Módulo 4:
Mejorando la Exactitud

Carlos López Vázquez

carlos.lopez@ieee.org

Plan

- Introducción
- Revisión de herramientas estadísticas
- Detectando problemas
- Imputando valores ausentes
- Ejemplos

Condicionantes…

Éxito depende de:

- ➤Disponibilidad de Datos
- ➤Disponibilidad de Modelos
- ➤Sensibilidad de los Modelos
- ➤Capacitación de técnicos
- ➤Calidad de Datos
- ➤Otros

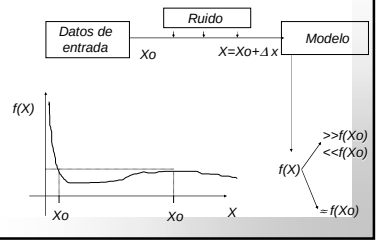
Disponibilidad de Datos

- ¡Siempre limitada!
  - ➤Atributos
  - ➤Resolución espacial
  - ➤Vigencia
  - ➤Niveles de Error
  - ➤Otros… ($, inexistencia)
- Normalmente ¡condicionan al modelo!

Disponibilidad de Modelos

- Modelo no es lo mismo que Realidad
- Siempre imperfectos
  - ➤Quizá "importados" de USA, etc. ☺
  - ➤Suelen faltar datos
    - – Datos sustitutivos (más otros modelos…)
  - ➤Poco plazo, poco presupuesto…
- Usualmente no validados
- Códigos complejos (CPU, disco, etc.)

Sensibilidad del Modelo



Sensibilidad del Modelo(2)

- Es específica al conjunto {Modelo,Datos}
- Un problema en sí mismo
  - ➤¿Qué tipo de *errores? ¿Cuántos?¿Dónde?*
  - ➤Enfoque Determinístico
  - ➤Enfoque Estocástico
- Ejs.: Viewshed area (Fisher)
- Ejs.: Goodchild para líneas

Capacitación de los técnicos

Idealmente deberían:

- Conocer del problema "físico"
- Conocer de los datos (propios y ajenos!)
- Conocer los modelos
- Capaces de criticar resultados

¡Es mucho conocer!

Condicionantes…

Éxito depende de:

- ✔Disponibilidad de Datos
- ✔Disponibilidad de Modelos
- ✔Sensibilidad de los Modelos
- ✔Capacitación de técnicos
- ➤Calidad de Datos
- ➤Otros

Calidad de Datos

- Completitud
- Exactitud
- Vigencia
- Linaje

Si no son "apropiados":

- Buscar fuentes alternativas
- Arremangarse…
  - Mejorar Exactitud
  - Cambiar de Modelo

Dos actores…

- Usuario:
  - Tomador de Datos
  - Sufridor de Consecuencias
  - +productos, con -fondos
- Productor:
  - Receptor de Críticas
  - Usualmente monopólico
  - +productos, con –fondos

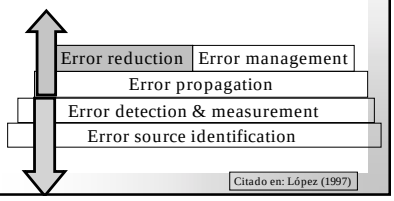
¡Qué pesimista!

Dos actores… (versión optimista)

- Usuario:
  - Especifica requerimientos
  - Preocupado por la Exactitud
  - No tiene acceso fluido a "la verdad"
  - Llevará la Culpa…
- Productor:
  - Observa estándares
  - Preocupado por la Exactitud
  - La "verdad" existe, pero es más cara
  - Llevará la Culpa…

¡Un problema común!

Una jerarquía de necesidades…



Citado en: López (1997)

[illegible]

Fuera de discusión ($, plazo de entrega, etc.)

¿Problema para algún PhD?

- [illegible] las relaciones cuantitativas
- Carencia de datos apropiados e independientes para validar
- Conocimiento insuficiente de la sensibilidad del modelo
- ¿Dónde están los outliers que importan?
- ¿Cómo imputar los valores ausentes?

