

FAST ONLINE CLASSIFICATION with SUPPORT VECTOR MACHINES

Şeyda Ertekin

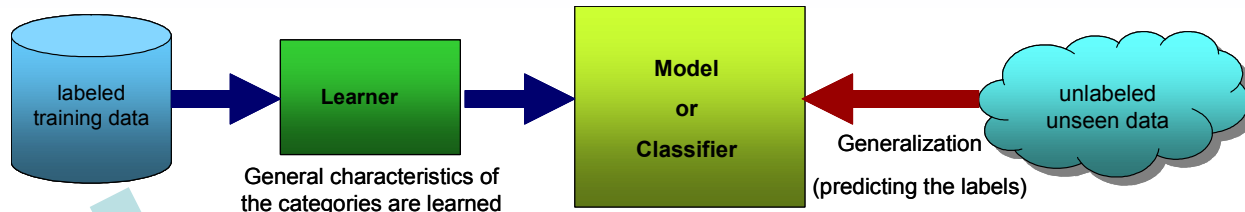
Computer Science & Eng.
Penn State University
University Park, PA

Léon Bottou

NEC Labs America
Princeton, NJ

C. Lee Giles

College of Information Sci. & Tech.
Penn State University
University Park, PA



Task: Classification

Motivation

How are we going to process them?

In 10 years: CPU speed x 100, disc size x 1000

We need machine learning algorithms which

- give **high classification accuracies**
- are **fast**
- can **scale to large datasets**

Online SVM: LASVM

- Reorganization of SMO
- Can deal with streaming data
- Has also an SV removal step
- Less memory demand
- Speed improvement

Base Algorithm: Support Vector Machines (SVMs)

Inseparable Case:

$$\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i \quad \text{with the constraint } \forall i \quad y_i f(x_i) \geq 1 - \xi_i$$

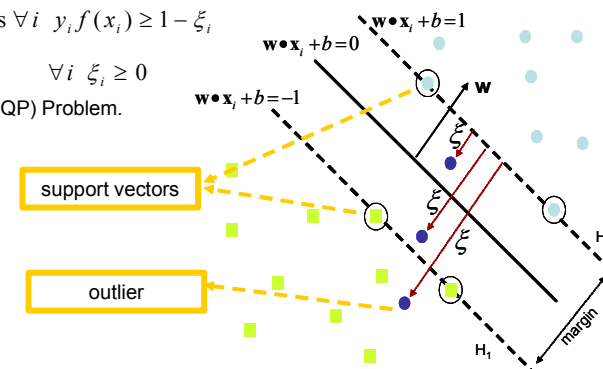
$$\forall i \quad \xi_i \geq 0$$

We need to solve SVM Quadratic Programming (QP) Problem.

Dual of the Convex Optimization Problem:

$$\max_{\alpha} W(\alpha) = \sum_i \alpha_i y_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j K(x_i, x_j)$$

$$\text{with the constraints } \begin{cases} \sum_i \alpha_i = 0 \\ A_i \leq \alpha_i \leq B_i \\ A_i = \min(0, C y_i) \\ B_i = \max(0, C y_i) \end{cases}$$



After solving QP, we get $\rightarrow f_o(x) = \sum_i \alpha_i \Phi(x_i) \Phi(x) + b$

Each α_i determines how much each training example influences the SVM solution.

$$\omega = \sum_{i=1}^n \alpha_i \Phi(x_i) \quad \begin{matrix} \alpha_i \neq 0 \text{ for support vectors} \\ \alpha_i = 0 \text{ for non support vectors} \end{matrix}$$

SVMs give very good classification accuracies but they may be quite costly with large datasets.

LASVM with noisy data

Online SVM with Active Learning

Online SVM with Non-Convex Loss Function

LASVM + Active Learning

Not all training examples are equally informative!

- How can we select the most informative one?
active learning
- Do we really have to search the entire training set?
Not really!

The randomized search first samples n random training examples and selects the best one among those n examples.

Hinge Loss:

Inseparable case $\min_{w,b} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l H(y_i f_\theta(x_i))$

$H(y_i f_\theta(x_i)) = \max(0, 1 - y_i f_\theta(x_i))$

No loss if $y_i f_\theta(x_i) > 1$

With the Hinge loss outliers are getting more attention than they should!

Fast learning especially with noisy data

Less support vectors, so testing is fast as well

Scalable to large datasets

Ramp Loss:

$J^s(\theta) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^l H_1(y_i f_\theta(x_i))$

minimum w must satisfy $w = \sum_{i=1}^L -C y_i H_1'(y_i f_\theta(x_i)) \Phi(x_i)$

$s = -1 \rightarrow$ outliers are not SVs anymore.

$s = 0 \rightarrow$ misclassified examples are not SVs anymore