Support Vector Machines

Instructor: Jessica Wu -- Harvey Mudd College

The instructor gratefully acknowledges Andrew Ng (Stanford), Eric Eaton (UPenn), David Sontag (NYU), Piyush Rai (Utah) and the many others who made their course materials freely available online.

SVM Basics

Learning Goals

- Describe the goal of SVMs
  - Start from perceptron and build graphical intuition
- Setup formal SVM optimization problem

Recall the Perceptron

Assume there is a linear classifier

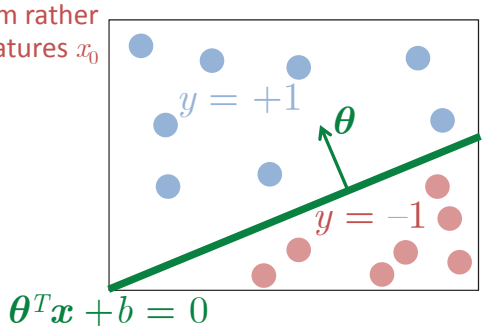
Start with simple learner and analyze with it does

Let $y \in \{-1, +1\}$ use explicit bias term rather than folding into features x_0

$$h_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(\boldsymbol{\theta}^T \boldsymbol{x} + b)$$

where

$$g(z) = \text{sgn}(z) \begin{cases} +1, & z \geq 0 \\ -1, & z < 0 \end{cases}$$

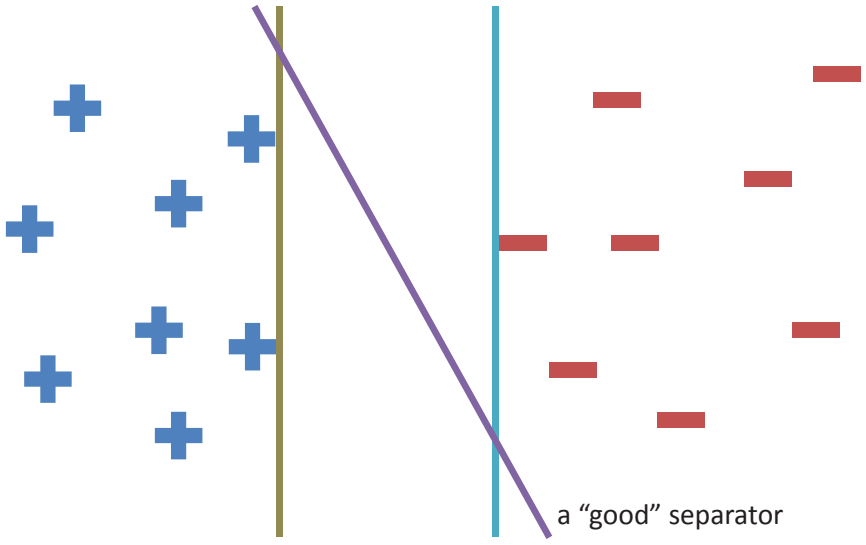


So if $\boldsymbol{\theta}^T \boldsymbol{x} + b \geq 0$, then $y = +1$

if $\boldsymbol{\theta}^T \boldsymbol{x} + b < 0$, then $y = -1$

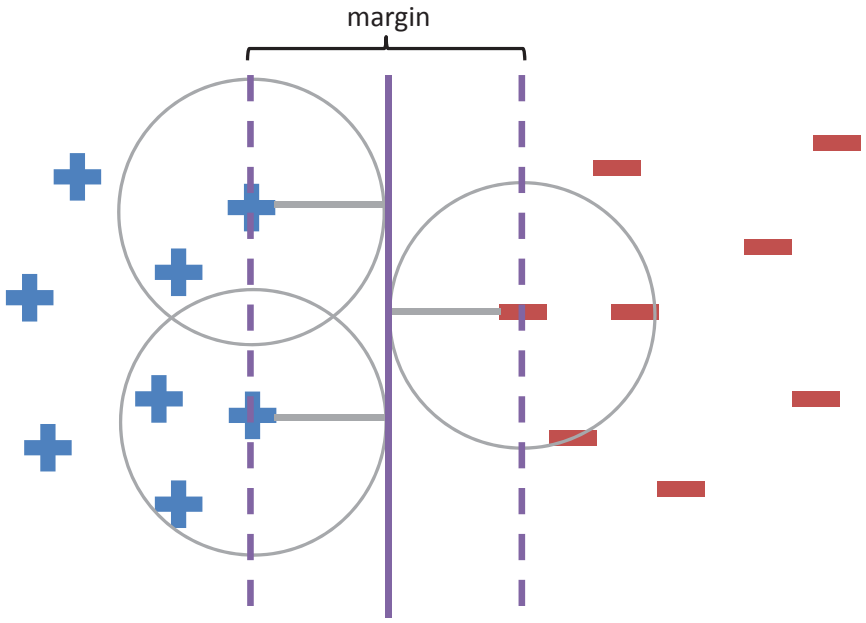
Optimal Linear Separator

Which linear separator is optimal?



Based on slide by Eric Eaton [originally by Tim Oates]

Maximizing the Margin



Based on slide by Eric Eaton [originally by Tim Oates]

Summary

- Perceptron finds one of many possible hyperplanes separating data (if one exists)
- Of possible choices, find one with **maximum margin**
  $\Rightarrow$ Support Vector Machines (SVMs)
