Optimal Data Distributions in Machine Learning

Thesis by

Carlos R. González

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

California Institute of Technology

Pasadena, California

2015

(Defended May 22, 2015)

For my little brother Daniel,

thanks for all the joy you bring. May this inspire you!

Acknowledgements

This work would not be possible with the help and support of very important people in my life.

I would like to thank my advisor, Dr. Yaser Abu-Mostafa, for his invaluable guidance, not only in my research but also in so many aspects of life. I am very thankful to have received hints of his wisdom throughout these five years. The road was bumpy in many turns we did not expect, but your perseverance helped me through.

In the same way, I was lucky to have not only one advisor, but two advisors. Thank you so much Dr. Pietro Perona for all your help, both in and outside of academic life. Your never-ending curiosity and interest for so many topics will always be an inspiring example in my research career.

To my family, thank you for your unconditional support and love throughout my *whole* life. The strong values and ethics you instilled in me have payed great dividends. I hope to continue making you proud.

To my friends, thanks for making life happier, especially in the tough times.

Abstract

In the first part of the thesis we explore three fundamental questions that arise naturally when we conceive a machine learning scenario where the training and test distributions can differ. Contrary to conventional wisdom, we show that in fact mismatched training and test distribution can yield better out-of-sample performance. This optimal performance can be obtained by training with the dual distribution. This optimal training distribution depends on the test distribution set by the problem, but not on the target function that we want to learn. We show how to obtain this distribution in both discrete and continuous input spaces, as well as how to approximate it in a practical scenario. Benefits of using this distribution are exemplified in both synthetic and real data sets.

In order to apply the dual distribution in the supervised learning scenario where the training data set is fixed, it is necessary to use weights to make the sample appear as if it came from the dual distribution. We explore the negative effect that weighting a sample can have. The theoretical decomposition of the use of weights regarding its effect on the out-of-sample error is easy to understand but not actionable in practice, as the quantities involved cannot be computed. Hence, we propose the Targeted Weighting algorithm that determines if, for a given set of weights, the out-of-sample performance will improve or not in a practical setting. This is necessary as the setting assumes there are no labeled points distributed according to the test distribution, only unlabeled samples.

Finally, we propose a new class of matching algorithms that can be used to match the training set to a desired distribution, such as the dual distribution (or the test distribution). These algorithms can be applied to very large datasets, and we show how they lead to improved performance in a large real dataset such as the Netflix dataset. Their computational complexity is the main reason for their advantage over previous algorithms proposed in the covariate shift literature.

In the second part of the thesis we apply Machine Learning to the problem of behavior recognition. We develop a specific behavior classifier to study fly aggression, and we develop a system that allows analyzing behavior in videos of animals, with minimal supervision. The system, which we call CUBA (Caltech Unsupervised Behavior Analysis), allows detecting movemes, actions, and stories from time series describing the position of animals in videos. The method summarizes the data, as well as it provides biologists with a mathematical tool to test new hypotheses. Other benefits of CUBA include

finding classifiers for specific behaviors without the need for annotation, as well as providing means to discriminate groups of animals, for example, according to their genetic line.