UNIVERSITAT POLITÈCNICA DE CATALUNYA

FACULTAT D'INFORMÀTICA DE BARCELONA

Extension Report: Stop Condition

Analysis of the SVM-RFE algorithm for feature selection

Author:
Robert PLANAS

Director:
Luis A. BELANCHE

Bachelor Degree in Informatics Engineering
Specialization: Computing

April 23, 2021

Contents

Chapter 1

Stop Condition

This modification intends to find the optimal number of features that still performs reasonably well in terms of accuracy (i.e. stop condition) in a non-expensive manner.

1.1 Description and reasoning

Such objective requires specifying a trade-off between accuracy and size of the feature subset. We could use cross-validation together with a function such as $a =$ accuracy, $p =$ percentage of selected features, $f(a, p) = t_0 a + (1 - p)$. Notice that a new parameter t_0 is required to determine the importance of one term over the other. Cross-validation however is a slow technique, as it requires to execute the whole SVM-RFE algorithm many times.

The alternative we propose here is to use the weights w calculated at every iteration of the SVM-RFE algorithm as an approximation of the accuracy difference. We hypothesize that greater values in the ranking criteria of the eliminated features will result in a greater loss of accuracy. That is, when eliminating k features, we calculate the commutative sum of the ranking criteria of the eliminated features and, if above a certain threshold, we stop.

1.2 Pseudocode formalization

Definitions:

- $X_0 = [\vec{x_0}, \vec{x_1}, \ldots, \vec{x_k}]^T$ list of observations.
- $\vec{y} = [y_1, y_2, \ldots, y_k]^T$ list of labels.

Algorithm 1: SVM-RFE with Stop Condition

Input: t, t_0 `// ` t ` = step, ` t_0 ` = threshold, ` $0 \leq t_0$
Output: $\vec{r}$
Data: $X_0, \vec{y}$

$\vec{s} = [1, 2, \ldots, n]$ `// subset of surviving features`
$\vec{r} = []$ `// feature ranked list`
$q = 0$ `// stop condition`
while $|\vec{s}| > 0 \wedge q > s$ **do**
   `/* Restrict training examples to good feature indices */`
   $X = X_0(:, \vec{s})$
   `/* Train the classifier */`
   $\vec{\alpha} = \texttt{SVM-train}(X, y)$
   `/* Compute the weight vector of dimension length` $|\vec{s}|$ `*/`
   $\vec{w} = \sum_k \vec{\alpha}_k \vec{y}_k \vec{x}_k$
   `/* Compute the ranking criteria */`
   $\vec{c} = [(w_i)^2 \text{ for all } i]$
   `/* Find the` t `features with the smallest ranking criterion */`
   $\vec{f} = \texttt{argsort}(\vec{c})(:t)$
   `/* Sum selected ranking criteria to determine stop cond. */`
   $q = \sum_i f_i$
   `/* Update the feature ranking list */`
   $\vec{r} = [\vec{s}(\vec{f}), \ldots \vec{r}]$
   `/* Eliminate the features with the` t `smallest ranking criterion */`
   $\vec{s} = [[\ldots \vec{s}(1 : f_i - 1), \ldots \vec{s}(f_i + 1 : |\vec{s}|)] \text{ for all } i]$
end

1.3 Results