Radial basis networks

Víctor Mijangos

Facultad de Ingeniería

UNAM

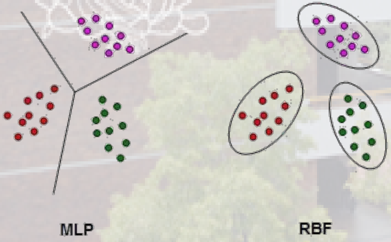
Radial Basis Network

Datos no linealmente separables

La limitación del perceptrón es encontrarse con conjuntos de datos no convexos.

Las redes multicapa pueden lidiar con este problema proyectando los datos a otro espacio. En este espacio, las fronteras de decisión entre clases son lineales.

Pero si en lugar de buscar fronteras de decisión lineales, buscamos que sean **radiales**, podremos encontrar la solución a problemas no lineales.



Interpolación y RBF

Problema de interpolación: Dado un conjunto $\{x : x \in \mathbb{R}^d\}$ y un conjunto finito de números reales $\{a_i : a_i \in \mathbb{R}, i = 1, \ldots, N\}$, el problema de interpolación consiste en escoger una función $\psi : \mathbb{R}^d \to \mathbb{R}$ que cumpla la condición:

$$\psi(x) = a_i$$

El método de **Radial Basis Function** [1] construye un espacio lineal que depende de la posición relativa de los puntos de acuerdo a una métrica. Así:

$$\psi(x) = \sum_{j=1}^{N} \lambda_j \phi(||x - c_j||)$$

Donde $c_j \in \mathbb{R}^d$ son los centroides de la función de base radial.

Radial Basis Function

La **función radial** depende de la **norma** en el espacio $||\cdot||$ y de **centros radiales** c_j.

- Entre más **cercano** esté un punto del centro c_j **mayor** será el valor de ψ.
- Entre más **lejano** esté un punto del centro c_j **menor** será el valor de ψ.

Así, podemos considerar una clasificación en términos de esta función. En el caso **binario**, podemos definirla como:

$$f(x) = \begin{cases} 1 & \text{si } \psi(x) > b \\ 0 & \text{si } \psi(x) \leq b \end{cases}$$

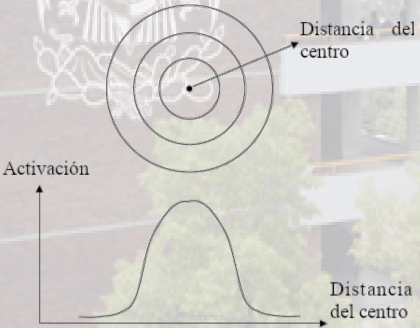
Donde b es un umbral que representa el valor que debe superarse para que un punto se considere dentro de una clase.

Radial Basis Function

En la función radial, λ es una serie de parámetros. La función $\phi(||x - c_j))$ puede determinarse como **una función de probabilidad**:

$$p(x \in \mathcal{N}(c_j)) = exp\{-\frac{1}{2}||x - c_j||^2\}$$

Esta función describe una especie de distribución **normal**, donnde c_j es la media.



Radial Basis Network

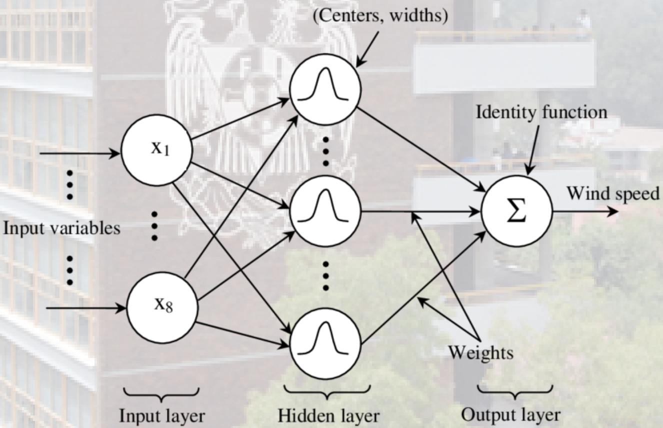
Ahora, bien, bajo esta perspectiva probabilística, la función radial queda determinada como sigue:

$$\psi(x) = \sum_{j=1}^{k} \lambda_j exp\{-\frac{1}{2}||x - c_j||^2\}$$

Sustituyendo λ_j por w_j y junto con la función de decisión que hemos definido, esta función puede verse como una red con una capa oculta. A este tipo de redes se les conoce como **Radial Basis Networks** (RBN).

Radial Basis Network

La red consta de una capa oculta **probabilística**, mientras que la capa de salida puede verse como un **perceptrón**.



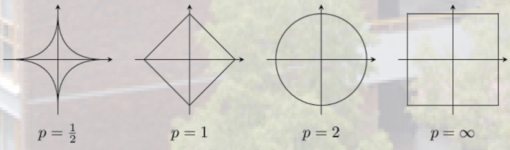
Métricas en RBN

El modelo de RBN depende de una **métrica euclideana** en base a la que se determina la cercania o lejanía del centro.

