BackProp específico para clasificación

- Se suele utilizar la función de activación *softmax* en la capa de salida.

- Dado un conjunto de entrenamiento $S = \{(\boldsymbol{x}_1, \boldsymbol{t}_1), \ldots, (\boldsymbol{x}_N, \boldsymbol{t}_N)\}$, con $\boldsymbol{x}_n \in \mathbb{R}^{M_0}$, $\boldsymbol{t}_n \in \{0,1\}^{M_2}$, $(M_2 \equiv C)$, esto permite establecer como criterio de optimización, alternativo al error cuadrático, la entropía cruzada:

$$q_S(\boldsymbol{\Theta}) = -\frac{1}{N}\sum_{n=1}^{N}\sum_{l=1}^{M_2} t_{n,l}\ \log s_l^2(\boldsymbol{x}_n;\boldsymbol{\Theta})$$

- *Problema de entrenamiento*: encontrar $\boldsymbol{\Theta}$ tal que la entropía cruzada sea mínima.

  Solución: DESCENSO POR GRADIENTE:

$$\Delta\theta_{ij}^l = -\rho\,\frac{\partial q_S(\boldsymbol{\Theta})}{\partial\theta_{ij}^l} \qquad 1 \le l \le 2,\ \ 1 \le i \le M_l,\ \ 0 \le j \le M_{l-1}$$

Derivación del BackProp para clasificación

- For $N = 1$, $q_S(\boldsymbol{\Theta}) = -\sum_{l=1}^{M_2} t_l \ \log s_l^2(\boldsymbol{x}; \boldsymbol{\Theta})$
- Actualización de los pesos de la capa de salida $\theta_{i,j}^2$ para una muestra genérica $(\boldsymbol{x}, \boldsymbol{t}) \equiv (\boldsymbol{x}_n, \boldsymbol{t}_n)$

$$q(\boldsymbol{\Theta}) \equiv q_n(\boldsymbol{\Theta}) \ = \ -\sum_{l=1}^{M_2} \left(t_l \log s_l^2 \right); \quad s_l^2 = g\left(\phi_l^2\right); \quad \phi_l^2 = \sum_{m=0}^{M_1} \theta_{lm}^2 s_m^1$$

$$\frac{\partial q}{\partial \theta_{ij}^2} = \frac{\partial q}{\partial s_i^2} \frac{\partial s_i^2}{\partial \theta_{ij}^2} = \frac{\partial q}{\partial s_i^2} \frac{d\, s_i^2}{d\, \phi_i^2} \frac{\partial \phi_i^2}{\partial \theta_{ij}^2}$$

$$\downarrow \quad \downarrow \quad \downarrow$$

$$= -\left(\frac{t_i}{s_i^2} \quad g'(\phi_i^2)\right) \ s_j^1 \stackrel{\text{def}}{=} -\delta_i^2 \ s_j^1$$

$$\frac{\partial q}{\partial \theta_{ij}^2} = -\delta_i^2 \ s_j^1 \,, \qquad \delta_i^2 \stackrel{\text{def}}{=} \frac{t_i}{s_i^2} \ g'(\phi_i^2)$$

$$\Delta\theta_{i,j}^2 = -\rho \ \frac{\partial q_S}{\partial \theta_{i,j}^2} = \rho \ \frac{t_i}{s_i^2} \ f'(z_i^2) \ s_j^1 = \rho \ \delta_i^2 \ s_j^1$$

- Actualización de los pesos de la capa oculta $\theta_{i,j}^1$: idéntica al BackProp para regresión.

Algoritmo BackProp para clasificación

Entrada: Una topología, datos de entrenamiento S, un factor de aprendizaje ρ, pesos iniciales θ_{ij}^l $1 \leq l \leq L, 1 \leq i \leq M_l, 1 \leq j \leq M_{l-1}$ y condiciones de convergencia

Salidas: Pesos de las conexiones que minimizan el error cuadrático medio de S

Método:

Mientras no converja

Para $1 \leq l \leq L, 1 \leq i \leq M_l, 0 \leq j \leq M_{l-1}$, inicializar $\Delta\theta_{ij}^l = 0$ Fin-para

Para toda muestra de aprendizaje $(\boldsymbol{x}, \boldsymbol{t}) \in S$

Desde la capa de entrada a la capa de salida ($l = 1, \ldots, L$):

Para $1 \leq i \leq M_l$: Calcular $\phi_i^l(\boldsymbol{x})$ y $s_i^l(\boldsymbol{x}) = g(\phi_i^l(\boldsymbol{x}))$ Fin-para

Desde la salida a la entrada ($l = L, \ldots, 1$) ,

Para cada nodo ($1 \leq i \leq M_l$)

Calcular $\delta_i^l(\boldsymbol{x}) = \begin{cases} g'(\phi_i^l(\boldsymbol{x})) \dfrac{t_i}{s_i^L(\boldsymbol{x})} & \text{si } l == L \\ g'(\phi_i^l(\boldsymbol{x})) \left(\sum_r \delta_r^{l+1}(\boldsymbol{x})\, \theta_{ri}^{l+1}\right) & \text{en otro caso} \end{cases}$

Para $0 \leq j \leq M_{l-1}$: Calcular: $\Delta\theta_{ij}^l$ += $\delta_i^l(\boldsymbol{x})\, s_j^{l-1}(\boldsymbol{x})$ Fin-para

Fin-para cada nodo

Fin-para cada muestra

Para $1 \leq l \leq L, 1 \leq i \leq M_l, 0 \leq j \leq M_{l-1}$, Actualizar pesos: θ_{ij}^l += $\rho\, \Delta\theta_{ij}^l$ Fin-para

Fin-mientras

Algoritmo BACKPROP (incremental) para clasificación

Entrada: Una topología, datos de entrenamiento S, un factor de aprendizaje ρ, pesos iniciales θ_{ij}^l $1 \leq l \leq L, 1 \leq i \leq M_l, 1 \leq j \leq M_{l-1}$ y condiciones de convergencia

Salidas: Pesos de las conexiones que minimizan el error cuadrático medio de S

Método:

Mientras no converja

Para toda muestra de aprendizaje $(\boldsymbol{x}, \boldsymbol{t}) \in S$

Desde la capa de entrada a la capa de salida ($l = 1, \ldots, L$):

Para $1 \leq i \leq M_l$, calcular $\phi_i^l(\boldsymbol{x})$ y $s_i^l(\boldsymbol{x}) = g(\phi_i^l(\boldsymbol{x}))$ Fin-para

Desde la salida a la entrada ($l = L, \ldots, 1$) ,

Para cada nodo ($1 \leq i \leq M_l$)

Calcular $\delta_i^l(\boldsymbol{x}) = \begin{cases} g'(\phi_i^l(\boldsymbol{x}))\ \dfrac{t_i}{s_i^L(\boldsymbol{x})} & \text{si } l == L \\ g'(\phi_i^l(\boldsymbol{x}))\ (\sum_r \delta_r^{l+1}(\boldsymbol{x})\ \theta_{ri}^{l+1}) & \text{en otro caso} \end{cases}$

Para $0 \leq j \leq M_{l-1}$, calcular: $\Delta\theta_{ij}^l = \delta_i^l(\boldsymbol{x})\ s_j^{l-1}(\boldsymbol{x})$ Fin-para

Fin-para

Para $1 \leq l \leq L, 1 \leq i \leq M_l, 0 \leq j \leq M_{l-1}$: actualizar θ_{ij}^l += $\rho\ \Delta\theta_{ij}^l$ Fin-para

Fin-para

Fin-mientras