SVMs with Slack
(Not Linearly Separable)

Rishabh Iyer
University of Texas at Dallas

Slides adapted from Nick Rouzzi

SVMs with Slack (Remove Linear Separability)

- Allow misclassification
  - Penalize misclassification linearly (just like in the perceptron algorithm)
    - Again, easier to work with than counting misclassifications
    - Objective stays convex
  - Will let us handle data that isn't linearly separable!
  - Idea: Take the constraints into the main objective
    - The objective function then becomes exactly like what we have seen in Perceptron/Linear Regression

SVMs with Slack

$$\min_{w,b,\xi} \left[\frac{1}{2}\|w\|^2 + c\sum_i \xi_i \right]$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

With Linear Sep:

$$y_i\left(w^T x^{(i)} + b\right) \geq 1$$

SVMs with Slack

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_i \xi_i$$

such that

$$y_i(w^T x^{(i)} + b) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

Potentially allows some points to be misclassified/inside the margin

SVMs with Slack

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_i \xi_i$$

Constant c determines degree to which slack is penalized

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

SVMs with Slack

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_i \xi_i$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

- How does this objective change with c?

SVMs with Slack

$$\min_{w,b,\xi} \underbrace{\frac{1}{2}\|w\|^2}_{\text{Margin}} + c \underbrace{\sum_i \xi_i}_{\text{Train Error.}}$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

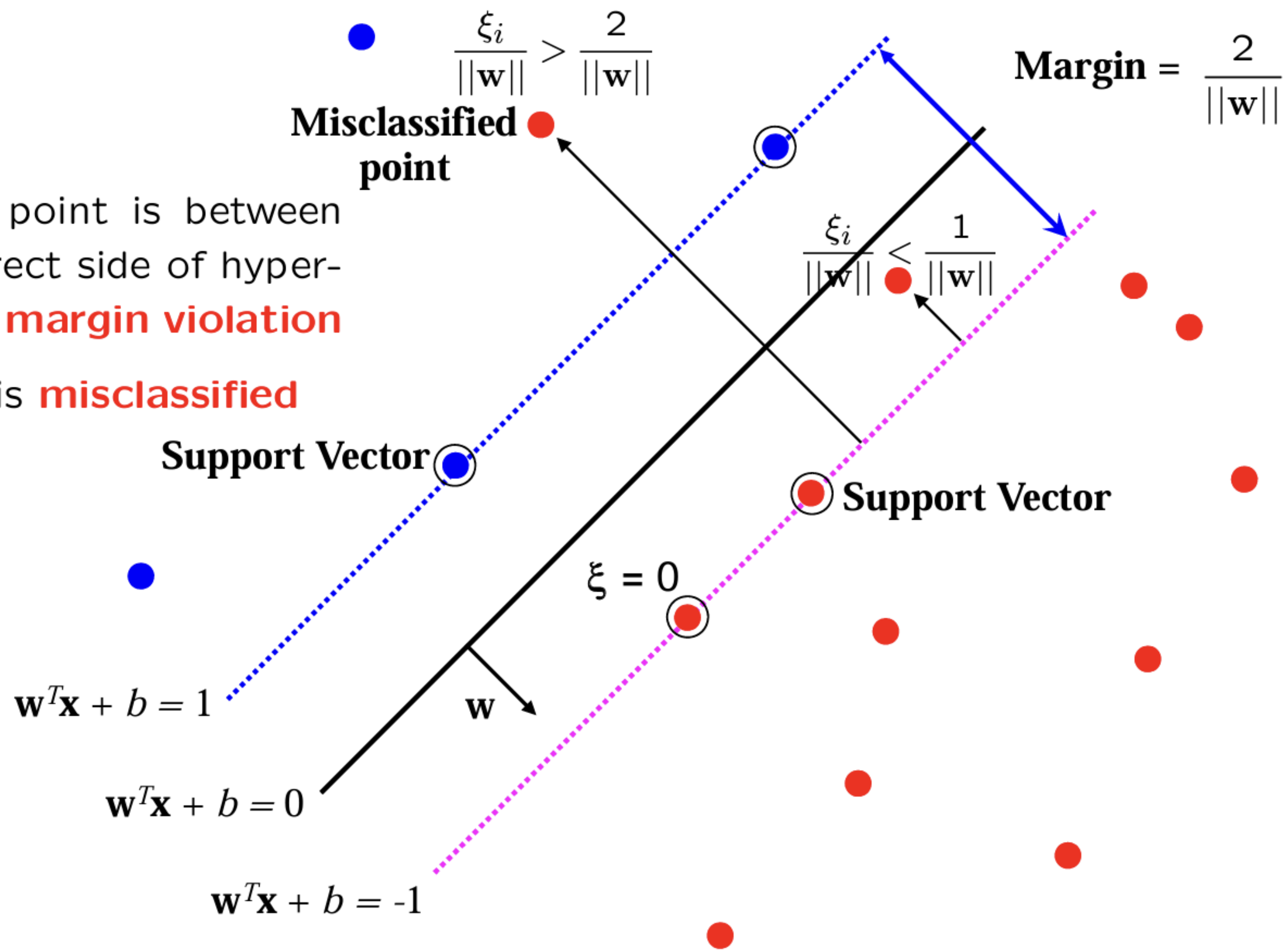
- How does this objective change with c?
  - As $c \to \infty$, requires a perfect classifier
  - As $c \to 0$, allows arbitrary classifiers (i.e., ignores the data)

