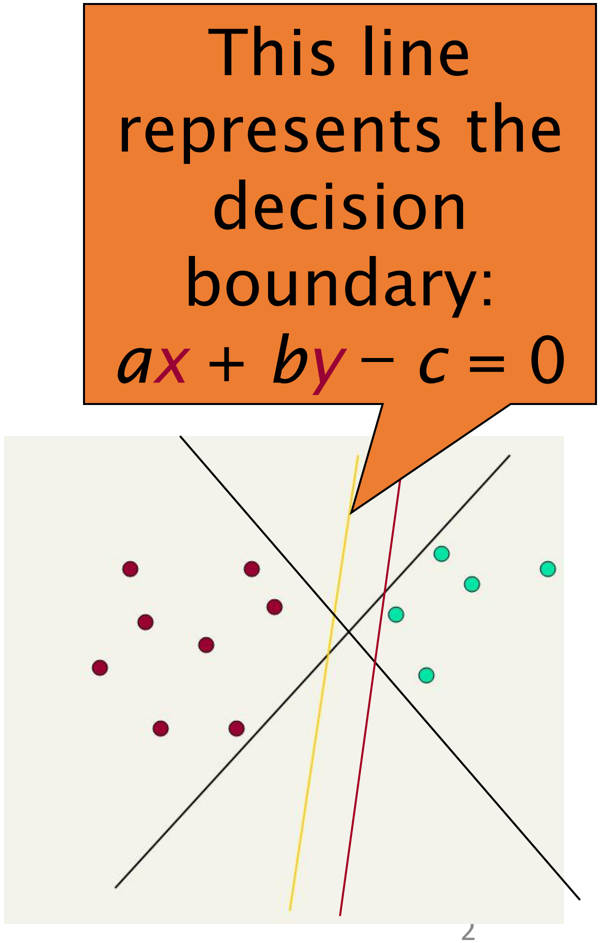
Support Vector Machines

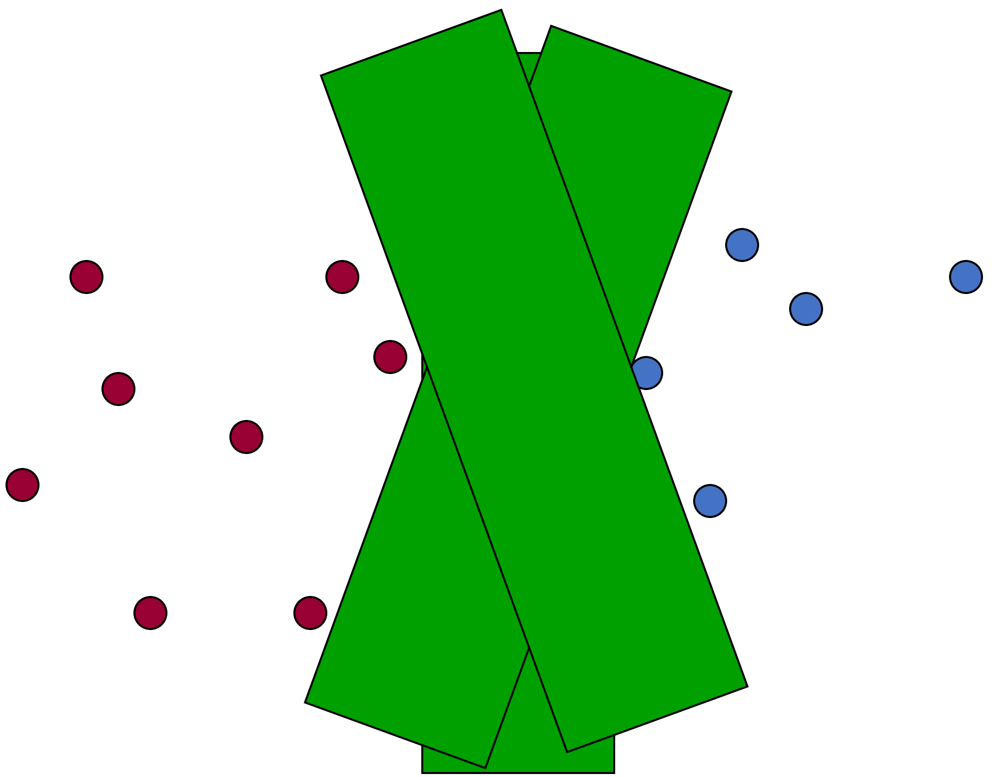
Linear classifiers: Which Hyperplane?

