7.1.3 Multi-class SVMs

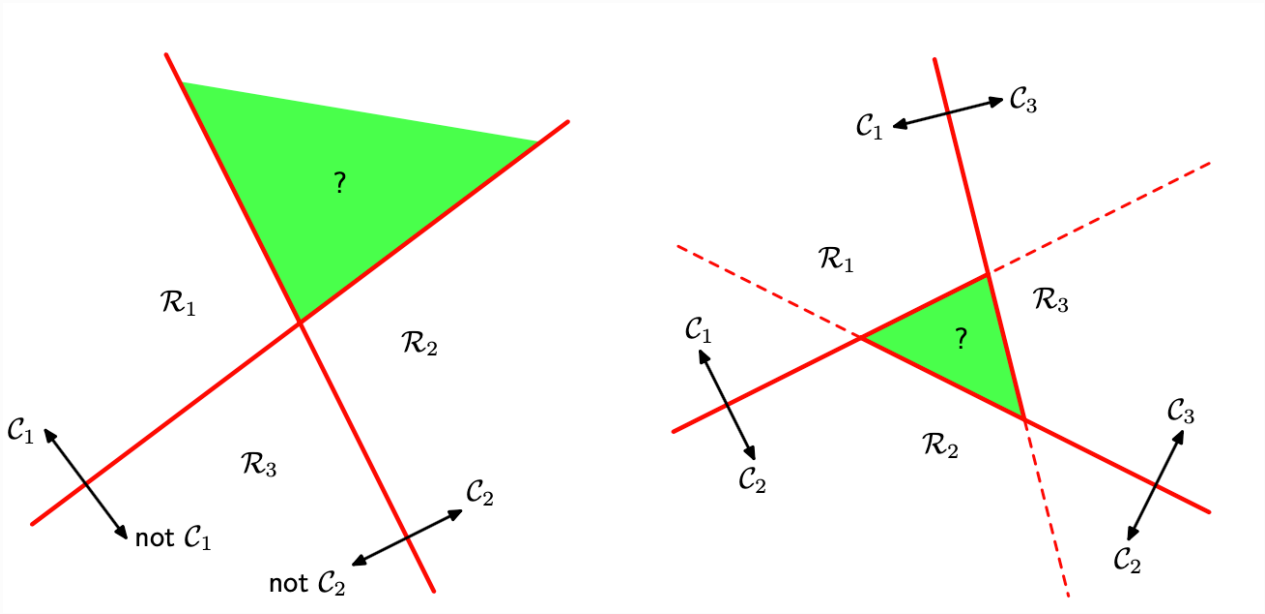
The support vector machine is fundamentally a two-class classifier. However, we solve $K > 2$ classes.

Various methods have been proposed for combining mult-two-class SVMs to build a multi-class classifier.

Famous method is the one : rest approach which constructs K separate SVMs, in which

the k_{th} model $y_k(x)$ is trained using the data from class C_k as the positive examples and

the data from the remaining $K - 1$ classes as the negative examples.



However, in Figure, we saw that using the decisions of the decisions of the individual classifiers can lead to inconsistent results in which an input is assigned to multiple classes simultaneously. (input이 여러 클래스에 동시에 할당되면 결과값이 비 일관적으로 나올수도 있다.)

This problem is sometimes addressed by making predictions for new inputs x using $y(x) = \max_k y_k(x)$.

This heuristic approach suffers from the problem that the different classifiers were trained on different tasks, and there is no guarantee that the real-valued quantities $y_k(x)$ for different classifiers will have appropriate scales. (각각의 분류기가 다른 문제를 training 하고, 서로다른 분류기가 적절한 스케일을 가질지 모르겠다.)

Another problem with the one vs rest approach is that the training sets are imbalanced.

Problem

1. Different classifiers were trained on different tasks. => 전체를 다 한다음에 평균치로 각각 할당 하

면 되지 않나?(속도)
2. Different classifiers not guarantee appropriate scales. => scaling 하면 되지 않나?
3. Training sets are imbalanced. 1/n 하면 되지 않나?

EX)

Approach 1. (다른 모습의 one vs the rest)

We have ten classes each with equal numbers of training dat points,

then the individual classifiers are trained

data sets 90% : - and 10% : + examples, and the symmetry of the original problem is lost.

A variant of the one vs rest scheme was proposed by Lee

who modify the target values so that the positive class has target +1 and the negative class has target $\frac{-1}{K-1}$.

Weston and Watkins define a single objective function for training all K SVMs simultaneously,

based on maximizing the margin from each to remaining classes.

However, this can result in much slower training because, instead solving K separate optimization problems each over N data points with an overall cost of $O(KN^2)$,

a single optimization problem of size $(K-1)N$ must be solved giving an overall cost of $O(K^2N^2)$.

Approach 2. (One vs one)

Train $\frac{K(K-1)}{2}$ different 2-class SVMs on all possible pairs of classes, and then so classify test points according to which class has the highest number of 'votes'.

$\frac{K(K-1)}{2}$ 다른 2-class SVM 들을 모든 class의 pair들에 대해 조져, 그런뒤 test points를 투표가 많이된 class를 따라서 분류해

As in Figure, this can lead to ambiguities in the resulting classification.

Also, train, test =>expensive computation.

Approach 2-2.(DAGSVM)

Organizing the pairwise classifiers into a directed acyclic graph.

For K classes, the DAGSVM has a total of $\frac{K(K-1)}{2}$ classifiers, and to classify a new test point only $K-1$ pairwise classifiers need to be evaluated, with the particular classifiers used depending on which path through the graph is traversed.

Approach 2-3. (Based on error-correcting output codes) generalization of Approach 2

Developed by 디터리치 and 바키리(to multi class problem) and 알웨인 applied to support vector machines.

More general partitions of the classes are used to train the individual classifiers.

K class themselves are represented as particular sets of responses from the two-class classifiers chosen,

and together with a suitable decoding scheme,

this gives robustness to errors and to ambiguity in the outputs of the individual classifiers.

Although the application of SVMs to multi class classification problems remains an open issue,

Despite of ad-hoc formulation and practical limitations, in practice the 1vs rest approach is the most used,

Single-class SVM to solve and unsupervised problem related to probability density estimation

These methods aim to find a smooth boundary enclosing a region of high density.

The boundary is chosen to represent a quantile of the densest, i.e,

the probability that a data point drawn from the distribution will land inside that region is given by a fixed number between 0 and 1 that is specified in advance.

This is more restricted problem than estimating the full density but may be sufficient in specific applications.

Two approaches to this problem using support vector machines have been proposed.

Approach 1.스콜코프 algorithm

Tries to find a hyperplane that separates all but a fixed fraction ν of the training data

from the origin while at the same time maximizing the distance of the hyperplane from the origin,

Approach 2. Tax and Duin algorithm

look for the smallest sphere in feature space that contains all but a fraction ν of the data points.

For kernels $k(x, x')$ that are functions only of $x - x'$, the two algorithms are equivalent.