SVMs with Slack: Illustration

$$\xi_i \geq 0$$

for $0 < \xi \leq 1$ point is between margin and correct side of hyper-plane. This is a **margin violation**

for $\xi > 1$ point is **misclassified**



SVMs with Slack

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_i \xi_i$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

- How should we pick c?

SVMs with Slack

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_i \xi_i$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

- How should we pick c?
  - Divide the data into three pieces training, testing, and validation
  - Use the validation set to tune the value of the hyperparameter c

Select Hyperparameters on Val set

Training (Tr)	Test (Te)	Val (V)
70%	15%	15%

Model: (w, b)

Hyper-params: c

Range of c: $C \in [10^{-5}, 10^{-4}, 10^{-3}, \ldots, 10^{4}, 10^{5}]$

for $c \in C$:

$\quad w_c, b_c = \text{Train}(Tr, c)$

$\quad Acc_c = \text{Test}(V, (w_c, b_c))$

$$c^* = \max_c Acc_c$$

Final model $= (w_{c^*}, b_{c^*})$

Get test accuracy on Test set.

↳ Generalization of model at Deployment.

Evaluation Methodology

- General learning strategy
  - Build a classifier using the training data
  - Select hyperparameters using validation data
  - Evaluate the chosen model with the selected hyperparameters on the test data → what we expect the model to perform at deployment.

How can we tell if we overfit the training data?

validation data.

ML in Practice

- Gather Data + Labels
- Select feature vectors
- Randomly split into three groups
  - Training set
  - Validation set
  - Test set
- Experimentation cycle
  - Select a “good” hypothesis from the hypothesis space
  - Tune hyper-parameters using validation set
  - Compute accuracy on test set (fraction of correctly classified instances)

SVMs with Slack

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_i \xi_i$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

- What is the optimal value of ξ for fixed w and b?

SVMs with Slack

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_i \xi_i$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

- What is the optimal value of ξ for fixed w and b?
  - If $y_i\left(w^T x^{(i)} + b\right) \geq 1$, then $\xi_i = 0$
  - If $y_i\left(w^T x^{(i)} + b\right) < 1$, then $\xi_i = 1 - y_i\left(w^T x^{(i)} + b\right)$

SVMs with Slack

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + c\sum_i \xi_i$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

- We can formulate this slightly differently
  - $\xi_i = \max\{0, 1 - y_i\left(w^T x^{(i)} + b\right)\}$
  - Does this look familiar?
  - Hinge loss provides an upper bound on Hamming loss

0/1 loss

Hinge Loss Formulation

- Obtain a new objective by substituting in for ξ

$$\min_{w,b} \underbrace{\frac{1}{2}\|w\|^2}_{\text{Generalization}} + c\sum_i \max\{0, \underbrace{1 - y_i\left(w^T x^{(i)} + b\right)}_{\text{Train error.}}\}$$

Can minimize with gradient descent!

