Kernel Logistic Regression for Phoneme Recognition

P. Karsmakers[a,b], K. Pelckmans[b], J.A.K. Suykens[b], H. Van hamme[c]

[a]K.H.Kempen (association K.U.Leuven), IIBT, kleinhoefstraat 4, B-2440 Geel

[b]K.U.Leuven, ESAT/SCD-SISTA, Kasteelpark Arenberg 10, B-3001 Heverlee

[c]K.U.Leuven, ESAT/PSI-SPEECH, Kasteelpark Arenberg 10, B-3001 Heverlee

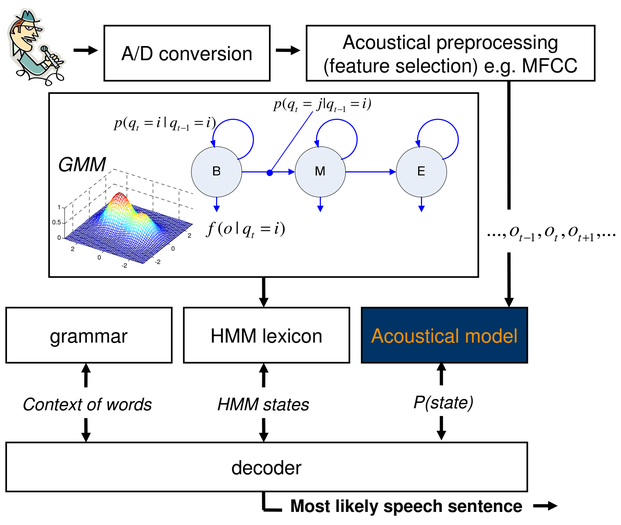
1. Problem statement

This research studies the extension of a multiclass logistic regression technique for the task of phoneme recognition. Herefor, a kernel version is derived based on a penalized likelihood criterion. The choice of this approach over an empirical risk minimization approach as performed by the Support Vector Machines (SVMs), is that the former yields probabilistic outcomes instead of a binary decision. This is particularly important in this subtask of speech recognition as it permits a proper integration of the phoneme recognition module in the full sequence. Specifically, it allows for a proper connection to a Hidden Markov Model (HMM) which makes different words out of a sequence of phonemes.

Consider a set of samples $D = \{(x_i, y_i)\}_{i=1}^{N} \subset \mathbb{R}^{d+1} \times \{1, \ldots, C\}$

Let the inputvectors x_i be speech signal MFCCs and the outputvectors y_i phonemelabels.

2. Classical automatic speech recognizer



3. Kernel Logistic Regression (K≥2)

In multi-class logistic regression the a-posteriori probability of class membership is modeled via the linear function.

$$f(x) = \beta^T x$$

To be able to interpret the output *f(x)* as a probability estimate logit stochastic models are used

$$\begin{cases} P(Y = 1 \mid X = x) = \frac{\exp(\beta_1^T x)}{1 + \sum_{c=1}^{C-1} \exp(\beta_c^T x)} \\ P(Y = 2 \mid X = x) = \frac{\exp(\beta_2^T x)}{1 + \sum_{c=1}^{C-1} \exp(\beta_c^T x)} \\ \vdots \\ P(Y = C-1 \mid X = x) = \frac{1}{1 + \sum_{c=1}^{C-1} \exp(\beta_c^T x)} \end{cases}$$

where the $\{\beta_c\}_{c=1}^{C-1}$ denote the different parameters of the *C-1* linear models.

The class membership of a new point *x'* can be made by the Bayesian classification rule which is given by

$$arg\max_{c \in 1,2,\ldots,C} \; P(Y = c | X = x')$$

The common method to infer the parameters is via the use of penalized maximum likelihood which can be written as

$$\mathcal{L}_\lambda(\beta) = \prod_{i=1}^{n} P(Y = y_i | X = x_i) \prod_{c=1}^{C-1} (\beta_c^T \beta_c)^\lambda$$

Most often, a Newton-Raphson based strategy is used to optimize the loglikelihood. It is well-known that this procedure can be rewritten in terms of an iteratively re-weighted least squares (IRLS) algorithm which consists of two steps:

For k = 1, … , L
1. *Compute regularized WLS*
2. *Recompute weights W*

4. Weigthed LS-SVM

Here we study the nonlinear extension to kernel machines where the inputs *x* are mapped to a high dimensional space via $\varphi(\cdot)$.
Now, both steps can be easily reformulated in terms of a weighted LS-SVM.

$$\min_{s^k, e^k} \frac{1}{2} e^{k^T} (W^k) e^k + \frac{\lambda}{2} ||\beta^k||^2$$
$$such\ that\ z^k = \Phi s^k + e^k$$

with s^k the k'th Newton-Raphson step, $\beta^L = \sum_{k=1}^{L} s^k$ and $\Phi = [\varphi(x_1); ...; \varphi(x_N)]$ where $\varphi : \mathbb{R}^d \to \mathbb{R}^{D_\varphi}$

The dual solution can be represented by

$$\hat{f}(x') = \sum_{i=1}^{N} \frac{1}{\lambda} \hat{\alpha}_i K(x_i, x')$$

where $\hat{\alpha} = (\hat{\alpha_1}, \ldots, \hat{\alpha_N})^T$ is solved by a linear system

5. Large scale algorithm

The main drawback of kernel logistic regression is that all training vectors are necessary in the final model. To become a sparse solution we have to achieve a resulting kernel expansion based on only a subset *S* of the training data.

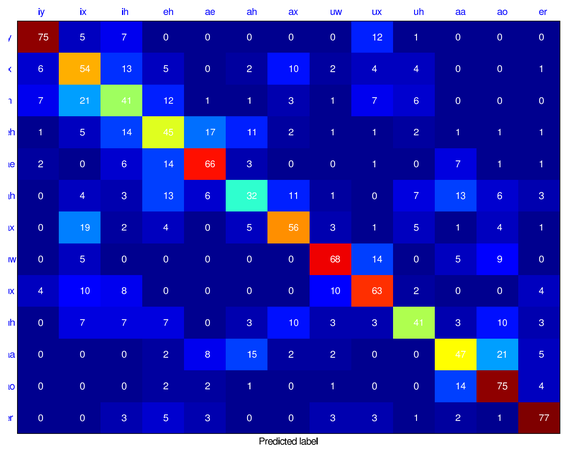
$$\hat{f}(x') = \sum_{i=1}^{|S|} \frac{1}{\lambda} \alpha_i K(x_i, x') \quad |S| \ll N$$

The size of the Hessian in the dual form is proportional to the number of datapoints, we suggest an approach where each optimization step is based on a well-chosen subset of the available dataset (equivalent to stochastic gradient descent). This automatically leads to a sparse solution.

While IRLS performs a Newton-Raphson optimization strategy, we resort to a version where one descends per parameterset β_c. Though the convergence becomes slightly worse, each step can be calculated much faster when using the dual representation.

6. TIMIT dataset

Experiments are carried out on the TIMIT dataset which consists of quasi-phonetically balanced American English training sentences, segmented on the phoneme level.
The confusion matrix is used to evaluate the discriminative ability of the classifier.



Acknowledgements

-(KP): BOF PDM/05/161, FWO grant V4.090.05N; - (SCD:) GOA AMBioRICS, CoE EF/05/006, (FWO): G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, (ICCoS, ANMMM, MLDM); (IWT): GBOU (McKnow), Eureka-Flite2 IUAP P5/22,PODO-II,FP5-Quprodis; ERNSI; - (JS) and (BDM) are full professosr at K.U.Leuven Belgium, respectively.

References

J.A.K. Suykens, T. van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle. "Least Squares Support Vector Machines", *World Scientific*, Singapore, 2002

J. Zhu, T. Hastie, "Kernel Logistic Regression and the Import Vector Machine"*, Advances in Neural Information Processing Systems*, vol. 14, 2002

Further information

Peter Karsmakers
K.U.Leuven – ESAT/SCD-SISTA
Kasteelpark Arenberg 10
B-3001 Leuven (Heverlee), Belgium
Peter.Karsmakers@esat.kuleuven.be