UNIVERSIDAD DE SANTIAGO DE CHILE

FACULTAD DE INGENIERIA

DEPARTAMENTO DE INGENIERIA INFORMÁTICA

Propuesta de Tesis

Informe

Nombre: Felipe Alberto Reyes González

Programa: Magíster en Ingeniería Informática

Profesor patrocinante: Victor Parada

Cel.: 890 26 317

email: felipe.reyesg@usach.cl

21 de abril de 2017

Tabla de contenido

1 La retropropagación

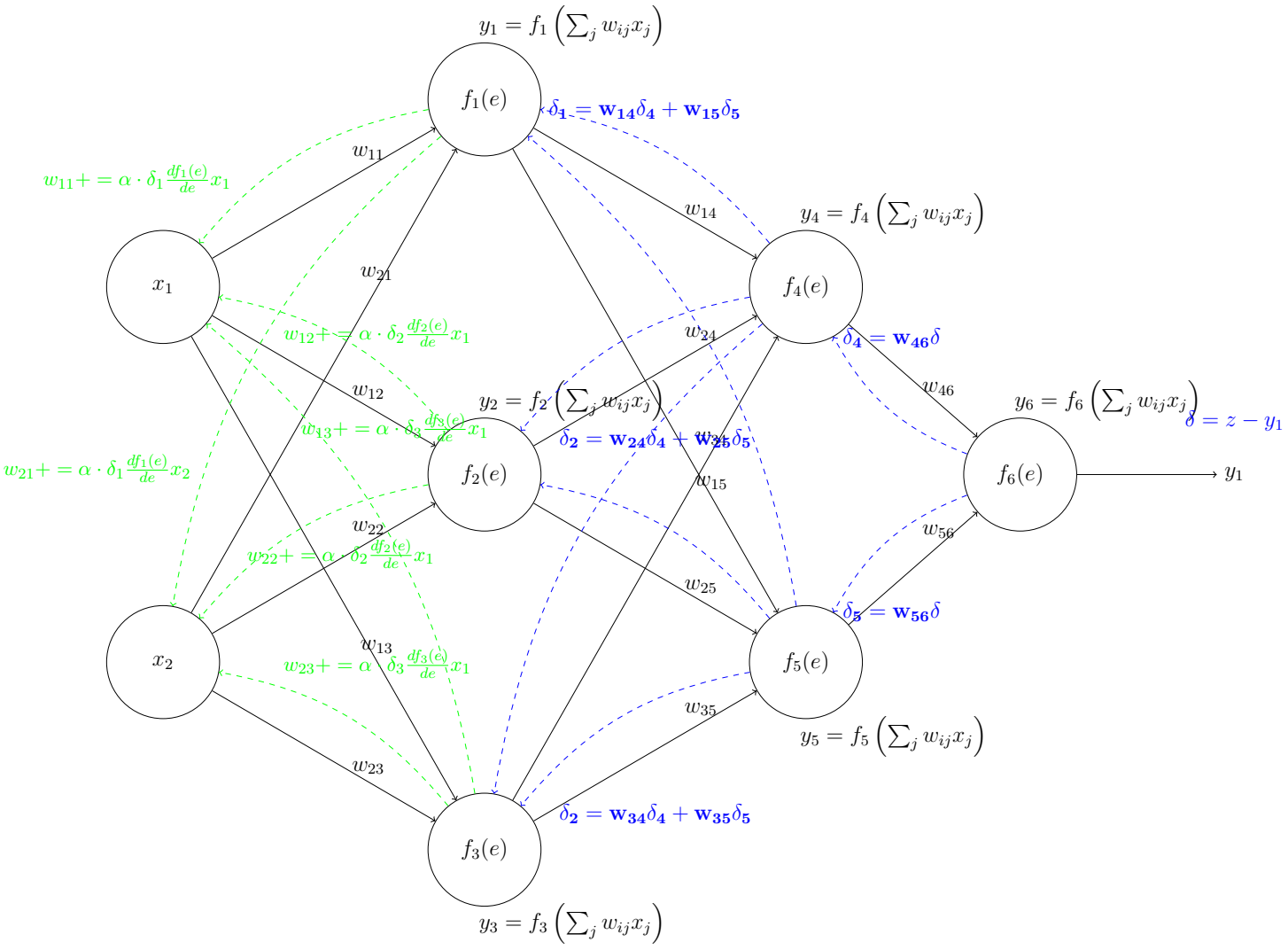

Figura 1: Algoritmo de retropropagación

2 El gradiente descendente

El gradiente descendente busca los punto $p \in \Omega$ donde funciones del tipo $f : \Omega \subseteq \mathbb{R}^m \to \mathbb{R}$ alcanzan su mínimo. La idea de este método se basa en que si f es una función diferenciable en todo su dominio Ω, entonces la derivada de f es un punto $p \in \Omega$ en dirección de un vector unitario $v \in \mathbb{R}^m$ se define como

$$df_p(v) = \nabla f(p)v$$

Observe que la magnitud de la ecuación es

$$|df_p(v)| = ||\nabla f(p)||||v|| \cos\theta = ||\nabla f(p) \cos\theta$$

Dicha magnitud es máxima cuando $\theta = 2n\pi, n \in \mathbb{Z}$. Es decir, para que $|df_d(v)|$ sea máxima, los vectores $\nabla f(p)$ y v debe ser paralelo. De esta manera, la función f crece más rápidamente en la dirección del vector $\nabla f(p)$ y decrece más rápidamente en la dirección del vectro $-\nabla f(p)$. Dicha situación sugiere que la dirección negativa del gradiente $-\nabla f(p)$ es una buena dirección de búsqueda para encontrar el minimizador de la función f.