- Lots of possible solutions for *a, b, c.*
- Some methods find a separating hyperplane, but not the optimal one [according to some criterion of expected goodness]
  - E.g., perceptron
- Support Vector Machine (SVM) finds an optimal* solution.
  - Maximizes the distance between the hyperplane and the “difficult points” close to decision boundary
  - One intuition: if there are no points near the decision surface, then there are no very uncertain classification decisions



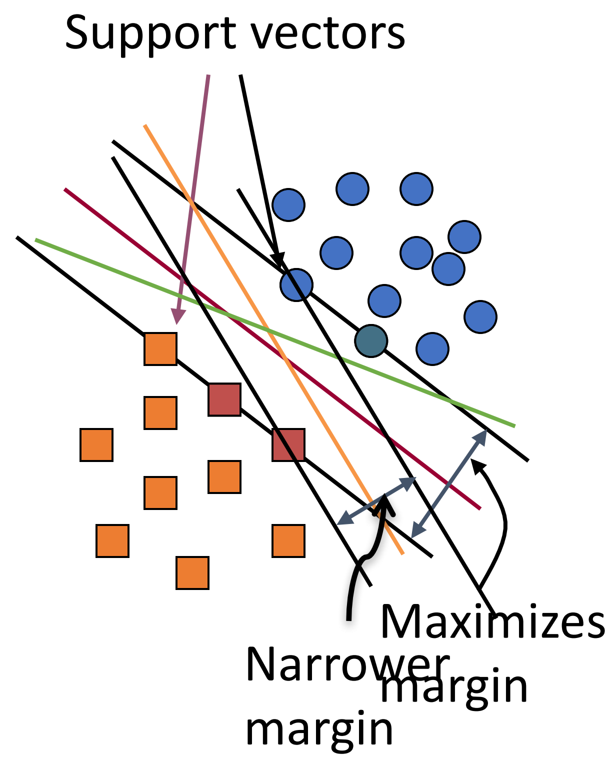
Another intuition

- If you have to place a fat separator between classes, you have fewer choices



Support Vector Machine (SVM)

- SVMs maximize the *margin* around the separating hyperplane.
    - A.k.a. large margin classifiers
- The decision function is fully specified by a subset of training samples, *the support vectors*.
- Solving SVMs is a *quadratic programming* problem
- Seen by many as the most successful current text classification method*



*but other discriminative methods often perform very similarly

Lagrange Multipliers

Constrained Optimization

Want to minimize some function $f(x)$, but there are some *constraints* on the values of x.

Method 1 (Dual Problem)

Solve a *dual optimization problem* where the constraints are nicer, and where it is easier to implement gradient descent.

Method 2 (Exact Solution)

Solve the *Lagrangian* system of equations.

Equality Constraints

Problem.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \ldots, h_l(x) = 0 \end{array}$$

Lagrangian.

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Example.

$$\begin{array}{ll} \text{minimize} & f(x) = n_1 \log x_1 + \cdots + n_d \log x_d \\ \text{subject to} & h(x) = x_1 + \cdots + x_d - 1 = 0 \end{array}$$

$$L(x, \lambda) = n_1 \log x_1 + \cdots + n_d \log x_d + \lambda(x_1 + \cdots + x_d - 1)$$

Two-Player Game

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

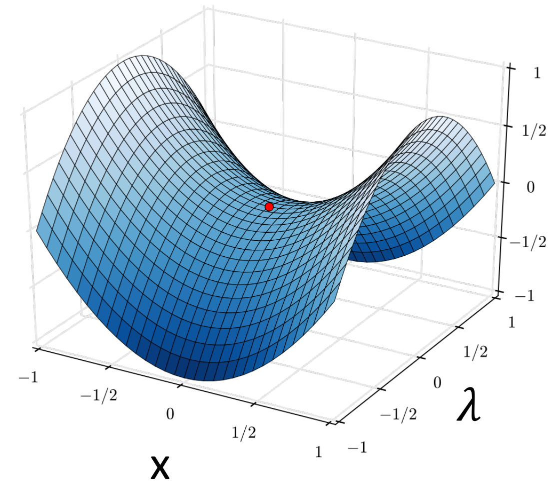
Rules.