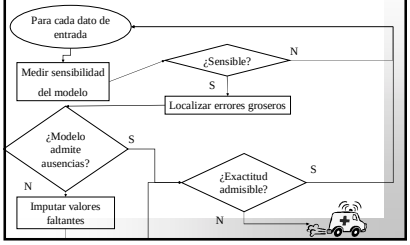
lo posible…

El proceso requeriría…

- Evaluar la sensibilidad del modelo
- Localizar errores groseros (outliers)
- Asignar valores apropiados para los outliers y/o los faltantes

¡Casi nada de ello está previsto en un GIS corriente!

Diagrama de decisión



Análisis de Sensibilidad

- No pueden analizarse *todos* los modelos
- Ejemplos:
  - Modelo hidrológico de una cuenca
    - Inputs: lluvia diaria, caudal en ríos, uso del terreno, geología, etc.
    - Outputs: caudal en ríos, niveles en las presas
  - Modelo de contaminación de aire
    - Inputs: inventario de emisores, viento de superficie, MDE, etc.
    - Outputs: mapas de niveles de inmisión

¿Sensibilidad…?¿Cómo?

- Propagación analítica
  - Taylor
  - Aritmética de Intervalos
- Monte Carlo

Temas:

- ¿Generación de "errores"?
  - Tamaño, localización, correlación…
- ¿Generación de outliers?
  - ¿PDF?, ¿modelo del error?

Fuente: Burrough & McDonnell (1998)

Expansión de Taylor

- Nos restringimos a modelos que son función del punto x (para facilitar, *think raster*)
  - ➤Excluímos buffers, ventanas, topologías, etc.
- Si el modelo puede ponerse como

  $U=g(A_1,A_2,\ldots,A_n)$

siendo A_i atributo cuantitativo sujeto a error
- Se conocen $\langle A_i \rangle$ y $\mathrm{var}(A_i)$; también $\mathrm{var}(A_i,A_j)$
- Si g(.) es lineal, entonces es fácil

Propagación en el Caso lineal

Si $U = g(A_i) = \sum_{i=1}^{n} b_i A_i$

y los A_i no están correlacionados, entonces

$$\langle U \rangle = \sum_{i=1}^{n} b_i \langle A_i \rangle$$

y $$\mathrm{var}(U) = \sum_{i=1}^{n} b_i^2 \,\mathrm{var}(A_i)$$

Si hay correlación, entonces

$$\mathrm{var}(U) = \sum_{i=1}^{n}\sum_{j=1}^{n} b_i b_j \sqrt{\mathrm{var}(A_i)\,\mathrm{var}(A_j)}\,\rho(i,j)$$

Caso más general

- Linealiza la función g(.)
- Taylor al primer orden

$$\delta U = \sum_{i=1}^{n} \frac{\partial g}{\partial A_i} \delta A_i + \ldots$$

- ¡Equivale a una función g lineal! ➔caso conocido
- Algunos autores llegan hasta segundo orden
- O dicen que llegan…☺

Pros y Contras

- Ventajas:
  - ➤Es una fórmula analítica
  - ➤Eficaz en términos de CPU
  - ➤Maneja correlación espacial
- Problemas: se trata de una *aproximación*
  - ➤¿Será buena? ¿mala?
  - ➤¿De dónde saco las derivadas parciales?
  - ➤Es fácil si hay normalidad $N(0,\sigma)$
    - En algunos casos el error no tiene media cero
  - ➤¿Cómo estimar la correlación espacial del error?

Cálculo de derivadas parciales

- A mano, cualquiera *podría*
  - ➤Sólo modelos chicos, relativamente simples
- Soluciones de hoy
  - ➤Álgebra simbólica (Maple, Derive, etc.)
  - ➤Procesadores de Código fuente
    - ADOL-C / ADOL-F
    - Tapenade
  - ➤Sobrecarga de operadores
    - Matlab+ADMAT
    - C++, F90, etc.