- Why?
  - Good according to intuition, theory, and practice
- Some history
  - SVM became famous when, in handwriting recognition task using images as input, it gave accuracy comparable to neural network with hand-designed features

Review: Margins

For training set $\left\{\left(\boldsymbol{x}^{(i)}, y^{(i)}\right)\right\}_{i=1}^{n}$ and parameter vector $\boldsymbol{\theta}$:

- **functional margin** of i^{th} example

  $$\hat{\gamma}^{(i)} = y^{(i)}\left(\boldsymbol{\theta}^T\boldsymbol{x} + b\right)$$

  – positive if $\boldsymbol{\theta}$ classifies $\boldsymbol{x}^{(i)}$ correctly

  – absolute value = "confidence" in predicted label (or "mis-confidence")

- **geometric margin** of i^{th} example

  $$\gamma^{(i)} = \frac{\hat{\gamma}^{(i)}}{\|\boldsymbol{\theta}\|} = \frac{y^{(i)}\left(\boldsymbol{\theta}^T\boldsymbol{x} + b\right)}{\|\boldsymbol{\theta}\|}$$

  – signed distance of $\boldsymbol{x}^{(i)}$ to hyperplane (positive if classified correctly)

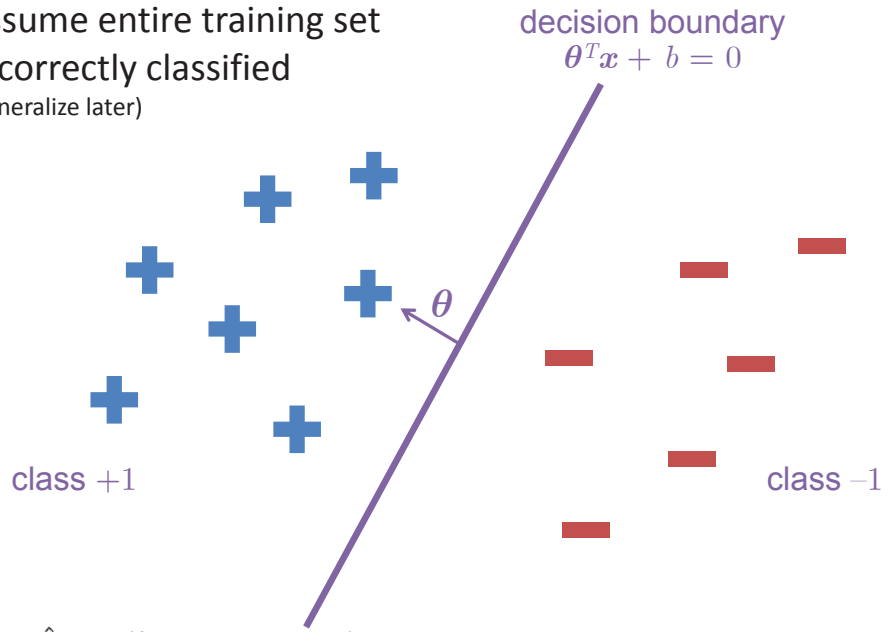
- **margin** over training set

  – minimum functional margin $\hat{\gamma} = \min_i \hat{\gamma}^{(i)}$

  – minimum geometric margin $\gamma = \min_i \gamma^{(i)}$

Maximum-Margin Classifier

Assume entire training set is correctly classified
(generalize later)



enforce $\hat{\gamma} = 1$ (functional margin)

then $\gamma = \frac{\hat{\gamma}}{\|\boldsymbol{\theta}\|} = \frac{1}{\|\boldsymbol{\theta}\|}$ (geometric margin)