- You get to choose the value of x.
  Your goal is to minimize $L(x, \lambda)$.
- Your adversary gets to choose the value of λ.
  His goal is to maximize $L(x, \lambda)$.

Primal Game

$$L(x,\lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Primal Game. You go first.



Your Strategy.

- Ensure that $h_1(x) = 0, \ldots, h_l(x) = 0$.
- Find x that minimizes $f(x)$.

Final Score. $p^* = \min_x \max_\lambda L(x,\lambda)$

The optimal x^*, λ^* are *saddle points* of $L(x,\lambda)$.

Dual Game

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

Dual Game. You go second.

Adversary's Strategy.

- For each λ, compute $\ell(\lambda) = \min_x L(x, \lambda)$
- Find λ that maximizes $\ell(\lambda)$.

Final Score. $d^* = \max_\lambda \min_x L(x, \lambda)$

Max-Min Inequality

Primal. $p^* = \min_x \max_\lambda L(x, \lambda)$

Dual. $d^* = \max_\lambda \min_x L(x, \lambda)$

"you do better if you have the last say"

$$\boxed{\begin{aligned} p^* &= \min_x \max_\lambda L(x, \lambda) \\ &\geq \max_\lambda \min_x L(x, \lambda) = d^* \end{aligned}}$$

If $p^* = d^*$, we can solve the primal by solving the dual.

MAX-MIN INEQUALITY

Example.

	$x = 1$	$x = 2$
$\lambda = 1$	1	4
$\lambda = 2$	3	2

Primal. $p^* = \min_x \max_\lambda L(x, \lambda) = 3$

Dual. $d^* = \max_\lambda \min_x L(x, \lambda) = 2$

EXACT SOLUTION

Problem.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & h_1(x) = 0, \dots, h_l(x) = 0 \end{array}$$

Lagrange multipliers.

1. Write down the Lagrangian.

$$L(x, \lambda) = f(x) + \lambda_1 h_1(x) + \cdots + \lambda_l h_l(x)$$

2. Solve for critical points x, λ.

$$\nabla_x L(x, \lambda) = 0, \quad h_1(x) = 0, \dots, h_l(x) = 0$$

3. Pick critical point which gives global minimum.

Example

$$\begin{array}{ll} \text{minimize} & f(x) = n_1 \log x_1 + \cdots + n_d \log x_d \\ \text{subject to} & h(x) = x_1 + \cdots + x_d - 1 = 0 \end{array}$$

Lagrangian

$$L(x, \lambda) = n_1 \log x_1 + \cdots + n_d \log x_d + \lambda(x_1 + \cdots + x_d - 1)$$

Critical points

$$\begin{aligned} 0 &= x_1 + \cdots + x_d - 1 \\ 0 &= n_i / x_i + \lambda \end{aligned} \quad \Rightarrow \quad \begin{aligned} (-\lambda) &= n_1 + \cdots + n_d \\ x_i &= n_i / (-\lambda) \end{aligned}$$

Inequality Constraints (Primal-Dual)

Primal Problem.

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_1(x) \leq 0, \dots, g_m(x) \leq 0 \end{array}$$

Lagrangian.

$$L(x, \alpha) = f(x) + \alpha_1 g_1(x) + \cdots + \alpha_m g_m(x)$$

Dual Problem.

$$\begin{array}{ll} \text{maximize} & \ell(\alpha) \\ \text{subject to} & \alpha_1 \geq 0, \dots, \alpha_m \geq 0 \end{array} \qquad \text{where } \ell(\alpha) = \min_{x \in \mathbb{R}^d} L(x, \alpha)$$

Box constraints are easier to work with!