Perceptron: $\min_{w,b} \sum_i \max\left(0, -y_i\left(w^T x^{(i)} + b\right)\right)$

Hinge Loss Formulation

- Obtain a new objective by substituting in for ξ

$$\min_{w,b} \underbrace{\frac{1}{2}\|w\|^2}_{\text{Penalty to prevent overfitting}} + \underbrace{c\sum_{i} \max\{0, 1 - y_i\left(w^T x^{(i)} + b\right)\}}_{\text{Hinge loss}}$$

(Minimizing Loss on Train Set]

REGULARIZATION

- Until now, we have seen the following optimization problems:

$$\min_{w,b} \sum_i L\left(f\left(x^{(i)}, w, b\right), y_i\right)$$

- In the case of Linear regression, L was the squared loss
- In Perceptron, L was Perceptron Loss
- The regularized version of this is:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + c \sum_i L\left(f\left(x^{(i)}, w, b\right), y_i\right)$$

- c is a hyper-parameter (again, to be tunes on validation set)

Perceptron vs Hinge vs Square vs Zero-One Loss

Hinge loss $\geq$ 0/1 Loss Everywhere

Imbalanced Data

- If the data is imbalanced (i.e., more positive examples than negative examples), may want to evenly distribute the error between the two classes

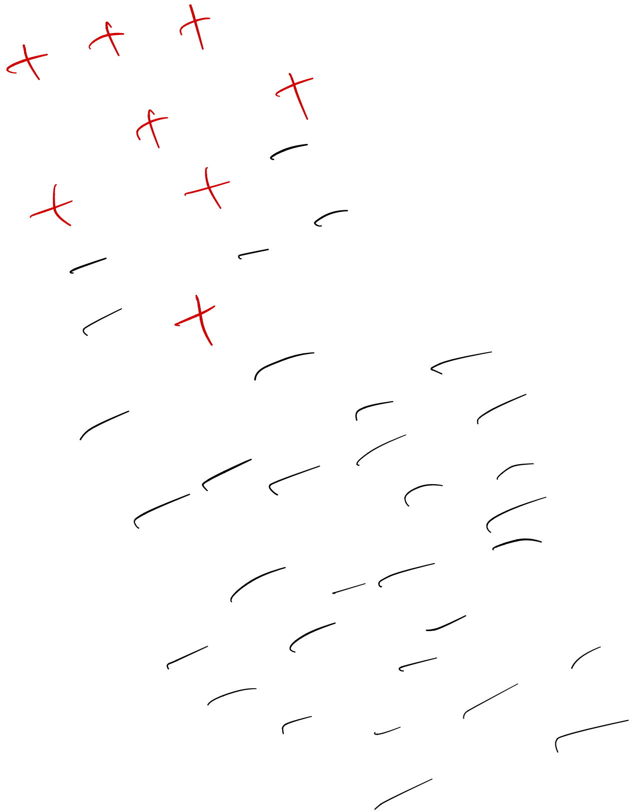
$$\min_{w,b,\xi} \frac{1}{2}||w||^2 + \frac{c}{N_+} \sum_{i:y_i=1} \xi_i + \frac{c}{N_-} \sum_{i:y_i=-1} \xi_i$$

such that

$$y_i\left(w^T x^{(i)} + b\right) \geq 1 - \xi_i, \text{for all } i$$

$$\xi_i \geq 0, \text{for all } i$$

Why is Imbalance a Problem?



Generalization

- We argued, intuitively, that SVMs generalize better than the perceptron algorithm
  - How can we make this precise?

There exists a theoretical connection between Margin & Generalization.

Roadmap

- Where are we headed?
  - Other simple hypothesis spaces for supervised learning
    - k nearest neighbor
    - Decision trees
  - Probabilistic Methods
    - Bayesian Methods
    - Naïve Bayes
    - Logistic Regression