Kernelizing Expectation Criteria

Kedar Bellare

May 29, 2009

1 Introduction

I propose a simple extension to the deterministic annealing for semi-supervised kernel machines [1] that has the following properties:

1. Encodes the *clustering assumption* of certain semi-supervised techniques [1, 2]: the assumption that the decision boundary passes through low-density regions.
2. Allows a practitioner to express complex *expectation criteria* similar to recent semi-supervised methods [3, 4, 5, 6].
3. Can be used with different kernel functions.

1.1 Semi-supervised Kernel Methods

We borrow some notation from Sindhwani et. al. [1].

Given l labeled examples $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^{l}$ and u unlabeled examples $\mathcal{D}' = \{\mathbf{x}'_j\}_{j=1}^{u}$, we seek a real-valued function f^* and a labeling $\mathbf{y}'^* = \{y_1'^*, y_2'^*, \ldots, y_u'^*\} \in \{-1, +1\}^u$ for the unlabeled data, by solving:

$$(f^*, \mathbf{y}'^*) = \arg\min_{f, \mathbf{y}'} \frac{1}{2}\|f\|_K^2 + C\sum_{i=1}^{l} L(y_i f(\mathbf{x}_i)) + C'\sum_{j=1}^{u} L(y'_j f(\mathbf{x}'_j))$$

$$\text{subject to: } \frac{1}{u}\sum_{j=1}^{u} \max(0, y'_j) = r \textbf{ (Balanced label constraint)}, \tag{1}$$

where $L(\cdot) : \mathbb{R} \to \mathbb{R}$ is a loss function, $f \in \mathcal{H}_K$ where $\mathcal{H}_K$ is a RKHS of functions with kernel K and C, C' are real-valued parameters that weight the contribution of losses on labeled and unlabeled data respectively. In most applications, $f(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}$ is the dot product between a weight vector $\mathbf{w}$ and the input vector $\mathbf{x}$, and $L(yf(\mathbf{x})) = \max(0, 1 - yf(\mathbf{x}))$ is the hinge loss.

1.2 Deterministic Annealing

Deterministic annealing for semi-supervised kernel machines [1] relaxes the objective above and modifies the optimization problem as:

$$
\begin{aligned}
(f^*, \mathbf{p}^*; T) \quad = \quad & \arg\min_{f,\mathbf{p}} \frac{1}{2}\|f\|_K^2 + C\sum_{i=1}^{l} L(y_i f(\mathbf{x}_i)) \\
& + C' \sum_{j=1}^{u} \Big[p_j L(f(\mathbf{x}'_j)) + (1-p_j) L(-f(\mathbf{x}'_j)) \Big] \\
& + T \sum_{j=1}^{u} \Big[p_j \log(p_j) + (1-p_j)\log(1-p_j) \Big] \qquad (2)
\end{aligned}
$$

where T is the temperature and $\mathbf{p} = (p_1, \ldots, p_u)$ where p_j may be interpreted as the probability that $y'_j = 1$. A higher temperature T smoothes the optimization surface. At a lower temperature T the probability values p_j are peaked at $\{0, 1\}$.

In addition to the objective above, we add a class balancing constraint:

$$
\frac{1}{u}\sum_{j=1}^{u} p_j = r, \qquad (3)
$$

where r is a user-provided parameter.

1.3 Expectation Criteria

Notice that the above label balancing constraint can be viewed as an expectation criterion $E_{p_j}[\phi(\mathbf{x}'_j)] = r$ for the default feature $\phi(\mathbf{x}'_j) = 1, \forall j = 1 \ldots u$. We can generalize this criterion to be $|E_{p_j}[\phi_k(\mathbf{x}'_j)] - r_k| \leq \epsilon_k$, for feature functions $(\phi_k(\mathbf{x}'))_{k=1}^{K}$, $(r_k)_{k=1}^{K}$ are the user-provided target values and ϵ_k is the errors in the measurement of feature $\phi_k(\mathbf{x}')$. Note that $\epsilon = 0$ for the default feature. (See Liang et. al. [6] for more details on how such constraints can be used).

1.4 Alternate Optimization

We can optimize the objective in Equation (2) by alternating between finding function f and probabilities $\mathbf{p}$. The first optimization over f for fixed $\mathbf{p}$ is the same as Sindhwani et. al. [1]. The second optimization over $\mathbf{p}$ (for fixed f) represents a maximum entropy problem given user-provided constraints $|E_{p_j}[\phi_k(\mathbf{x}'_j)] - r_k| \leq \epsilon_k$. Solving for p_j we get:

$$
p_j = \frac{1}{1 + \exp\left(\frac{g_j - \sum_{k=1}^{K} \gamma_k \phi_k(\mathbf{x}'_j)}{T}\right)}, \qquad (4)
$$

for dual parameters $\gamma = (\gamma_k)_{k=1}^K$ and $g_j = C'[L(f(\mathbf{x}'_j)) - L(-f(\mathbf{x}'_j))]$. We can find these dual parameters by optimizing:

$$\gamma^* = \arg\max_{\gamma} \sum_{k=1}^{K} \gamma_k r_k - \sum_{j=1}^{u} \log \left[1 + \exp\left(\frac{\sum_{k=1}^{K} \gamma_k \phi_k(\mathbf{x}'_j) - g_j}{T} \right) \right] - \sum_{k=1}^{K} \epsilon_k |\gamma_k|. \tag{5}$$

References

[1] V. Sindhwani, S. S. Keerthi and O. Chapelle. Deterministic Annealing for Semi-supervised Kernel Machines. *International Conference on Machine Learning (ICML)*, 2006.

[2] T. Joachims. Transductive Inference for Text Classification using Support Vector Machines. *International Conference on Machine Learning (ICML)*, 1999.

[3] J. Graca, K. Ganchev and B. Taskar. Expectation Maximization and Posterior Constraints. *Neural Information Processing Systems (NIPS)*, 2008.

[4] G. Mann and A. McCallum. Generalized Expectation Criteria for Semi-Supervised Learning of Conditional Random Fields. *ACL*, 2008.

[5] K. Bellare, G. Druck and A. McCallum. Alternating Projections for Learning with Expectation Constraints. *Uncertainty in Artificial Intelligence (UAI)*, 2009.

[6] P. Liang, M. I. Jordan and D. Klein. Learning from measurements in exponential families. *International Conference on Machine Learning (ICML)*, 2009.