Eso no es todo…

- En general, los errores son *función de punto* y no constantes espaciales
  - ➤Ej.: interpolación
- Eso afecta a la estimación de δA_i
- El procedimiento estándar es Kriging
  - ➤¡Pero Kriging no genera outliers!
- ¿Cómo generar errores groseros?
  - ➤Yet to be solved…

Fuente: B. Schneider

Aritmética de Intervalos

- También analítico
- Equivalente a un "peor caso"
- Notación: Si $\underline{a_i} \le A_i \le \underline{A_i}$➔ $[A_i]=[\underline{a_i}, \underline{A_i}]$
- Ej.:
  - ➤Suma: $S=A_i+B_i$; $[S]=[\underline{a_i}+\underline{b_i}, \underline{A_i}+\underline{B_i}]$
  - ➤Producto: $P=A_i*B_i$;
    $[P]=[\min(\underline{a_i}\underline{b_i}, \underline{a_i}\underline{B_i}, \underline{A_i}\underline{b_i}, \underline{A_i}\underline{B_i}), \max(\text{ídem})]$
- Automatizable
  - ➤C++, F90, etc.

Pros y Contras

- Cotas exactas y estrictas
  - ➤Quizá inalcanzables…
  - ➤*Estricto* es quizá requerido en algunos casos
- Eficaz en tiempo de CPU
- No requiere normalidad (ignora PDF)
- No require diferenciabilidad
- Problemas:
  - ➤No provee PDF del intervalo
  - ➤No maneja correlación espacial

Método de Monte Carlo

- Monte Carlo ←→ azar (!)
- Enfoque estadístico, no determinístico
- Idea: repita para k=1,N

  Generar realizaciones A_i, i=1,m

  Calcule y guarde $U_k=g(A_i)$
- Luego procese los U_k generados, calculando media, varianza, etc.
- La gracia es que $\mathrm{var}(U_k) \sim 1/\mathrm{sqrt}(N)$

Detalles…

- ¿Cómo *generar realizaciones*?
  - ➤Asumir independencia espacial
    - – Normal, media μ y varianza σ
    - – demasiado fácil… y no realista
  - ➤Modelar correlación espacial
    - – No es simple; normalmente ¡hay que adivinarla!
      - Error reportado como RMS, percentil 90, etc.
      - Nada de localización espacial
    - – Krigeado: simulación condicional
- Nada de esto es trivial…

Más detalles…

- Método de MC es *CPU intensive*
  - ➤Hoy día hay CPU… y antes no
  - ➤La CPU no es el mayor problema
- La función g(.) no se aproxima; se la usa directamente
- La distribución de U_k se estima mejor
- MC puede mejorarse con *bootstrapping*

El proceso requeriría…

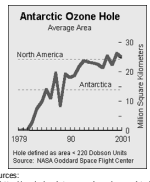
- ✔Evaluar la sensibilidad del modelo
- Localizar errores groseros (outliers)
- Asignar valores apropiados para los outliers y/o los faltantes

¿Qué es un *outlier*?

- Hay varias definiciones algo ambiguas
- *Un outlier es un valor que produce resultados inusuales (de baja probabilidad) al aplicarle cierto modelo conceptual*
  - ➤Ej.: test de normalidad
- Suele traducirse como dato *aberrante*
- No requiere la existencia de un *valor verdadero*

¿Detección automática de Outliers?

- La Historia del agujero de Ozono
- En 1985 Farman, Gardinar y Shanklin estaban confundidos al analizar registros tomados por la misión Británica en la Antártida mostrando que los niveles de ozono habían bajado 10%
- ¿Porqué el satélite Nimbus 7, equipado con instrumentos específicos para registrar niveles de ozono no había registrado ese descenso tan pronunciado?
- ¡Las concentraciones de ozono registradas por el satélite eran tan bajas que fueron tratadas como outliers y descartadas por un programa!



