Chapter 4

EXTENSIONS TO MED

Each problem that I solved became a rule which served afterwards to solve other problems.
René Descartes, 1596-1650

In the previous chapter we saw how the MED approach to learning can combine generative modeling with discriminative methods, such as SVMs. In this chapter we explore extensions of this framework spanning a wide variety of learning scenarios. One resounding theme is the introduction of further (possibly intermediate) variables in the discriminant function \( L(X; \Theta) \), and solving for an augmented distribution \( P(\Theta, \ldots) \) involving these new terms (Figure 4.1). The resulting partition function typically involves more integrals, but as long as it is analytic, the number of Lagrange multipliers and the complexity of the optimization will remain basically unchanged, as when we introduced slack variables in Section 3. Once again, we note that as we add more distributions to the prior, we must be careful to balance their competing goals (i.e. their variances) evenly so that we still derive meaningful information from each component of the aggregate prior (the model prior, the margin prior, and the many further priors we will introduce shortly).

Figure 4.2 depicts the many different scenarios that MED can handle. Some extensions such as multi-class classification can be treated as multiple binary classification constraints [80] or through error-correcting codes [41]. In this chapter we explicate the case where the labels are no longer discrete but continuous, as in regression. Once again (as in binary classification), we find that MED subsumes SVM regression. Following
that, we discuss structure learning (as opposed to parameter estimation), and in particular, feature selection. This leads to the more general problems of kernel selection and meta-learning, where multi-task SVM models share a common feature or kernel selection configuration. We also discuss the use of partially labeled examples and transduction (for both classification and regression). Finally, we arrive at a very important generalization which requires special treatment on its own: the extension to mixture models (i.e. mixtures of the exponential family) and latent modeling (discussed Chapter 5).

1. Multiclass Classification

There are several different approaches to extending binary classification to multi-class problems (for example, error correcting output codes [41]), each with its own benefits and drawbacks.

It is straightforward to perform multi-class discriminative density estimation by adding extra classification constraints. For \( T \) input points, the binary case merely requires \( T \) inequalities of the form: \( y_t \mathcal{L}(X_t; \Theta) - \gamma_t \geq 0 \). In a multi-class setting, constraints are needed based on the pairwise log-likelihood ratios of the generative model of the correct class and that of all the other classes. In other words, in a three-class problem \( (A, B, C) \) with three models \( (\theta_A, \theta_B, \theta_C) \), if \( y_t = A \), the log-likelihood of model \( \theta_A \) must dominate. This leads to the following two classification