Inequality Constraints (Exact Soln)

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_1(x) \le 0, \dots, g_m(x) \le 0 \end{array}$$

Lagrangian.

$$L(x, \alpha) = f(x) + \alpha_1 g_1(x) + \cdots + \alpha_m g_m(x)$$

Solve for x, α satisfying

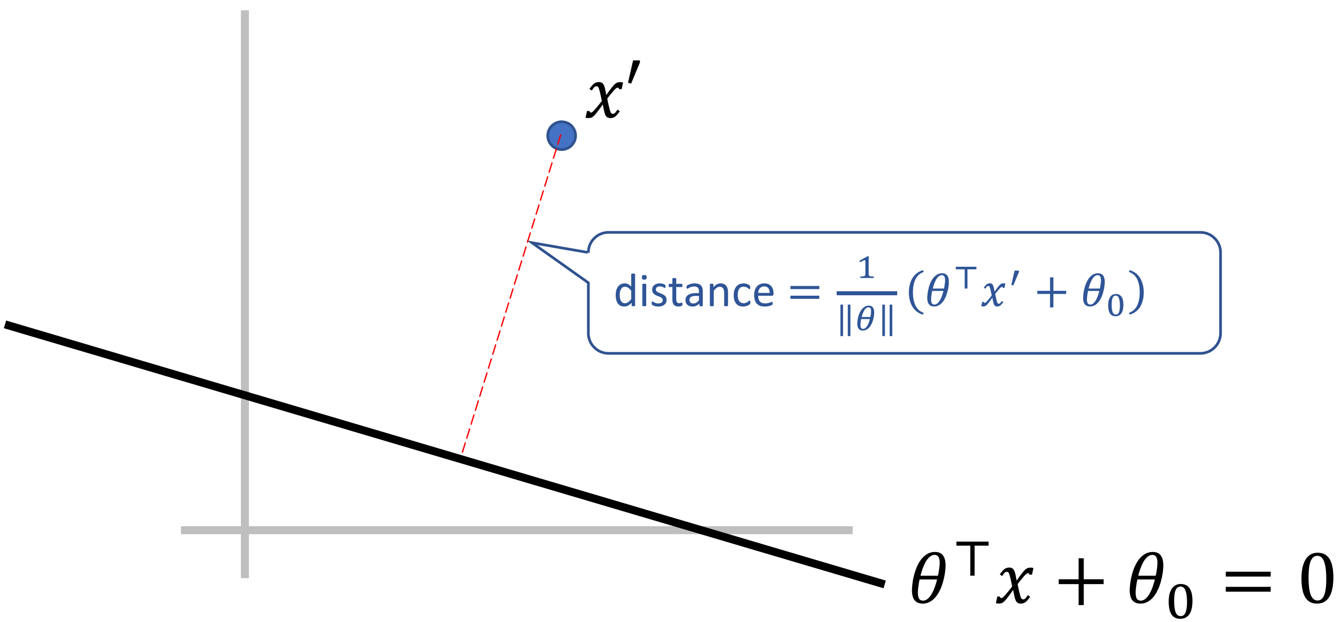
1. $\nabla_x L(x, \alpha) = 0$
2. $g_1(x) \le 0, \dots, g_m(x) \le 0$
3. $\alpha_1 \ge 0, \dots, \alpha_m \ge 0$
4. $\alpha_1 g_1(x) = 0, \dots, \alpha_m g_m(x) = 0$