SVM Exercise

Consider the following training data:

x_1	x_2	label
1	1	+
2	2	+
2	0	+
0	0	–
1	0	–
0	1	–

- Plot these points. Are the classes $\{+, -\}$ linearly separable?
- Plot the maximum-margin hyperplane by inspection. What is the associated weight vector θ of this hyperplane? Identify the support vectors.
- If you remove one of the support vectors, does the size of the optimal margin decrease, stay the same, or increase?
- *Extra*: Is your answer above true for any data set? Provide a counterexample or give a short proof.

Example adapted from Russell Norvig

(This slide intentionally left blank.)

Generalization

- Large margin $\Rightarrow$ good generalization
- Can we justify this more formally?

  reminder: margin $\gamma = 1/||\boldsymbol{\theta}||$

  large margin

  $\Rightarrow$

- Want more formality?

  – learning theory ☺

Optimization Problem

$$\min_{\boldsymbol{\theta},b} \frac{1}{2}\|\boldsymbol{\theta}\|^2$$

maximum margin = $\max 1/||\boldsymbol{\theta}||$
$\Rightarrow \min ||\boldsymbol{\theta}||$ (non-convex)
$\Rightarrow \min \frac{1}{2} ||\boldsymbol{\theta}||^2$

$$\text{s.t. } y^{(i)}\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \text{ for } i = 1, \ldots, n$$

Note

- convex quadratic objective with linear constraints (n of them)

$\Rightarrow$ Quadratic program (QP)

Polynomial-time algorithms exist for solving QPs

SVMs

Learning Goals

- Solving the SVM optimization problem
  - using Lagrange multipliers (leads to dual problem)
- Allowing misclassified examples
  - using slack variables (leads to soft-margin SVM)
- Describe the SVM loss function
- Allowing non-linear decision boundaries
  - using kernels

Primal and Dual SVMs

Primal

weights associated with features

$$\min_{\boldsymbol{\theta},b} \frac{1}{2}\|\boldsymbol{\theta}\|^2$$

$$\text{s.t. } y^{(i)}\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \text{ for } i = 1,\ldots,n$$

⇓ **MATH! (Lagrange multipliers)**

Dual

weights associated with examples

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y^{(i)}y^{(j)}\underbrace{\langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\rangle}_{\text{dot product } \langle x,y\rangle = x^Ty}$$

$$\text{s.t. } \alpha_i \geq 0 \qquad \text{for } i = 1,\ldots,n$$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

Math Details

For math to work,

Karush-Kuhn-Tucker (KKT) conditions must hold:

for optimal $\boldsymbol{\theta}$, b, $\boldsymbol{\alpha}$,

complementary slackness

$$\alpha_i \left[1 - y^{(i)} \left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + b\right)\right] = 0$$

$\Rightarrow$ if $\alpha_i > 0$, then $y^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + b) = 1$

i.e. if weight for example $i > 0$,

then $\boldsymbol{x}^{(i)}$ lies on a margin boundary ($\boldsymbol{x}^{(i)}$ is a SV)

$\Rightarrow$ if $y^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + b) > 1$, then $\alpha_i = 0$

i.e. if $\boldsymbol{x}^{(i)}$ does not lie a margin boundary ($\boldsymbol{x}^{(i)}$ is not a SV)

then weight for example $i = 0$

Solving SVM Problem

Solve dual in lieu of primal

- Solve for $\boldsymbol{\alpha}$ directly
- How?
  - coordinate descent, sequential minimal optimization (SMO)

Compute primal solution from dual

- to obtain $\boldsymbol{\theta}$,

$$\boldsymbol{\theta} = \sum_{i=1}^{n} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}$$

  - can sum over SVs since only these examples have $\alpha_i > 0$
  - remember, we have not added $x_0 = 1$

- to obtain b, find a SV $\boldsymbol{x}^{(i)}$, solve

$$y^{(i)} \left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + b\right) = 1$$

$$y^{(i)} y^{(i)} \left(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + b\right) = y^{(i)}$$

$$b = y^{(i)} - \boldsymbol{\theta}^T \boldsymbol{x}^{(i)}$$

Sparsity and Generalization

$\alpha_i > 0$ only for SVs

SVs $<<<$ # training examples

$\Rightarrow$ sparseness leads to better generalization

Why?

