

Learning piecewise linear classifiers via Dirichlet Process

Xuejun Liao, Ya Xue

April 22, 2005

1 Piecewise linear classifier

A piecewise linear classifier is a classifier that generates a piecewise linear decision boundary. Consider the binary problem where the labeled data $\mathcal{D} = \cup_{j=1}^J \mathcal{D}_j$ and $\mathcal{D}_j = \{(\mathbf{x}_{ij}, y_{ij}) : i = 1, \dots, n_j\}$, $y_{ij} \in \{0, 1\}$. We assume that each \mathcal{D}_j contains at least one positive sample and one negative sample. The decomposition is obtained by a parsing algorithm that will be described later on. Associated with each \mathcal{D}_j there is a linear classifier parameterized by $\mathbf{w}_j \in \mathbb{R}^d$. We assume each \mathcal{D}_j is primitive and the samples in it are always linearly separable, so that there exists \mathbf{w}_j assuring $\mathbf{w}_j^T \mathbf{x}_{ij} \geq 0$ if $y_{ij} = 1$ and $\mathbf{w}_j^T \mathbf{x}_{ij} < 0$ if $y_{ij} = 0$.

It turns out that many of \mathbf{w}_j will become identical to each other, yielding a small number of distinct representative \mathbf{w} 's. Each distinct \mathbf{w} represents a linear boundary that separates a subset of the samples in \mathcal{D} and the ensemble of \mathbf{w} gives us the piecewise linear boundary. In this report we present the method of how to use Dirichlet Process to learn \mathbf{w}_{ij} , and automatically discover the piecewise linear boundary from \mathbf{w}_{ij} .

2 Posterior Distribution of \mathbf{w}_j and z_{ij} Given the Base Measure

Let the base prior measure

$$p(\mathbf{w}_j) = \mathcal{N}(\mathbf{w}_j; 0, \sigma^2 I)$$

and the data likelihood

$$P(y_{ij} = 1 | \mathbf{w}_j) = \int_{-\infty}^{\mathbf{w}_j^T \mathbf{x}_{ij}} \mathcal{N}(\varepsilon; 0, 1) d\varepsilon = \int_{-\infty}^0 \mathcal{N}(z; -\mathbf{w}_j^T \mathbf{x}_{ij}, 1) dz = \int_0^{\infty} \mathcal{N}(z; \mathbf{w}_j^T \mathbf{x}_{ij}, 1) dz$$

Let $z_{ij} \sim \mathcal{N}(z_{ij}; \mathbf{w}_j^T \mathbf{x}_{ij}, 1)$, then

$$P(y_{ij} | z_{ij}) = \mathbf{1}(z_{ij} \geq 0)y_{ij} + \mathbf{1}(z_{ij} < 0)(1 - y_{ij})$$

where $\mathbf{1}(\cdot)$ is the indicator function, and

$$p(z_{ij} | y_{ij}, \mathbf{w}_j) \propto P(y_{ij} | z_{ij})p(z_{ij} | \mathbf{w}_j) = [\mathbf{1}(z_{ij} \geq 0)y_{ij} + \mathbf{1}(z_{ij} < 0)(1 - y_{ij})] \mathcal{N}(z_{ij}; \mathbf{w}_j^T \mathbf{x}_{ij}, 1)$$

or

$$p(z_{ij} | y_{ij}, \mathbf{w}_j) \propto \begin{cases} \mathbf{1}(z_{ij} \geq 0) \mathcal{N}(z_{ij}; \mathbf{w}_j^T \mathbf{x}_{ij}, 1), & \text{if } y_{ij} = 1 \\ \mathbf{1}(z_{ij} < 0) \mathcal{N}(z_{ij}; \mathbf{w}_j^T \mathbf{x}_{ij}, 1), & \text{if } y_{ij} = 0 \end{cases} \quad (1)$$

which shows that $p(z_{ij}|y_{ij}, \mathbf{w}_j)$ is a Gaussian distribution truncated either above zero or below zero, depending on whether $y_{ij} = 1$ or $y_{ij} = 0$.