Karush-Kuhn-Tucker (KKT) Conditions

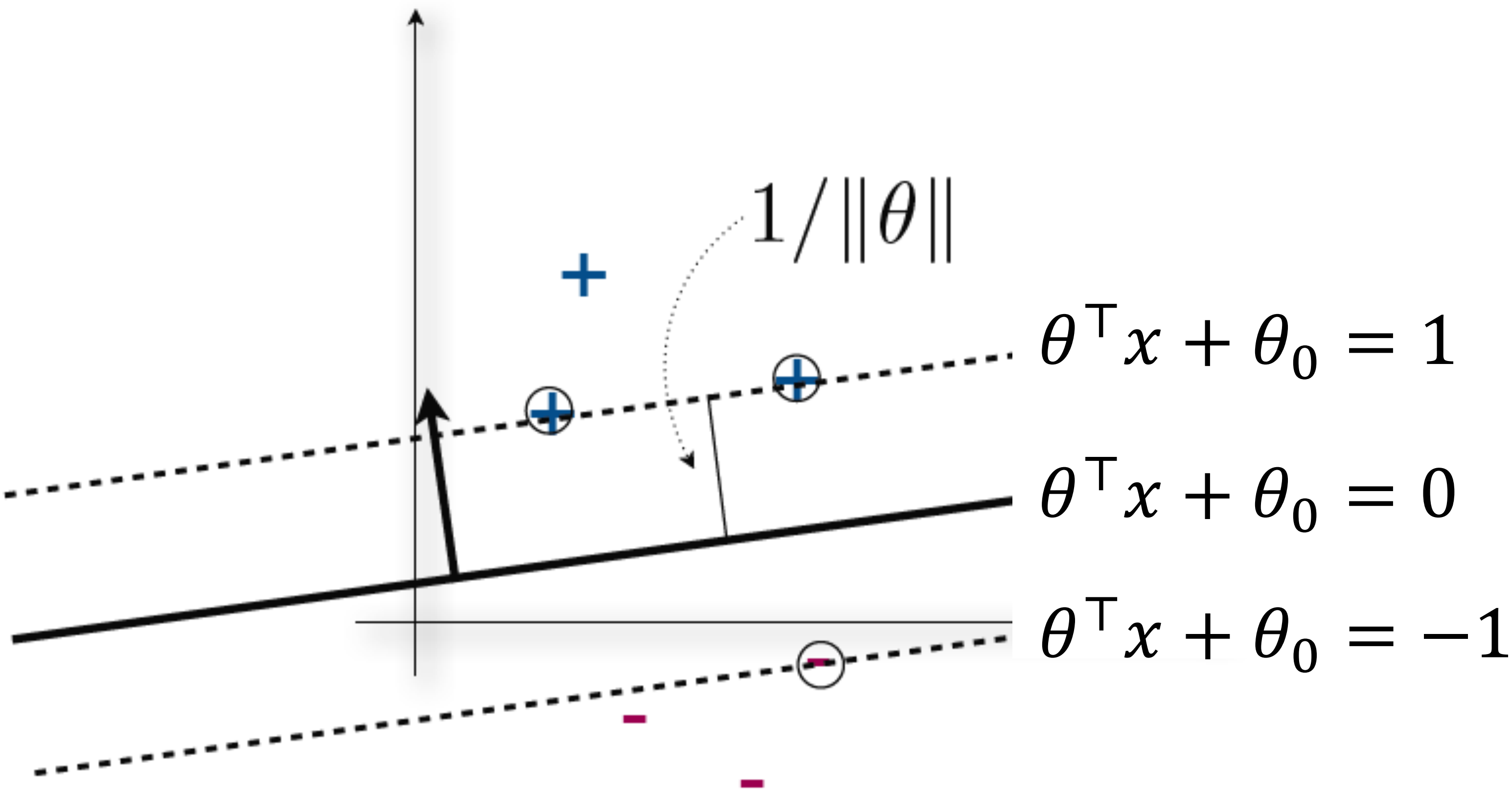
Complementary Slackness

Maximum Margins

Computing the Margin



Computing the Margin



Maximum Margin

Unfortunately, this only applies to data that is linearly separable.

Our goal is to

maximize $1/\|\theta\|$

subject to $y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

Or equivalently,

minimize $\frac{1}{2}\|\theta\|^2$

subject to $y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

Lagrangian

Drop θ_0 for now

minimize $\quad \frac{1}{2}\|\theta\|^2$

subject to $\quad y(\theta^\top x) \geq 1$ for all data (x, y)

Lagrangian. $\quad L(\theta, \alpha) = \frac{1}{2}\|\theta\|^2 + \sum_{(x,y)} \alpha_{x,y} \left(1 - y(\theta^\top x)\right)$

To find $\ell(\alpha) = \min_\theta L(\theta, \alpha)$, we solve

$$0 = \nabla_\theta L(\theta, \alpha) = \theta - \sum_{(x,y)} \alpha_{x,y}\, yx$$

to get $\theta = \sum_{(x,y)} \alpha_{x,y}\, yx$. Substituting into $L(\theta, \alpha)$ gives

$$\ell(\alpha) = \sum_{(x,y)} \alpha_{x,y} - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, yy'(x^\top x').$$

Primal-Dual

It can be shown that the primal and dual problems are equivalent (*strong duality*).