El concepto de métrica se puede ampliar. Una generalización común son las métricas basadas en normas L^p, determinadas por:

$$||x - y||_p = \left(\sum_{i=1}^{d} |x_i - y_i|^p\right)^{\frac{1}{p}}$$

Con $1 \leq p \leq \infty$.



Aprendizaje en las RBNs

Estimación de parámetros

Las RBN son capaces de solucionar problemas no linealmente separables, como el problema XOR.

Sin embargo, a diferencia de las redes FeedForward, la capa oculta determina k distribuciones de probabilidad.

Entonces, además de estimar los pesos de las conexiones (en la salida) se deben estimar los parámetros de las distribuciones.

Capa oculta

La capa oculta representa el mayor problema. Podemos pensarla **independientemente** de la capa de salida.

Así, el problema se vuelve en estimar los parámetros $C = \{c_1, ..., c_k\}$, las **medias** de las distribuciones, que se acoplen a la distribución.

Para esto, es común utilizar un algoritmo de **clustering**. Específicamente k-means [2].

Función de riesgo en capa oculta

Dado que la función en la capa oculta está definida como una distribución, podemos ver que la función a minimizar está dada por:

$$\begin{aligned}
\hat{C} &= \arg\min_{C} -\frac{1}{N}\sum_{x}\sum_{j=1}^{k} \ln \phi(||x - c_j||) \\
&= \arg\min_{C} -\frac{1}{N}\sum_{x}\sum_{j=1}^{k} \ln exp\{-\frac{1}{2}||x - c_j||^2\} \\
&= \arg\min_{C} \frac{1}{2N}\sum_{x}\sum_{j=1}^{k} ||x - c_j||^2
\end{aligned}$$

Estimación de parámetros en la capa oculta

El problema así definido se puede interpretar como **minimización de la varianza** para cada distribución.
Tenemos que:

$$\nabla_{c_j} \frac{1}{2N} \sum_{x} \sum_{j=1}^{k} ||x - c_j||^2 = \frac{1}{N} \sum_{x} x - c_j$$

E iguland a cero el gradiennte obtenemos que:

$$c_j = \frac{1}{N} \sum_{x} x$$

Esto es **la media empírica**. Sin embargo, ya que existen varias distribuciones, esta media es únicamente entre los x más cercanos a c_j.

Método de k means

El problema se reduce a **estimar** un conjunto de **distribución normal** $N(\mu_j, \sigma_j^2), j = 1, 2, \ldots, k$ con parámetros:

$$\mu = c_j$$
$$\sigma^2 = \sum_x ||x - c_j||^2$$

Este problema de minimización puede resolverse por medio del algoritmo de *k*-means.

Si bien *k*-means se ve como un algoritmo de clustering, los centroides de este algoritmmo definen una media de la **distribución de un cluster**.

Algoritmo de k means

Inicialización: Se inicializan los centroides $C = \{c_1, c_2, \ldots, c_k\}$ a leatoriamente.

Inducción: Para cada ejemplo x se calcula:

$$d(x, c_j) = ||x - c_j||$$

Se estiman los clusters como:

$$clust(x) = \arg\min_j d(x, c_j)$$

Se mueven los centroides con respecto a la siguiente regla:

$$c_j \leftarrow \frac{1}{N} \sum_{clust(x)=j} x$$

Finalización: Se finaliza cuando no se hagan nuevas asignaciones de los centroides.

Obtención de la capa oculta

La capa oculta de la RBN estará definida por los parámetros $\hat{C} = \{c_1, ..., c_k\}$ estimados por el algoritmo de k-means.

La capa oculta en esta red está definida por:

$$h_j = exp\{-\frac{1}{2}||x - c_j||^2\}$$

Es decir, depende de c_j y se puede interpretar como una función probabilística.

La probabilidad de un dato dependerá de la cercanía a c_j. Así el radio de c_j define el valor de la unidad oculta.

Cálculo de la capa de sálida

Para obtener los valores de capa de salida se determina la función:

$$\psi(x) = \sum_{i=1}^{k} w_i h_i$$

donde w_i son pesos en la capa de salida. Así, puede definirse la función de clasificación como:

$$f(x) = \begin{cases} 1 & \text{si } \psi(x) > 0 \\ 0 & \text{si } \psi(x) \leq 0 \end{cases}$$

Los pesos de la capa de salida pueden estimarse por medio del algoritmo del perceptrón como:

$$w_i \leftarrow w_i - \eta(\phi(x) - y)h_i$$

References

David S Broomhead and David Lowe.
Radial basis functions, multi-variable functional interpolation and adaptive networks.
Technical report, Royal Signals and Radar Establishment Malvern (United Kingdom), 1988.

James MacQueen et al.
Some methods for classification and analysis of multivariate observations.
In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

The End