Sources:
http://exploringdata.cqu.edu.au/ozone.html
http://www.epa.gov/ozone/science/hole/size.html

Algunos detalles…

- ¿Quién dice que es un outlier?
- En ocasiones no está claro
  - ➤Dicotómico (ej.: digitado desde papel)
  - ➤[mal, quizá mal, no sé, quizá bien, bien]
  - ➤Lógica borrosa (*fuzzy*)
- Literatura estadística
  - ➤Conjuntos pequeños
  - ➤Errores sintéticos
  - ➤Cálculos pesados

Más detalles…

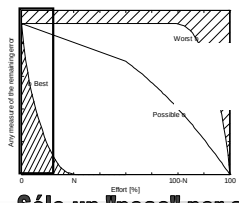
- ¿Qué método usar para detectar?
  - ➤Requiere definir relación "*mejor que*"
  - ➤Podría automatizarse
- Casos analizados
  - ➤Dicotómicos
  - ➤Inspector "*perfecto*"

Tipos de errores

- Error Tipo I: Dato clasificado como erróneo siendo correcto
- Error Tipo II: Dato clasificado como correcto siendo erróneo

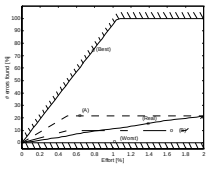
- Ventajas: *el tamaño no importa*
- Desventaja: *el tamaño podría importar*
- Se necesitarán otros estimadores

El proceso de detección



Sólo un "poco" por ciento

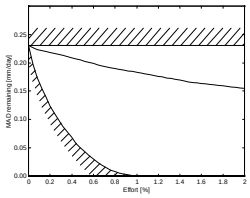
¿Cómo comparar métodos?



$$e_I = 1 - \frac{df}{dx}\frac{N}{100}$$

$$e_{II} = \left(\frac{100-f}{100}\right)\frac{N}{100}$$

Cuando el tamaño importa…

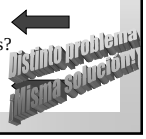


¿Cómo sería esto *automático*?

- Dados: tipología de datos, valores, un generador de errores y N métodos
  - repetir muchas veces
    - Generar simulación de errores y contaminar banco
    - Para i=1,N
      - Aplicar método i
      - Contabilizar estadísticos de éxito
  - hasta lograr estabilidad estadística
  - Elegir el método con mejor resultado
- Implementado en Matlab
- Quizá costoso (CPU)

Un segundo problema…

- ¿Qué hacer con los datos erróneos?
  - *Digitar de nuevo*
  - *Ir al campo a observar nuevamente*
  - *Resignarse*
  - *Eliminarlos*
  - *Sustituirlos*
- ¿Qué hacer con los faltantes?
  - *Ignorarlos*
  - *Sustituirlos*



¿Cómo sería esto automático?

- Dados: tipología de datos, valores, un generador de huecos y N métodos
  - repetir muchas veces
    - Generar simulación de huecos y modificar el banco
    - Para i=1,N
      - Aplicar método i
      - Contabilizar estadísticos de éxito
  - hasta lograr estabilidad estadística
  - Elegir el método con mejor resultado
- Quizá costoso (CPU)
- Implementado

Imputar es más simple…

- Imputar es un problema más clásico
  - Interpolación
  - Vecino más cercano
  - Etc.
- Varias funciones ya disponibles en GIS
- Sólo hay que simular ausencias
  - ¡No es trivial!
  - ¿Al azar?,¿en rachas?, etc.

¿Cómo comparar métodos?

