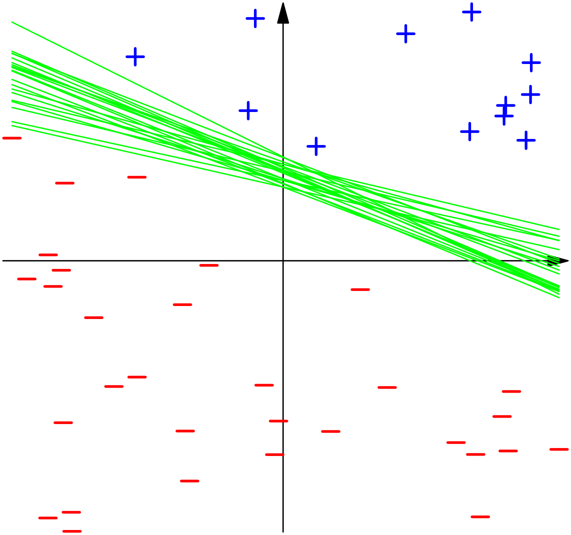
Definition

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, $y = \pm 1$. The set $\{v \in \mathbb{R}^d \mid w^T v + a = 0\}$ is called a *separating hyperplane*.

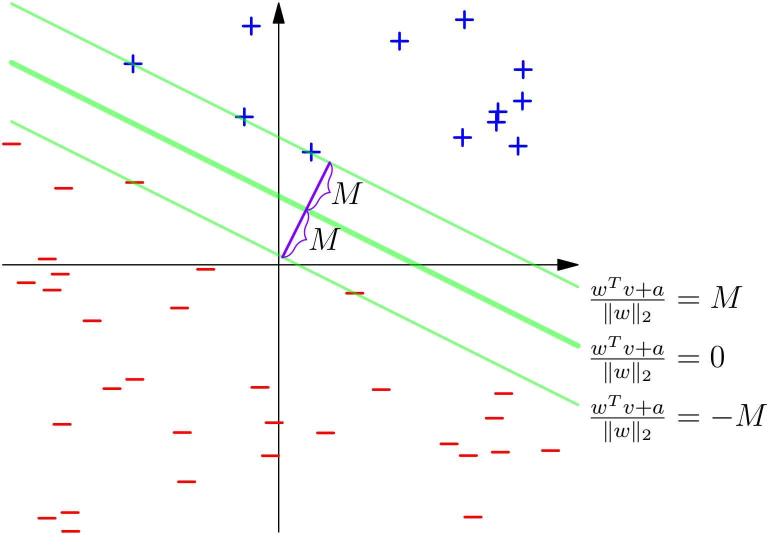
Linearly Separable Data



Many Separating Hyperplanes Exist



Maximum Margin Separating Hyperplane



Maximizing the Margin

We can rewrite this in a more standard form:

$$\begin{aligned} \text{maximize}_{w,a,M} \quad & M \\ \text{subject to} \quad & \frac{y_i(w^T x_i + a)}{\|w\|_2} \geq M \quad \text{for all } i. \end{aligned}$$

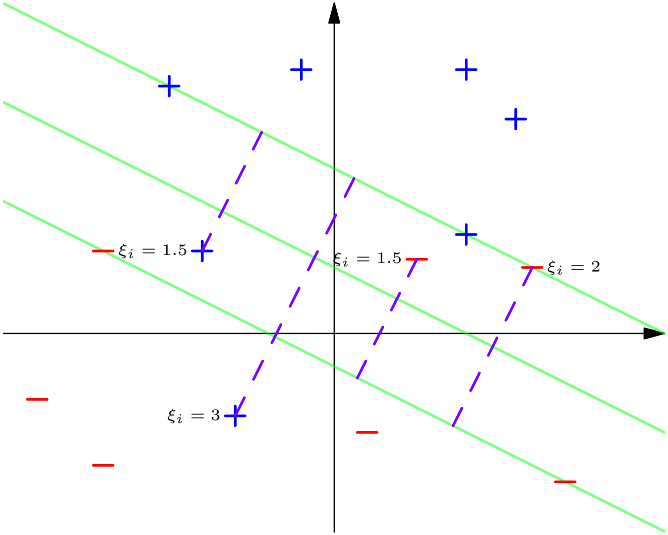
Let's fix the norm $\|w\|_2$ to $1/M$ to obtain:

$$\begin{aligned} \text{maximize} \quad & \frac{1}{\|w\|_2} \\ \text{subject to} \quad & y_i(w^T x_i + b) \geq 1 \quad \text{for all } i \end{aligned}$$

It's equivalent to solving the minimization problem

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{subject to} \quad & y_i(w^T x_i + b) \geq 1 \quad \text{for all } i \end{aligned}$$

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



Soft Margin SVM

Introduce **slack variables**:

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

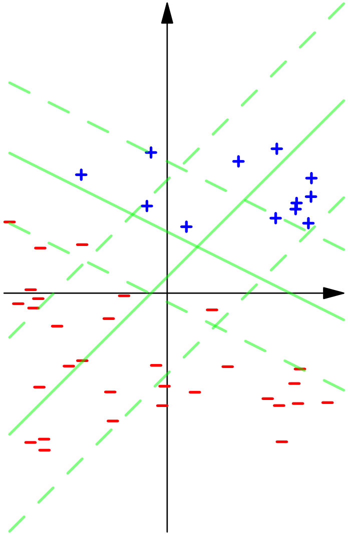
- If $\xi_i = 0\ \forall i$, it's reduced to hard SVM.
- If $\xi_i > 0$, we have misclassified an example i.e. it is on the wrong side of the hyperplane
- C controls the penalty for each misclassfication.

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

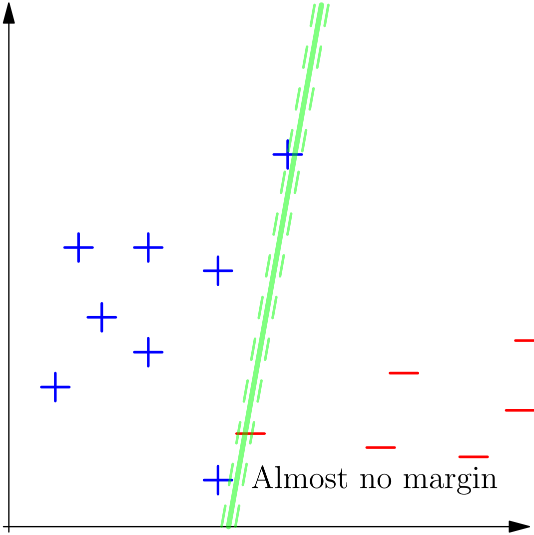
1. If your data is linearly separable, which SVM (hard margin or soft margin) would you use?
2. Consider the optimization problem:

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

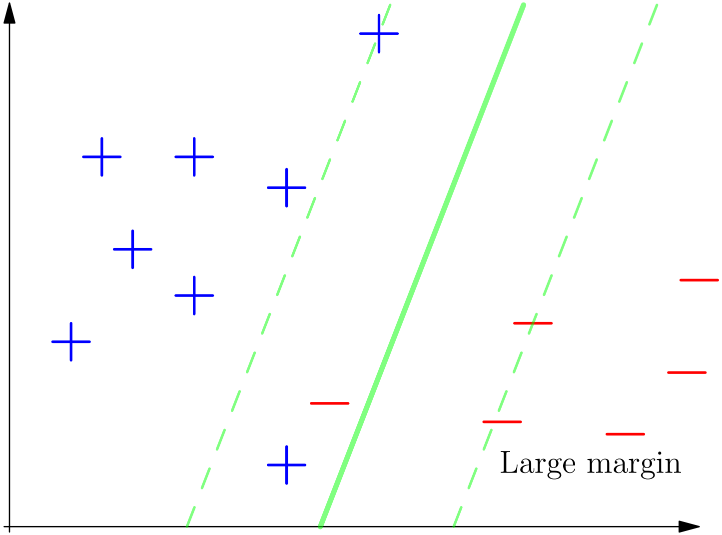
Optimize Over Cases Where Margin Is At Least $1/r$



