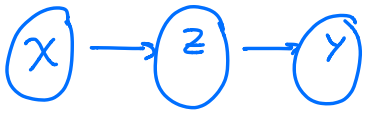
Neural Networks



$z_m = \sigma\{\alpha_{0m} + \alpha_m^T X\}$ for σ activation.

Define $T_k = \beta_{0k} + \beta_k^T z$

$Y = g(T) \quad z = \begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix} \quad k = 1$

$$Y = g\left(\beta_0 + \sum_{m=1}^{m} \beta_m \sigma(\alpha_{0m} + \alpha_m^T x)\right)$$

Fitting Neural Nets

Gradient Descent – training error

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 = f(\theta)$$

Fixing an i. $F_i(\alpha, \beta) = (y_i - f(x_i))^2$

Compute $\nabla_{\alpha,\beta} F_i(\vec{\alpha}, \vec{\beta})$

set $\theta^{(j+1)} = \theta^{(j)} - \gamma^{(j)} \nabla F_i(\alpha, \beta)$ iterate until convergence.

$\gamma^{(j)}$: tunable step size

More specifically

$$\frac{\partial R_i}{\partial \theta} = \begin{bmatrix} \frac{\partial R_i}{\partial \alpha} \\ \frac{\partial R_i}{\partial \beta} \end{bmatrix} = \begin{bmatrix} 2(y_i - f(x_i)) \frac{d}{d\alpha} f(x_i) \\ 2(y_i - f(x_i)) \frac{d}{d\beta} f(x_i) \end{bmatrix}$$

$$= \begin{bmatrix} -2(y_i - g(\beta^T z_i)) g'(\beta^T z_i)(z_i)_m \\ -2(y_i - h(\alpha^T x_i)) h'(\alpha^T x_i)(x_i)_m \end{bmatrix}$$

Define $\delta_i = -2(y_i - g(\beta^T z_i)) \cdot g'(\beta^T z)$

$s_{mi} = -2(y_i - h(\alpha^T x_i)) h'(\alpha^T x_i)$

Forward Pass: Update (α, β) and compute $\hat{f}(x_i)$

Backward Pass: Update gradients (δ_i, s_{mi}).

Regularization:

$$\hat{\theta} = \text{argmin } R(\theta) + \lambda J(\theta)$$

build this into gradient updates.

Support Vector Machines

What if we don't have perfect separation?

$H = \{x : f(x) = x^T\beta + \beta_0 = 0\}$

Classifier: $G(x) = \text{sgn}(x^T\beta + \beta_0)$

When we assumed perfectly separable

$$(\hat{\beta}_0, \hat{\beta}) = \underset{(\beta_0,\beta)}{\text{argmax}} \frac{1}{\|\beta\|} \quad \text{s.t.} \quad y_i(x_i^T\beta + \beta_0) \geq 1$$

Relax the problem so that:

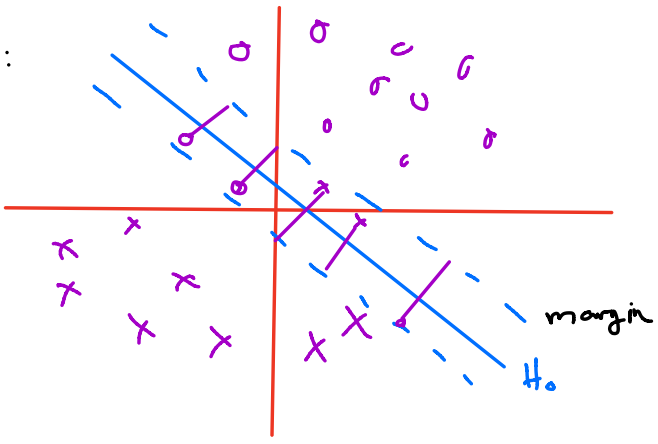
$$\underset{(\beta,\beta_0)}{\text{argmin}} L_P = \underset{(\beta,\beta_0)}{\text{argmin}} \frac{1}{2}\|\beta\|^2 - \underbrace{\sum_i \alpha_i (y_i(x_i^T\beta + \beta_0) - 1)}_{\text{Lagrange multiplier}}$$

Translating into its dual

$$L_D = \underset{\alpha}{\text{argmax}} \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{N} \alpha_i \alpha_k y_i y_k x_i^T x_k$$

s.t. $\alpha_i \geq 0$ KKT: $\alpha_i y_i [x_i^T\beta + \beta_0 - 1] = 0$

Geometrically:



So the penalty L = sum of position of points on incorrect sides of their margin divided by β.

(a) fix margin $\frac{1}{\|\beta\|}$; after that minimize penalty.

Lagragian formulation

$\underset{\beta, \beta_0}{\arg\max} \frac{1}{\|\beta\|}$ $\quad y_i(x_i^T\beta + \beta_0) \geq 1 - \xi_i$ slack variable

$\sum_{i=1}^{n} \xi_i < D$