Primal.

minimize $\frac{1}{2}\|\theta\|^2$

subject to $y(\theta^\top x) \geq 1$ for all data (x, y)

Dual.

maximize $\sum_{(x,y)} \alpha_{x,y} - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, y y' (x^\top x')$

subject to $\alpha_{x,y} \geq 0$ for all (x, y)

After solving the dual to get the optimal $\alpha_{x,y}$'s,

we obtain the optimal θ using $\theta = \sum_{(x,y)} \alpha_{x,y}\, y x$.

Support Vectors

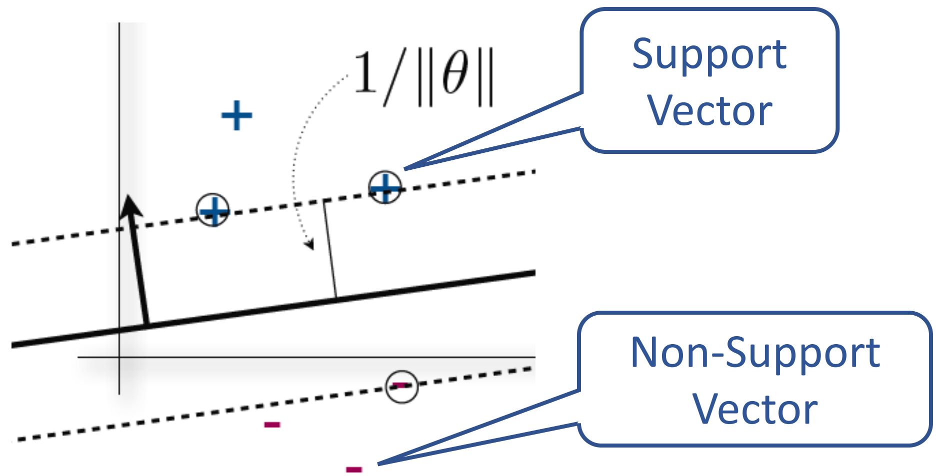
Complementary Slackness.

$\hat{\alpha}_{x,y} > 0$: $\quad y(\hat{\theta}^\top x) = 1 \quad$ Support Vectors

$\hat{\alpha}_{x,y} = 0$: $\quad y(\hat{\theta}^\top x) > 1 \quad$ Non-Support Vectors

Sparsity

Since very few data points are support vectors, most of the $\hat{\alpha}_{x,y}$ will be zero.



Kernel Trick

Learning.

$$\ell(\alpha) = \sum_{(x,y)} \alpha_{x,y} - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'}\, y y' (x^\top x')$$

Prediction.

$$h(x;\theta) = \operatorname{sign}(\theta^\top x) = \operatorname{sign}\left(\sum_{(x',y')} \alpha_{x',y'}\, y' (x^\top x')\right)$$

For the dual, we don't need the feature vectors x, x'.
Knowing just the dot products $(x^\top x')$ is enough.

Recall that $(x^\top x')$ is a measure of similarity between x and x'.
This similarity function is also called a *kernel*.

Extensions

SVM with OFFSET

Primal.

minimize $\frac{1}{2}\|\theta\|^2$

subject to $y(\theta^\top x + \theta_0) \geq 1$ for all data (x, y)

Dual.

maximize $\sum_{(x,y)} \alpha_{x,y} - \frac{1}{2}\sum_{(x,y)}\sum_{(x',y')} \alpha_{x,y}\, \alpha_{x',y'} y y' (x^\top x')$

subject to $\alpha_{x,y} \geq 0$ for all (x, y)