What is maximum LOOCV error?

(This slide intentionally left blank.)

(This slide intentionally left blank.)

SVMs

Learning Goals

- ✓ Solving the SVM optimization problem
  - using Lagrange multipliers (leads to dual problem)
- Allowing misclassified examples
  - using slack variables (leads to soft-margin SVM)
- Describe the SVM loss function
- Allowing non-linear decision boundaries
  - using kernels

Ideas?

What if Data is Not Linearly Separable?

Idea 1: Add more features

$$\boldsymbol{x} \rightarrow \quad \phi(\boldsymbol{x})$$

Problems

- What ϕ?
- Increasing dimensionality could lead to overfitting

What if Data is Not Linearly Separable?

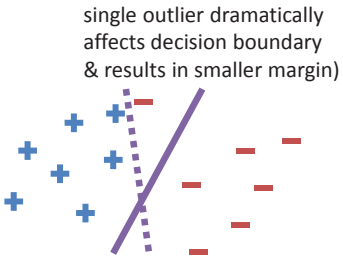
Idea 2: Allow some training data to be misclassified (be robust to outliers)

number of mistakes

$$\min_{\boldsymbol{\theta},b} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n} \mathbb{I}\left[\left[y^{(i)}\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b\right) \leq 0\right]\right]$$

trade-off mistakes

$$\text{s.t. } y^{(i)}\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 \quad \text{for some subset } i$$



Problems

- 0/1 loss not QP and NP-hard
- Does not distinguish near misses from really bad mistakes

Soft-Margin SVM

Idea 2 revised: Allow **slack**

$$\min_{\boldsymbol{\theta},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i$$

Greek xi (zaɪ, saɪ)

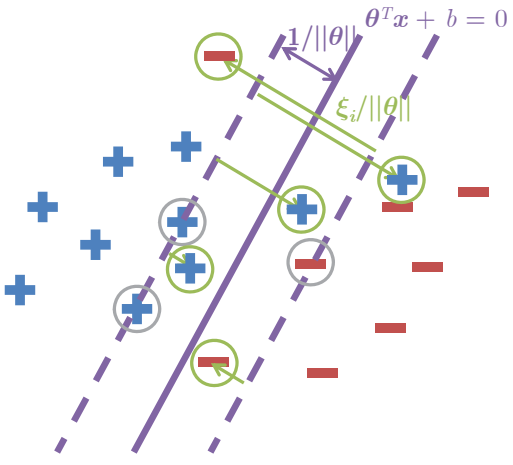
new variable ξ

relax constraints using **slack variables** ξ_i

$$\text{s.t. } y^{(i)}\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi_i \quad \text{for } i = 1,\ldots,n$$

$$\xi_i \geq 0 \quad \text{for } i = 1,\ldots,n$$

Soft-Margin SVM



- margin boundaries still at $\boldsymbol{\theta}^T\boldsymbol{x} + b = \pm 1$
- training example …
  - within margin region when $0 < \xi_i < 1$
  - misclassified when $\xi_i > 1$

- examples can have (functional) margin < 1
- if $\xi_i > 0$, pay penalty $C\xi_i$
- C is **slack penalty** [$C \geq 0$] trade-off between…
  - minimizing $||\boldsymbol{\theta}||^2$ (maximizing margin)
  - ensuring most examples have functional margin of at least 1
- small C [min ½$||\boldsymbol{\theta}||^2$]
- large C [min $\Sigma_i\ \xi_i$]

Dual Formulation

$$\max_{\boldsymbol{\alpha}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y^{(i)} y^{(j)} \langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)} \rangle$$

$$\text{s.t. } 0 \leq \alpha_i \leq C \quad \text{for } i = 1, \ldots, n$$

new constraint: add upper bound $\alpha_i \leq C$

$$\sum_{i=1}^{n} \alpha_i y^{(i)} = 0$$

intuition

- w/o slack

  α_i can increase without bound to prevent misclassification

- w/ slack

  upper bound of C limits α_i to allow misclassifications

Math Details

Different KKT conditions

$$\begin{aligned} \alpha_i = 0 \quad &\Rightarrow \quad y^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + b) > 1 \\ \alpha_i = C \quad &\Rightarrow \quad y^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + b) < 1 \\ 0 < \alpha_i < C \quad &\Rightarrow \quad y^{(i)}(\boldsymbol{\theta}^T \boldsymbol{x}^{(i)} + b) = 1 \end{aligned}$$

