**Empirical Risk Minimization**

In the case of supervised binary classification, the best classifier $h^*$ minimizes the error rate

$$h^* = \arg \min_{h \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} 1_{y_i \neq h(x_i)}$$

over a sample

$$\mathcal{S} = \{(x_i, y_i)\}_{i=1}^{m} \sim P = X \times Y.$$  

The 0-1 loss is not convex and not differentiable in 0 and has null gradient.

The 0-1 loss is usually replaced by surrogate losses that are:

- convex
- smooth relaxations of the 0-1 loss
- and the surrogate empirical risk is minimized:

$$\mathcal{R}_\phi(\mathcal{S}, h) = \frac{1}{m} \sum_{i=1}^{m} \phi(y_i, h(x_i))$$

with $\phi$ a surrogate loss.

**Empirical Surrogate β-risk Minimization**

Definition:

For any $\mathcal{S}$, $\phi$ and $h$, and for any non-negative real coefficients $\beta^1$ and $\beta^2$ defined for each instance $x_i \in \mathcal{S}$ such that $\beta^1 + \beta^2 = 1$, the empirical surrogate risk $\mathcal{R}_\phi(\mathcal{S}, h)$ can be rewritten as

$$\mathcal{R}_\phi(\mathcal{S}, h) = \mathcal{R}_\phi(\mathcal{S}, h, \beta)$$

where

$$\mathcal{R}_\phi(\mathcal{S}, h, \beta) = \frac{1}{m} \sum_{i=1}^{m} \sum_{t=1}^{(1,t)} \beta_t^1 \phi_t(\sigma(x_i)) + \frac{1}{m} \sum_{i=1}^{m} \beta_t^2 (y_i h(x_i))$$

and the first term is the empirical surrogate β-risk.

**Example: Soft-Margin β-SVM**

Direct generalization of a standard soft-margin SVM:

$$\arg \min_{\beta, \theta, \xi} \frac{1}{2} \|\theta\|_2^2 + c \sum_{i=1}^{m} (\beta_i^1 \xi_i^1 + \beta_i^2 \xi_i^2)$$

s.t. $\sigma(\theta^T x_i + b) \geq 1 - \xi_i$ for $\forall i = 1..m, \sigma \in \{-1,1\}$

$$\xi_i^1, \xi_i^2 \geq 0 \forall i = 1..m, \sigma \in \{-1,1\}$$

where $\theta \in X'$ is the vector defining the margin hyperplane and $b$ its offset, $\mu : X \rightarrow X'$ a mapping function and $c \in R$ a tuned hyper-parameter.

**Application to Semi-Supervised Learning**

Given a set $\mathcal{X}_l$ of labeled instances of size $m_l$ and a set $\mathcal{X}_u$ of unlabeled instances of size $m_u$:

**Iterative Algorithm**

1. initialize $\beta$

$$\forall i = 1..m_l \text{ and } \forall \sigma \in \{-1,1\}, \beta_i^1 = 1 \text{ if } \sigma = y_i, 0 \text{ otherwise}$$

$$\forall i = m_l+1..m_u \text{ and } \forall \sigma \in \{-1,1\}, \beta_i^1 = 0.5$$

2. iteratively learn:

- an optimal separator

$$h^{t+1} = \arg \min_{h \in \mathcal{H}} \sum_{i=1}^{m} \beta_t^1 \xi_t^1 (y_i h(x_i)) + \frac{1}{m} \sum_{i=1}^{m} \phi(y_i, h(x_i)) + \frac{1}{m} \sum_{i=1}^{m} \phi(\sigma(x_i), h(x_i)) + N(h)$$

- $\beta_t^1$ only for $X_l$

$$h^{t+1} = \arg \min_{h \in \mathcal{H}} \sum_{i=1}^{m} \beta_t^1 \phi_t(\sigma(x_i))$$

s.t. $\sum_{i=1}^{m} \phi(y_i h^{t+1}(x_i)) = 0$

$$\beta_t^1 + \beta_t^2 = 1, \beta_t^1 \geq 0, \beta_t^2 \geq 0 \forall i = m_l, m_u$$

**References**