Overfitting: Tight Margin With No Misclassifications



Training Error But Large Margin



SVM Lagrange Multipliers

Primal

$$
\begin{array}{ll}
\text{minimize} & \frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\
\text{subject to} & -\xi_i \le 0 \quad \text{for } i = 1,\ldots,n \\
& \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \le 0 \quad \text{for } i = 1,\ldots,n
\end{array}
$$

SVM Lagrange Multipliers

Primal

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{subject to} & -\xi_i \le 0 \quad \text{for } i = 1, \ldots, n \\ & \left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \le 0 \quad \text{for } i = 1, \ldots, n \end{array}$$

Subgradient Descent (HW 3)

SVM Lagrange Multipliers

Dual

$$L(w, b, \xi, \alpha, \lambda) = \frac{1}{2}||w||^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i\left(1 - y_i\left[w^T x_i + b\right] - \xi_i\right) + \sum_{i=1}^{n}\lambda_i\left(-\xi_i\right)$$

Lagrange Multiplier	Constraint
λ_i	-$\xi_i \leq 0$
α_i	$\left(1 - y_i\left[w^T x_i + b\right]\right) - \xi_i \leq 0$

The SVM Dual Problem

- By Slater's conditions, we have strong duality (Convex Optimization + Affine Constraints + Feasibility)
- We can draw some insights from complementary slackness.
  - If x^* is primal optimal and λ^* is dual optimal, $f_0(x^*) = g(\lambda^*)$
  - $f_0(x^*) = g(\lambda^*) = f_0(x^*) + \sum_{i=1}^{m} \lambda_i^* f_i(x^*)$
  - Each term in sum $\sum_{i=1}^{m} \lambda_i^* f_i(x^*)$ must actually be 0.
  - That is $\lambda_i > 0 \implies f_i(x^*) = 0$ and $f_i(x^*) < 0 \implies \lambda_i = 0 \quad \forall i$

The SVM Dual Problem

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

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

(First order conditions on the Lagrangian)

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 $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$.
  - α_i^*, y_i is scalar, so the optimum solution is in the span of the input examples

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 $w^* = \sum_{i=1}^{n} \alpha_i^* y_i x_i$.
- We also know that $\alpha_i^* \in [0, \frac{c}{n}]$, which is the 'weight' associated with each example. So C controls max weight on each example.

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	$((1 - y_i f(x_i)) - \xi_i) \leq 0$

- By strong duality, we must have complementary slackness. Each of $\sum_{i=1}^{m} \lambda_i^* f_i(x^*)$ must be 0:

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

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

  - Recall first order condition $\nabla_{\xi_i} L = 0$ gave us $\lambda_i^* = \frac{c}{n} - \alpha_i^*$

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 you're on the right side of the margin i.e slack $\xi_i^* = 0$ and we get $\alpha_i^* = 0$
- if $y_i f^*(x_i) < 1$, then a misclassification has occurred and slack $\xi_i^* > 0$, so $\alpha_i^* = \frac{c}{n}$

Consequences of Complementary Slackness

- By strong duality, we must have complementary slackness:

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

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

- We also know that $\alpha_i^* \in [0, \frac{c}{n}]$
- 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 $1 - y_i f^*(x_i) = 0$

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 $\alpha_i^* \in \left[0, \frac{c}{n}\right]$ as the 'weight' associated with that example

- In the case where $\alpha_i^* = 0$, there is no dependence on those example x_i
- 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 Results: Summary

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

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

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.