T-61.3050 Data Challenge Submission

Olli Jarva (69124C) & Jarno Rantanen (64913L)

Aalto University School of Science

`olli@jarva.fi` & `jarno@jrw.fi`

1 Solution in a nutshell

We're using Support Vector Machines (SVM's) for classifying the input data (instances of hand-written characters) into 26 classes corresponding to the english alphabet. Images are preprocessed by aligning them to the lower left corner. The SVM uses a Radial Basis Function (RBF) kernel. The optimal kernel function parameters are sought for using a grid-search and validated using k-fold cross-validation. With k=20, we achieved 90.8% classification performance over the training set (n=42152, [1]).

2 Detailed description

SVM's were chosen for their relative ease of use, while still being a robust and well-performing machine learning method for high-dimensional data (and number of samples » features). Implementing an SVM from scratch is relatively error-prone, so we're using the open-source `mlpy` [2] Python library for a ready-made SVM implementation based on `libsvm` [3].

We minimized the variance by aligning images to lower left corner. This resulted approximately 0.3% improvement in classification performance.

The Radial Basis Function (RBF) kernel was chosen simply because our literature review suggested them as a reasonable default choice with SVM [4, 5]. Also having only two parameters to the SVM helps in its training, as the search space is only two-dimensional (as opposed to the 4 dimensions of the polynomial kernel, for example).

Optimal parameters for the SVM (and the kernel function) present a maximization problem in a 2-dimensional search space: γ for kernel and C for SVM penalty parameter. Since an exhaustive search in this space isn't possible, a basic grid search was employed. The performance of the classifier tree in each grid point was evaluated using k-fold cross-validation. Cross-validation was chosen as it provides reasonable protection against overfitting against the training data.

We also tried building a *classifier tree* by having secondary classifiers for commonly misclassified pairs. However, after carefully selecting primary classifier parameters γ and C, single

classifier performed better than the classifier tree.

References

[1] OCR training set. http://users.ics.aalto.fi/kcho/MLBP2012/train.txt , retrieved at 18th Oct 2012. Instructions: https://noppa.aalto.fi/noppa/kurssi/t-61.3050/term_project.

[2] D. Albanese, R. Visintainer, S. Merler, S. Riccadonna, G. Jurman, and C. Furlanello. mlpy: Machine learning python, 2012. http://mlpy.sourceforge.net/ , retrieved at 18th Oct 2012.

[3] C.-C. Chang and C.-J. Lin. Libsvm – a library for support vector machines. http://www.csie.ntu.edu.tw/ cjlin/libsvm/ , retrieved at 18th Oct 2012.

[4] C.-W. Hsu, C.-C. Chang, and C.-J. Lin. A practical guide to support vector classication, 2010. http://www.csie.ntu.edu.tw/ cjlin/papers/guide/guide.pdf , retrieved at 18th Oct 2012.

[5] O. Invanciuc. Applications of support vector machines in chemistry. *Reviews in Computational Chemistry*, 23:291–400, 2007.