$\sum_{(x,y)} \alpha_{x,y}\, y = 0$

Parameters. $\hat{\theta} = \sum_{(x,y)} \alpha_{x,y}\, y x$

$\hat{\theta}_0 = y - \hat{\theta}^\top x$ where (x, y) is a support vector

SVM with Errors

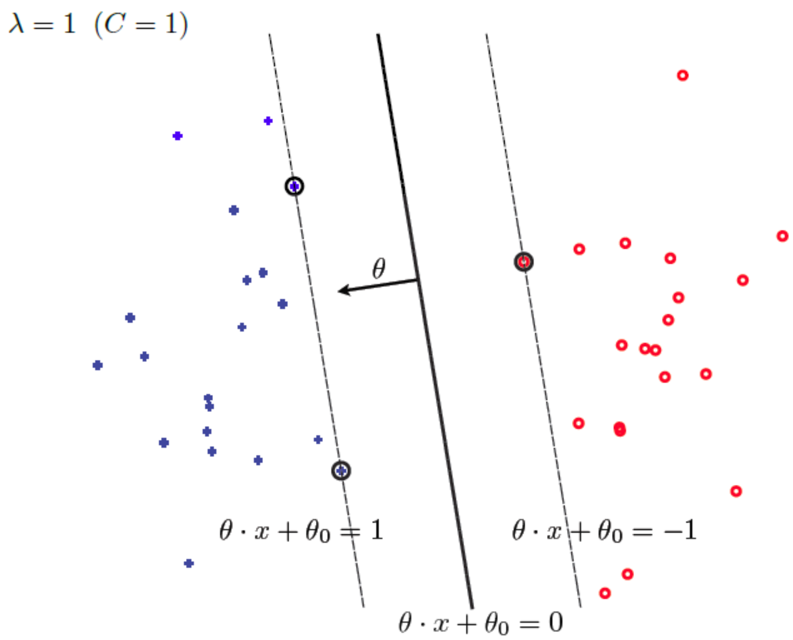
Primal.

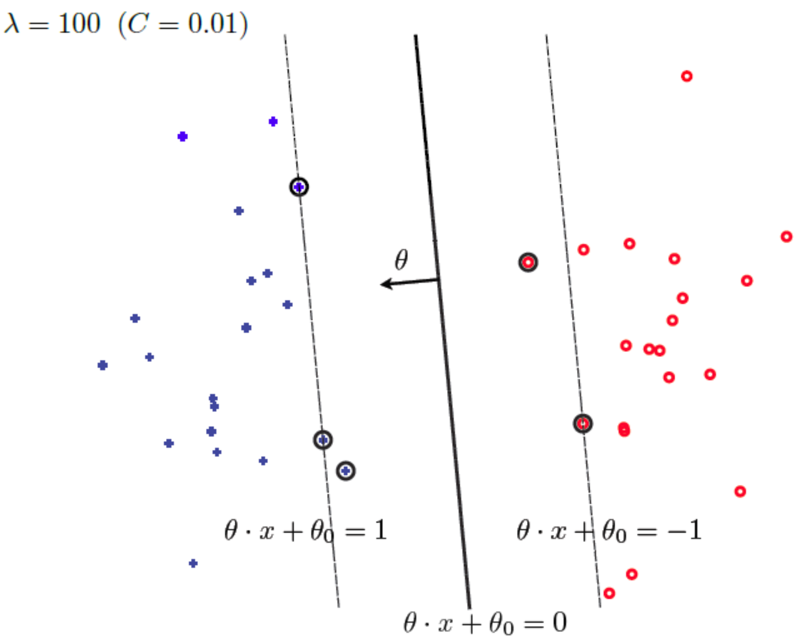
$$
\begin{array}{lll}
\text{minimize} & \frac{\lambda}{2}\|\theta\|^2 + \frac{1}{n}\sum_{(x,y)} \xi_{x,y} & \\
\text{subject to} & y(\theta^\top x + \theta_0) \geq 1 - \xi_{x,y} & \text{for all data } (x,y) \\
 & \xi_{x,y} \geq 0 & \text{for all data } (x,y)
\end{array}
$$

Slack variables allow constraints to be violated for a cost.

SVM with Errors

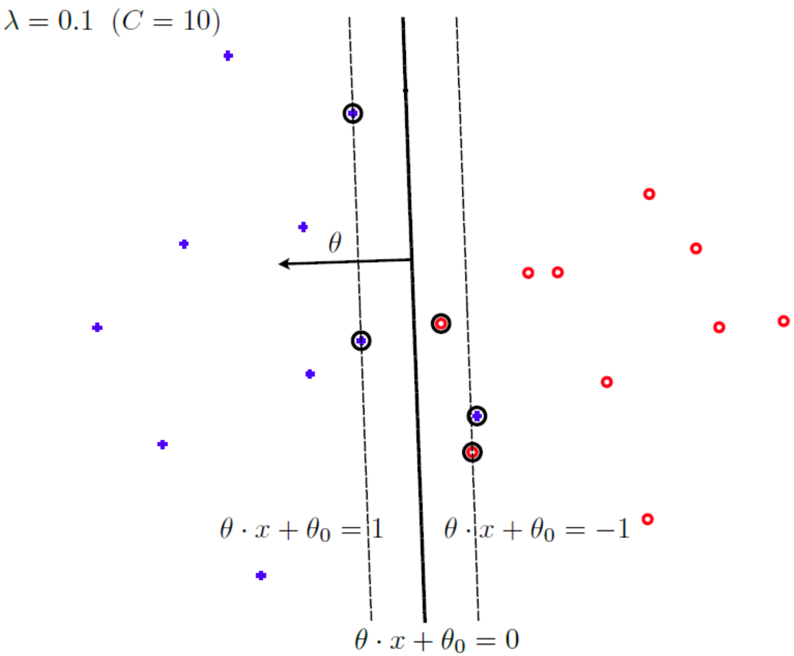
Linearly Separable.