$$\begin{aligned} p(\mathbf{w}_j|z_{.j}) &= \frac{1}{A_j} \prod_{i=1}^{n_j} p(z_{ij}|\mathbf{w}_j)p(\mathbf{w}_j) = \frac{1}{A_j} \prod_{i=1}^{n_j} \mathcal{N}(z_{ij}; \mathbf{w}_j^T \mathbf{x}_{ij}, 1) \mathcal{N}(\mathbf{w}_j; 0, \sigma^2 I) \\ &= \mathcal{N}(\mathbf{w}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \end{aligned} \quad (2)$$

where

$$A_j = (2\pi)^{d/2} |\boldsymbol{\Sigma}_j|^{1/2} \quad (3)$$

$$\boldsymbol{\Sigma}_j = \sum_{i=1}^{n_j} \mathbf{x}_{ij} \mathbf{x}_{ij}^T + \sigma^2 I$$

$$\boldsymbol{\mu}_j = \boldsymbol{\Sigma}_j^{-1} \sum_{i=1}^{n_j} z_{ij} \mathbf{x}_{ij}$$

3 Dirichlet Process

Let

$$\{\mathbf{w}_j\} \sim DP(g_0, \lambda) \quad (4)$$

where the base measure $g_0 = \mathcal{N}(\mathbf{w}_i; 0, \sigma^2 I)$ and the precision $\lambda > 0$. The Dirichlet process posterior is

$$p(\mathbf{w}_i|\mathbf{w}_{-i}) = \frac{\lambda \mathcal{N}(\mathbf{w}_i; 0, \sigma^2 I) + \sum_{k=1, k \neq i}^N \delta(\mathbf{w}_i - \mathbf{w}_k)}{\lambda + N - 1} \quad (5)$$

$$\begin{aligned} p(z_{.j}, \mathbf{w}_j|\mathbf{w}_{-j}) &= p(\mathbf{w}_j|\mathbf{w}_{-j})p(z_{.j}|\mathbf{w}_j) \\ &= \frac{\lambda \mathcal{N}(\mathbf{w}_j; 0, \sigma^2 I) + \sum_{k=1, k \neq j}^N \delta(\mathbf{w}_j - \mathbf{w}_k)}{\lambda + N - 1} \prod_{i=1}^{n_j} \mathcal{N}(z_{ij}; \mathbf{w}_j^T \mathbf{x}_{ij}, 1) \end{aligned} \quad (6)$$

Other relevant probability distributions are

$$p(z_{.j}|\mathbf{w}_{-j}) = \int p(\mathbf{w}_j|\mathbf{w}_{-j})p(z_{.j}|\mathbf{w}_j)d\mathbf{w}_j = \frac{\lambda A_j + \sum_{j=1, j \neq i}^N \prod_{i=1}^{n_j} \mathcal{N}(z_{ij}; \mathbf{w}_k^T \mathbf{x}_{ij}, 1)}{\lambda + N - 1} \quad (7)$$

$$p(\mathbf{w}_j|\mathbf{w}_{-j}, z_{.j}) = \gamma_0 \mathcal{N}(\mathbf{w}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) + \sum_{k=1, k \neq j}^N \gamma_k \delta(\mathbf{w}_j - \mathbf{w}_k) \quad (8)$$

$$\gamma_0 = \frac{\lambda A_j}{\lambda A_j + \sum_{k=1, k \neq j}^N \prod_{i=1}^{n_j} \mathcal{N}(z_{ij}; \mathbf{w}_k^T \mathbf{x}_{ij}, 1)} \quad (9)$$

$$\gamma_k = \frac{\prod_{i=1}^{n_j} \mathcal{N}(z_{ij}; \mathbf{w}_k^T \mathbf{x}_{ij}, 1)}{\lambda A_j + \sum_{k=1, k \neq j}^N \prod_{i=1}^{n_j} \mathcal{N}(z_{ij}; \mathbf{w}_k^T \mathbf{x}_{ij}, 1)} \quad (10)$$

Thus,

$$\mathbf{w}_j | \mathbf{w}_{-j}, z_{.j}, y \begin{cases} = \mathbf{w}_k, & \text{with probability } \gamma_k \\ \sim \mathcal{N}(\mathbf{w}_j; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), & \text{with probability } \gamma_0 \end{cases} \quad (11)$$

and from (1)

$$p(z_{ij} | y, \mathbf{w}) = p(z_{ij} | y_{ij}, \mathbf{w}_j) \propto \begin{cases} \mathbf{1}(z_{ij} \geq 0) \mathcal{N}(z_{ij}; \mathbf{w}_j^T \mathbf{x}_{ij}, 1), & \text{if } y_{ij} = 1 \\ \mathbf{1}(z_{ij} < 0) \mathcal{N}(z_{ij}; \mathbf{w}_j^T \mathbf{x}_{ij}, 1), & \text{if } y_{ij} = 0 \end{cases} \quad (12)$$

based on (11) and (12) which we perform Gibbs sampling of the posterior $\{\mathbf{w}_j\}, \{z_{ij}\} | \{y_{ij}\}$.

4 Example Results