- Midiendo discrepancia contra el valor conocido
- Hay un Método Óptimo
- No hay un Peor Método
- Se usan sustitutos *Naive*
  - *"El dato de ayer"*
  - *"El dato de al lado" "El más próximo"*
  - *El promedio espacial, la moda, etc.*
- Tema recurrente en la literatura
- Probablemente siga siéndolo…

Plan

- ✔ Introducción
- Revisión de herramientas estadísticas
- Detectando problemas
- Imputando valores ausentes
- Ejemplos

Herramientas estadísticas

- Seguro que ya las conocen
- Necesario refrescar un poco
- Al menos algo…☺
- Univariada
- Multivariada
- Componentes Principales
- Y además…
  - Redes Neuronales
  - Krigeado

Algo básico…

- La **función de distribución** F(x) de una variable aleatoria X se define como:

$$F(x) = PROB(X \le x)$$

- X se dice **discreta** si

$$PROB(X = x_i) = a_i \ge 0 \forall i \quad \text{y además} \quad \sum_i a_i = 1$$

- X se dice **continua** si PROB(X=x)=0
- La función de densidad de probabilidad f(t) está definida por

$$F(x) = \int_{-\infty}^{x} f(t)dt$$

Esperanza matemática…

- Se define como:
  - **Caso discreto** $\mu = E(x) = \sum_i x_i PROB(X = x_i)$
  - **Caso continuo** $\mu = E(x) = \int_{-\infty}^{+\infty} t.f(t)dt$
- También llamada *media*
- *Valor modal* o *moda*: x | f(x) es máxima
- *Mediana*: x | F(x)=0.5
- *Percentil* p: x | F(x)=p
- *Varianza* $\sigma^2 = E\left((X-\mu)^2\right) = \int_{-\infty}^{+\infty}(t-\mu)^2 f(t)dt$

Exactos, pero desconocidos

El amigo Gauss…

- La distribución Normal $N(\mu, \sigma^2)$

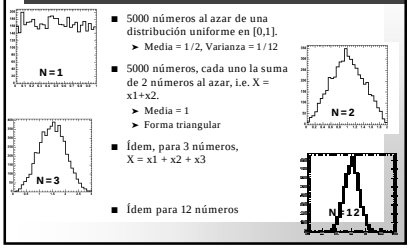
$$f(t) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-(t-\mu)^2/(2\sigma)\right] \quad t \in (-\infty, +\infty)$$

- La "versión estándar" es N(0,1)
- Teo. Central del Límite

$$y_1, y_2, ..., y_n \qquad \mu_i = E(y_i) \qquad \sigma_i^2 = E\left((y_i - \mu_i)^2\right)$$

$$Y = y_1 + y_2 + ... + y_n \qquad z = \frac{Y - \sum_{i=1}^{n} y_i}{\sqrt{\sum_{i=1}^{n} \sigma_i^2}} \quad n \to +\infty \quad \sim N(0,1)$$

Ilustración del Teo Central



- 5000 números al azar de una distribución uniforme en [0,1].
  - Media = 1/2, Varianza = 1/12
- 5000 números, cada uno la suma de 2 números al azar, i.e. X = x1+x2.
  - Media = 1
  - Forma triangular
- Ídem, para 3 números, X = x1 + x2 + x3
- Ídem para 12 números

Caso típico 1: datos iid.

i.e. $\mu_i = \mu$, $\sigma_i = \sigma$ para todo *i*

Teorema Central del Límite

$$\langle X \rangle = \sum_i \mu_i = N\mu \quad \Rightarrow \quad \langle \bar{x} \rangle = \frac{X}{N} = \mu$$

$$V(\bar{x}) = \sum_i V_i(\bar{x}) = \frac{1}{N^2}\sum_i V_i(X) = \frac{N\sigma^2}{N^2} = \frac{\sigma^2}{N}$$

$$\sigma(\bar{x}) = \frac{\sigma}{\sqrt{N}}$$ ← **Famosa ley de la raiz (N)**

Caso típico 2: igual media

i.e. $\mu_i = \mu$ para todo *i*

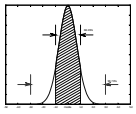
$$\bar{x} = \frac{\sum_i x_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}$$ **Promedio ponderado**

$$V(\bar{x}) = \frac{1}{\sum_i 1/\sigma_i^2} \Rightarrow \sigma^2(\bar{x}) = \frac{1}{\sum_i \frac{1}{\sigma_i^2}}$$

