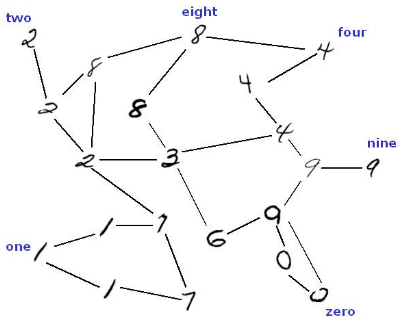

Figure 1. The random fields used in this work are constructed on labeled and unlabeled examples. We form a graph with weighted edges between instances (in this case scanned digits), with labeled data items appearing as special "boundary" points, and unlabeled points as "interior" points. We consider Gaussian random fields on this graph.

show how the extra evidence of class priors can help classification in Section 4. Alternatively, we may combine external classifiers using vertex weights or "assignment costs," as described in Section 5. Encouraging experimental results for synthetic data, digit classification, and text classification tasks are presented in Section 7. One difficulty with the random field approach is that the right choice of graph is often not entirely clear, and it may be desirable to learn it from data. In Section 6 we propose a method for learning these weights by entropy minimization, and show the algorithm's ability to perform feature selection to better characterize the data manifold.

2. Basic Framework

We suppose there are l labeled points $(x_1, y_1), \ldots, (x_l, y_l)$, and u unlabeled points $x_{l+1}, \ldots, x_{l+u}$; typically $l \ll u$. Let $n = l + u$ be the total number of data points. To begin, we assume the labels are binary: $y \in \{0, 1\}$. Consider a connected graph $G = (V, E)$ with nodes V corresponding to the n data points, with nodes $L = \{1, \ldots, l\}$ corresponding to the labeled points with labels $y_1, \ldots, y_l$, and nodes $U = \{l + 1, \ldots, l + u\}$ corresponding to the unlabeled points. Our task is to assign labels to nodes U. We assume an $n \times n$ symmetric weight matrix W on the edges of the graph is given. For example, when $x \in \mathbb{R}^m$, the weight matrix can be

$$w_{ij} = \exp\left(-\sum_{d=1}^{m} \frac{(x_{id} - x_{jd})^2}{\sigma_d^2}\right) \tag{1}$$

where x_{id} is the d-th component of instance x_i represented as a vector $x_i \in \mathbb{R}^m$, and $\sigma_1, \ldots, \sigma_m$ are length scale hyperparameters for each dimension. Thus, nearby points in Euclidean space are assigned large edge weight. Other weightings are possible, of course, and may be more appropriate when x is discrete or symbolic. For our purposes the matrix W fully specifies the data manifold structure (see Figure 1).

Our strategy is to first compute a *real-valued* function $f : V \longrightarrow \mathbb{R}$ on G with certain nice properties, and to then assign labels based on f. We constrain f to take values $f(i) = f_l(i) \equiv y_i$ on the labeled data $i = 1, \ldots, l$. Intuitively, we want unlabeled points that are nearby in the graph to have similar labels. This motivates the choice of the quadratic energy function

$$E(f) = \frac{1}{2} \sum_{i,j} w_{ij} \left(f(i) - f(j)\right)^2 \tag{2}$$

To assign a probability distribution on functions f, we form the Gaussian field $p_\beta(f) = \frac{e^{-\beta E(f)}}{Z_\beta}$, where β is an "inverse temperature" parameter, and Z_β is the partition function $Z_\beta = \int_{f|_L = f_l} \exp\left(-\beta E(f)\right) df$, which normalizes over all functions constrained to f_l on the labeled data.

It is not difficult to show that the minimum energy function $f = \arg\min_{f|_L = f_l} E(f)$ is *harmonic*; namely, it satisfies $\Delta f = 0$ on unlabeled data points U, and is equal to f_l on the labeled data points L. Here Δ is the *combinatorial Laplacian*, given in matrix form as $\Delta = D - W$ where $D = \text{diag}(d_i)$ is the diagonal matrix with entries $d_i = \sum_j w_{ij}$ and $W = [w_{ij}]$ is the weight matrix.

The harmonic property means that the value of f at each unlabeled data point is the average of f at neighboring points:

$$f(j) = \frac{1}{d_j} \sum_{i \sim j} w_{ij} f(i), \text{ for } j = l + 1, \ldots, l + u \tag{3}$$

which is consistent with our prior notion of smoothness of f with respect to the graph. Expressed slightly differently, $f = Pf$, where $P = D^{-1}W$. Because of the maximum principle of harmonic functions (Doyle & Snell, 1984), f is unique and is either a constant or it satisfies $0 < f(j) < 1$ for $j \in U$.

To compute the harmonic solution explicitly in terms of matrix operations, we split the weight matrix W (and similarly D, P) into 4 blocks after the lth row and column:

$$W = \begin{bmatrix} W_{ll} & W_{lu} \\ W_{ul} & W_{uu} \end{bmatrix} \tag{4}$$

Letting $f = \begin{bmatrix} f_l \\ f_u \end{bmatrix}$ where f_u denotes the values on the unlabeled data points, the harmonic solution $\Delta f = 0$ subject to $f|_L = f_l$ is given by

$$f_u = (D_{uu} - W_{uu})^{-1} W_{ul} f_l = (I - P_{uu})^{-1} P_{ul} f_l \tag{5}$$