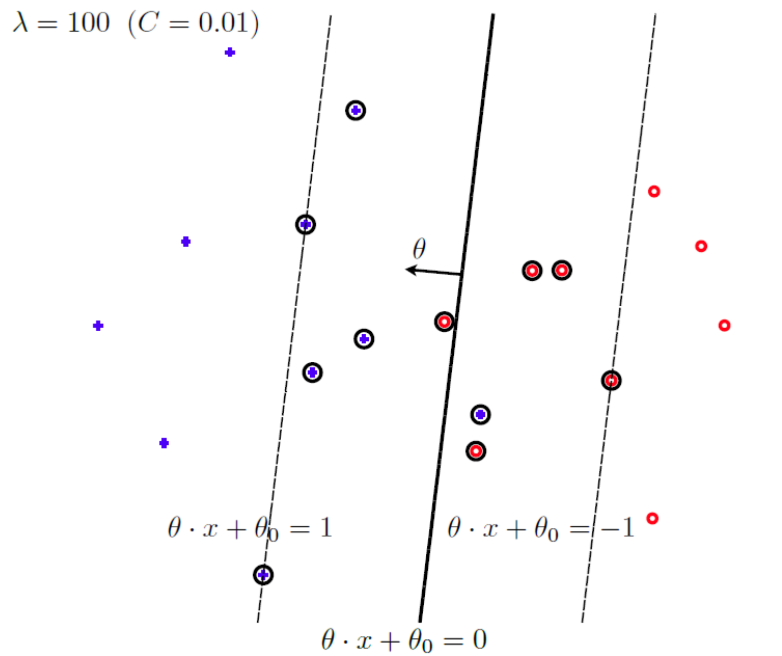




SVM with Errors

Not Linearly Separable.





SVM with Errors

Dual.

maximize $\quad \sum_{(x,y)} \alpha_{x,y} \; - \frac{1}{2} \sum_{(x,y)} \sum_{(x',y')} \alpha_{x,y} \, \alpha_{x',y'} y y' (x^\top x')$

subject to $\quad 1/\lambda \geq \alpha_{x,y} \geq 0$ for all (x, y)

$\sum_{(x,y)} \alpha_{x,y} \, y = 0$

Putting limits on what the adversary can do.

There are many efficient solvers for quadratic problems with box constraints.

Summary

- Lagrange Multipliers
  - Lagrangian
  - Primal-Dual Problems
  - Inequality Constraints
  - Complementary Slackness
- Support Vector Machines
  - Maximum Margins
  - Dual Problem
  - Support Vectors
  - Kernel Trick
- Regularization
  - Slack Variables
  - Regularized Hinge Loss
  - Bounded Multipliers

Intended Learning Outcomes

Support Vector Machines

- Write down the primal problem, and explain how it is derived from the maximum margin problem.
- Write down the dual problem, and identify the kernel. Describe how the optimal θ is derived from the $\alpha_{x,y}$'s. Describe in terms of the $\alpha_{x,y}$'s, how to do prediction.
- Define support vectors, both geometrically and in terms of the $\alpha_{x,y}$'s. Recognize that most of the $\alpha_{x,y}$'s are zero.

Intended Learning Outcomes

Extensions

- Describe the dual problem for the SVM with offset.
- Describe the primal problem for SVM with slack variables. Show that the primal is equivalent to regularized hinge loss. Explain how the regularizing parameter λ affects the margins. Describe the dual problem in terms of box constraints.