Types of support vectors

- hard-margin SVM has only one type of SV
  - points on margin boundaries
- soft-margin SVM has three types of SVs
  - points on margin boundaries

    $\xi_i = 0, \quad 0 < \alpha_i < C$
  - points within margin region but still on correct side

    $0 < \xi_i < 1$
  - points on wrong side of hyperplane (misclassified)

    $\xi_i \geq 1$

margin violators

$\alpha_i = C$

SVMs

Learning Goals

- ✓ Solving the SVM optimization problem
  - using Lagrange multipliers (leads to dual problem)
- ✓ Allowing misclassified examples
  - using slack variables (leads to soft-margin SVM)
- Describe the SVM loss function
- Allowing non-linear decision boundaries
  - using kernels

Deriving the Loss Function

Recall soft-margin SVM optimization problem

$$\min_{\boldsymbol{\theta},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{i=1}^{n}\xi_i$$

$$\text{s.t. } y^{(i)}\left(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b\right) \geq 1 - \xi_i \quad \text{for } i = 1,\ldots,n$$

$$\xi_i \geq 0 \quad \text{for } i = 1,\ldots,n$$

- if $y^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b) \geq 1$,
  then $\xi_i = 0$ (no penalty)
- if $y^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b) < 1$,
  then $\xi_i = 1 - y^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b)$ (linear penalty)

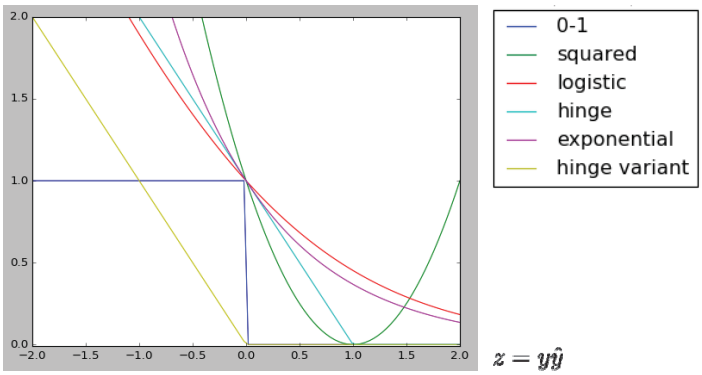
$\Rightarrow \xi_i = \max(0,\ 1 - y^{(i)}(\boldsymbol{\theta}^T\boldsymbol{x}^{(i)} + b))$

hinge loss (encoded in slack variables)

Loss Functions

$\hat{y}$ = raw prediction

- 0/1 $L(y, \hat{y}) = H(-y\hat{y})$ Heaviside step function
- squared $L(y, \hat{y}) = (1 - y\hat{y})^2$
- logistic $L(y, \hat{y}) = \frac{1}{\log 2} \log\left(1 + e^{-y\hat{y}}\right)$
- hinge $L(y, \hat{y}) = \max(0, 1 - y\hat{y})$
- exponential $L(y, \hat{y}) = e^{-y\hat{y}}$
- (variant of hinge loss) $L(y, \hat{y}) = \max(0, -y\hat{y})$



Soft-Margin SVM as Regularization

Let

$$L(y, \hat{y}) = \max(0, 1 - y\hat{y})$$

So

$$\min_{\boldsymbol{\theta}, b} \underbrace{\frac{1}{2}\|\boldsymbol{\theta}\|^2}_{\text{regularization}} + C \sum_{i=1}^{n} \underbrace{L\left(y^{(i)}, \hat{y}^{(i)}\right)}_{\text{empirical risk using hinge loss}}$$

note:
$\hat{y} = \boldsymbol{\theta}^T \boldsymbol{x} + b$ is "raw" output, not predicted class

Standard form of regularization

$$\min_{\boldsymbol{\theta}, b} \frac{\lambda}{2}\|\boldsymbol{\theta}\|^2 + \frac{1}{n} \sum_{i=1}^{n} L\left(y^{(i)}, \hat{y}^{(i)}\right)$$

So $C \propto \frac{1}{\lambda}$

low regularization
small λ, large C

high regularization
large λ, small C