



Final Assignment

DD3431 Machine Learning (PhD variant)

This final report for the DD3431 Machine Learning course describes the application of Multiclass Linear Support Vector Machines to a multi-class, multi-output classification problem modeling the resource allocation strategies for basestations in a 5G environment.

1 Introduction

2 Description of the data

The data used for this study corresponds to samples of the resource allocation for a communication system composed of exactly three basestations and up to a maximum of 20 user devices. The area of interest studied within this communication scenario is modeled as a map; more specifically, it represents an area of 2400 m² divided into a grid of 2m×2m cells. This grid is encoded as a matrix, with values between 0 and 2 indicating the occupancy status of each cell:

- 0 for empty cells,
- 1 for cells occupied by a user device,
- 2 for cells occupied by a basestation.

The input data for the learning algorithm corresponds then to the vectorized form of this matrix, i.e. a vector of 600¹ elements, each with a value $v_i \in \{0, 1, 2\}$.

On the other hand, the output data for the learning algorithm corresponds to encoded variables representing the basestation–user device association and the resource allocation information for each sample. The exact details of this encoding and the information carried within escape the scope of this report, but it is of importance to note that:

¹ $\frac{2400}{(2 \times 2)} = 600$

- there are 9 of these output variables per input sample (3 for each basestation in the model);
- these output variables are mutually independent;
- they have discrete integer values that range between 0 and 12 288;
- these values can be repeated across variables;
- finally, the order in which these values appear in the output is relevant (i.e. permutations of the same values correspond to different output classes!).

In summary, the structure of an arbitrary sample looks like the following:

$$\underbrace{0, 1, 0, 0, 2, 0, \dots, 0, 0, 0, 0, 1}_{600 \text{ input features}}, \underbrace{4567, 23, \dots, 1337}_{9 \text{ output labels}}$$

3 Choice of classifier

The dataset in question is part of ongoing research at the Department of Information Science and Engineering of the School of Electrical Engineering, and has already been modeled with great success using Random Forests and Neural Networks. The application of Support Vector Machines was then a natural step given the models' popularity in networking research literature; the specific application of the Linear Kernel for SVMs was a result of experimentation with the dataset, where initial experiments exposed the linearly separable nature of the output variables.

Support Vector Machines are binary classifiers though, and thus additional modifications are required to adapt these classifiers to multiclass problems as the one in question. Specifically, literature details the following techniques for adapting binary classifiers to multiclass applications [1, 2]:

- One-vs-One Method
- One-vs-Rest Method
- DAGSVM

4 Preparation of the input data

In total, 72 000 samples in the format detailed in section 2 were provided for the training of the model, with two thirds (48 000) of these used for fitting and the rest (24 000) used for validation purposes. The complete dataset was also used for 10-fold cross-validation.

References

- [1] Chih-Wei Hsu and Chih-Jen Lin. "A comparison of methods for multiclass support vector machines". In: *IEEE transactions on Neural Networks* 13.2 (2002), pp. 415–425.
- [2] D. M. J. Tax and R. P. W. Duin. "Using two-class classifiers for multiclass classification". In: *Object recognition supported by user interaction for service robots*. Vol. 2. 2002, 124–127 vol.2. DOI: 10.1109/ICPR.2002.1048253.