Kernel Logistic Regression and the Import Vector Machine

Ji Zhu
Department of Statistics
Stanford University
Stanford, CA 94305
jzhu@stat.stanford.edu

Trevor Hastie
Department of Statistics
Stanford University
Stanford, CA 94305
hastie@stat.stanford.edu

Abstract

The support vector machine (SVM) is known for its good performance in binary classification, but its extension to multi-class classification is still an on-going research issue. In this paper, we propose a new approach for classification, called the import vector machine (IVM), which is built on kernel logistic regression (KLR). We show that the IVM not only performs as well as the SVM in binary classification, but also can naturally be generalized to the multi-class case. Furthermore, the IVM provides an estimate of the underlying probability. Similar to the “support points” of the SVM, the IVM model uses only a fraction of the training data to index kernel basis functions, typically a much smaller fraction than the SVM. This gives the IVM a computational advantage over the SVM, especially when the size of the training data set is large.

1 Introduction

In standard classification problems, we are given a set of training data \((x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N)\), where the output \(y_i\) is qualitative and assumes values in a finite set \(C\). We wish to find a classification rule from the training data, so that when given a new input \(x\), we can assign a class \(c\) from \(C\) to it. Usually it is assumed that the training data are an independently and identically distributed sample from an unknown probability distribution \(P(X,Y)\).

The support vector machine (SVM) works well in binary classification, i.e. \(y \in \{0, 1\}\), but its appropriate extension to the multi-class case is still an on-going research issue. Another weakness of the SVM is that it only estimates \(\text{sign}(p(x) - 1/2)\), while the probability \(p(x)\) is often of interest itself, where \(p(x) = P(Y = 1 | X = x)\) is the conditional probability of a point being in class 1 given \(X = x\). In this paper, we propose a new approach, called the import vector machine (IVM), to address the classification problem. We show that the IVM not only performs as well as the SVM in binary classification, but also can naturally be generalized to the multi-class case. Furthermore, the IVM provides an estimate of the probability \(p(x)\). Similar to the “support points” of the SVM, the IVM model uses only a fraction of the training data to index the kernel basis functions. We call these training data import points. The computational cost of the SVM is \(O(N^3)\), while the computational cost of the IVM is \(O(N^2 q^2)\), where \(q\) is the number of import points. Since \(q\) does not tend to
increase as \( N \) increases, the IVM can be faster than the SVM, especially for large training
data sets. Empirical results show that the number of import points is usually much less than
the number of support points.

In section (2), we briefly review some results of the SVM for binary classification and
compare it with kernel logistic regression (KLR). In section (3), we propose our IVM
algorithm. In section (4), we show some simulation results. In section (5), we generalize
the IVM to the multi-class case.

2 Support vector machines and kernel logistic regression

The standard SVM produces a non-linear classification boundary in the original input space
by constructing a linear boundary in a transformed version of the original input space.
The dimension of the transformed space can be very large, even infinite in some cases.
This seemingly prohibitive computation is achieved through a positive definite reproducing
kernel \( K \), which gives the inner product in the transformed space.

Many people have noted the relationship between the SVM and regularized function es-
timation in the reproducing kernel Hilbert spaces (RKHS). An overview can be found in
Evgeniou et al. (1999), Hastie et al. (2001) and Wahba (1998). Fitting an SVM is equiva-
lent to minimizing:

\[
\frac{1}{N} \sum_{i=1}^{N} (1 - y_i f(x_i))^+ + \lambda \|f\|^2_{\mathcal{H}_K}.
\]

with \( f = b + h, \ h \in \mathcal{H}_K, \ b \in \mathcal{R} \). \( \mathcal{H}_K \) is the RKHS generated by the kernel \( K \). The
classification rule is given by \( \text{sign}[f] \).

By the representer theorem (Kimeldorf et al (1971)), the optimal \( f(x) \) has the form:

\[
f(x) = b + \sum_{i=1}^{N} a_i K(x, x_i).
\]

It often happens that a sizeable fraction of the \( N \) values of \( a_i \) can be zero. This is a
consequence of the truncation property of the first part of criterion (1). This seems to be an
attractive property, because only the points on the wrong side of the classification boundary,
and those on the right side but near the boundary have an influence in determining the
position of the boundary, and hence have non-zero \( a_i \)'s. The corresponding \( x_i \)'s are called
support points.

Notice that (1) has the form \( \text{loss} + \text{penalty} \). The loss function \((1 - yf)_+ \) is plotted in Figure
1, along with several traditional loss functions. As we can see, the negative log-likelihood
(NLL) of the binomial distribution has a similar shape to that of the SVM. If we replace
\((1 - yf)_+ \) in (1) with \( \ln(1 + e^{-yf}) \), the NLL of the binomial distribution, the problem
becomes a KLR problem. We expect that the fitted function performs similarly to the SVM
for binary classification.

There are two immediate advantages of making such a replacement: (a) Besides giving
a classification rule, the KLR also offers a natural estimate of the probability \( p(x) = e^f / (1 + e^f) \), while the SVM only estimates \( \text{sign}[p(x) - 1/2] \); (b) The KLR can natu-
 rally be generalized to the multi-class case through kernel multi-logit regression, whereas
this is not the case for the SVM. However, because the KLR compromises the hinge loss
function of the SVM, it no longer has the “support points” property; in other words, all the
\( a_i \)'s in (2) are non-zero.

KLR is a well studied problem; see Wahba et al. (1995) and references there; see also
Green et al. (1985) and Hastie et al. (1990).