Conclusion

In the first part of the thesis, we explored 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 in Chapter 2 that in fact mismatched training and test distribution can yield better out-of-sample performance. Thus, the natural step to follow is to find the optimal training distribution for a given test distribution, regardless of the target function that we want to learn. We called this distribution the dual distribution and showed how to obtain it in both discrete and continuous input spaces, as well as how to approximate it in a practical scenario in Chapter 3. Experiments on both synthetic and real data sets showed the benefits of using the dual distribution. Yet, in order to apply the dual distribution in the supervised learning scenario where the training data set is fixed, it was necessary to use weights to make the sample appear as if it came from the dual distribution. We then explored 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 proposed the Targeted Weighting algorithm in Chapter 4, 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. Experiments on real datasets showed the unanimous success of our proposed algorithm. Finally, we proposed a new class of matching algorithms in Chapter 5 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 showed how they led 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 in biological experiments. We developed a wing extension classifier needed to analyze thousands of legacy videos efficiently. These videos were part of an ongoing study of fly aggression. The classifier aided in the investigation that culminated in discovering the neuron and substance that is responsible for fly male aggression. We then moved on to the more complex problem of analyzing behavior using

minimal supervision for humans. To do this, we developed CUBA, which allows detecting movemes, actions, and stories from time series describing the position of animals in videos. The method allows summarizing the data, as well as it provides biologists a mathematical tool to test new hypotheses. When applied to real data, the system also allowed finding classifiers for behaviors such as hopping and freezing in flies without the need for annotation, and it also allowed discriminating groups of flies according to their genetic line.