Sea $f : \Omega \subseteq \mathbb{R} \to \mathbb{R}$, si f tiene un mínimo en p, para encontrar a p se construye una sucesión de punto $\{p_t\}$ tal que p_t converge a p. Para resolver esto, comenzamos en p_t y nos desplazamos una cantidad $-\lambda_t \nabla f(p_t)$ para encontrar el punto p_{t+1} más cercano a p, es decir:

$$p_{t+1} = p_t - \lambda_t \nabla f(p_t)$$

donde λ_t se selecciona de tal manera que $p_{t+1} \in \Omega$ y $f(p_t) \geq f(p_{t+1})$

El parámetro λ_t se seleccionara para maximizar la cantidad a la que decrece la función f en cada paso.

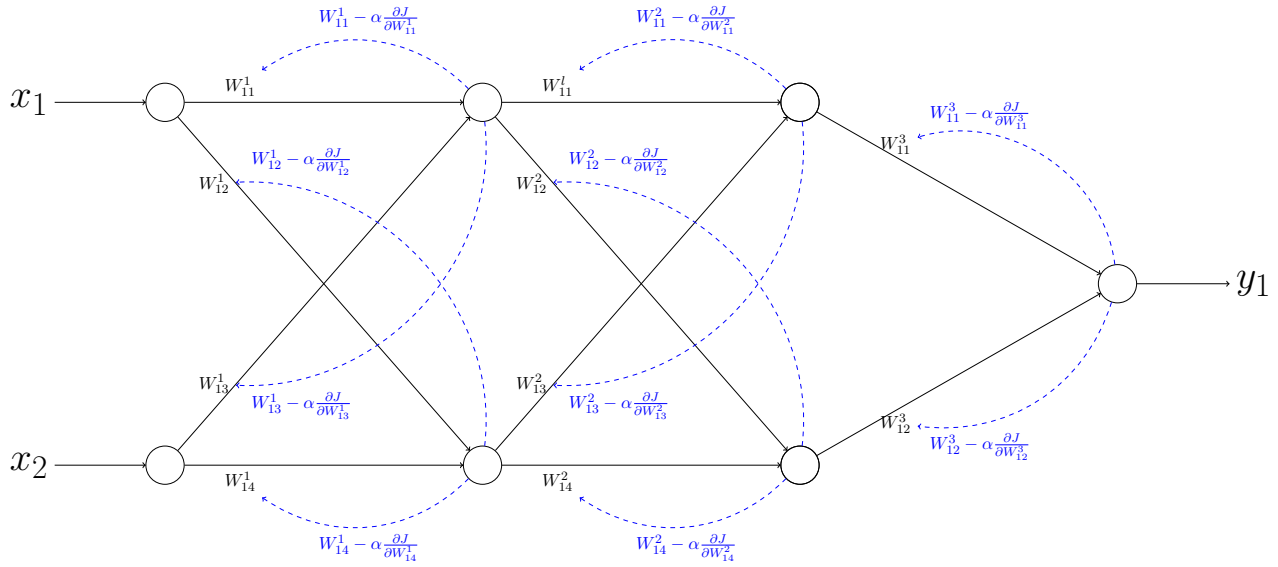

Figura 2: W^1_{ij} es el peso de la n-ésima neurona en la capa $l-1$ a la j-ésima neurona de la capa l de la red.

3 El problema del gradiente descendente

El proceso iterativo que implementan los algoritmos de optimización, lentamente se dirigen hacia un óptimo local, perturbando los pesos en una dirección deducida mediante el uso del gradiente, de tal manera que disminuye a la función de costo. El algoritmo de gradiente descendente actualiza los pesos por el negativo del gradiente ponderado por un valor escalar entre 0 y 1.

Algoritmo 1: Algoritmo del gradiente descendente

Data: C: La función de costo.

Result: Valor óptimo de la función de costo C

initialization;

while $\frac{\partial C}{\partial W^l_{ij}} \to 0$ **do**

  $W^l_{ij} = W^l_{ij} - \alpha\frac{\partial C}{\partial W^l_{ij}}$;

end

Se busca un algoritmo que permita encontrar pesos y sesgos para que la salida de la red aproxime los valores de $y(x)$ a los valores correspondientes con cada entrada x. De esta manera, será posible cuantificar qué tan bien se logra el objetivo mediante la función de costo

$$C(w,b) = \frac{1}{2n}\sum_{x} ||y(x) - a||^2$$

Donde w denota la colección de todos los pesos de la red, b es el sesgo, n

4 El gradiente descendente estocástico

El método del gradiente descendente estocástico (SGD) actualiza los parámetros en cada ejemplo x_i y etiquta y_i de la siguiente manera

$$\theta = \theta - \eta \nabla_\theta$$