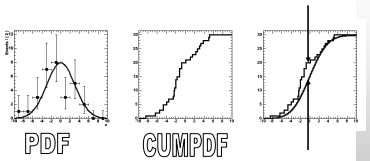
Fórmula del 'inverso de suma de inversos' para la varianza

La aplicación práctica…

- Dado el banco
  - Confirmar que es normal
    - Varios tests disponibles
    - Ej.: Test de Kolmogorov-Smirnov
  - Estimar la media
  - Estimar la varianza
- Dado un dato
  - Calcular anomalía $|x - \bar{x}|/s$
  - Comparar contra tabla



Kolmogorov-Smirnov



Bondad de ajuste: $d = \sqrt{N} \cdot \max|\text{cum}(x) - \text{cum}(p)|$

¡*No funciona para multivariado*!

Más formalmente...Test de Grubbs

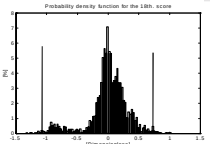
- Asumiendo datos normales
- Detectar un outlier por vez, removerlo, y repetir
  - H0: No hay outliers en los datos
  - HA: Hay al menos un outlier
- Estadístico de Grubbs $G = \frac{\max|X - \bar{X}|}{s}$
- Rechazar H0 si: $G > \frac{(N-1)}{\sqrt{N}} \sqrt{\frac{t^2_{(\alpha/N, N-2)}}{N-2+t^2_{(\alpha/N, N-2)}}}$

En general…

- *Estimar* los percentiles p y 1-p
  Criterio de López (¡¡!!):
- Si x está en [p,1-p]➔correcto
- Si no, x es *outlier*

También Rousseeuw (1991)

- Habría que considerar n, casos multimodales, etc.



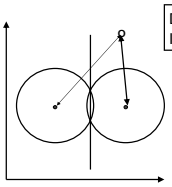
Multivariada…

- Ahora es con vectores $\underline{x}$…
- $\underline{\mu}=E(\underline{\mathbf{x}})$ es un vector
- σ^2 es ahora una matriz de covarianza **C**
- Anomalía era el escalar $(1/\sigma^2)*(x-\mu)^2$
- Ahora será $d^2=(\underline{x}-\underline{\mu})^T*C^{-1}*(\underline{x}-\underline{\mu})$, también escalar
  - ➤Se denomina *Distancia de Mahalanobis*

¿Porqué tan complicado?

Caso isotrópico

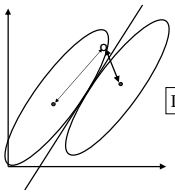
Distributiones Gaussianas Isotrópicas (igual varianza)



Mínima distancia euclídea

Fuente: Mahesan Niranjan

La distancia euclídea no siempre…



Distancia de Mahalanobis

$d^2=(\underline{x}-\underline{\mu})^T*C^{-1}*(\underline{x}-\underline{\mu})$

Análisis de Componentes Principales

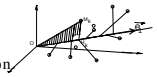
- Técnica corriente y popular
- Dada una tabla de m filas y n columnas, se "comprime" en otra de m filas y p columnas, p<n y en muchos casos p<<n
- Compresión *con pérdida*
- Se usa para reducir dimensionalidad del problema, conservando lo esencial de la varianza
  - ➤Imágenes multiespectrales
  - ➤Datos meteorológicos

ACP(2)

- Ilustración en R^3 para M_k-O
- Busco $\underline{e}_1$ tal que

$$\sum_{i=1}^{n} \overline{M_k H_k}^2$$

sea mínima



- Luego se repite en R^2 con (M_k-H_k), encontrándose $\underline{e}_2$
- En general hay n direcciones, ortogonales entre sí

ACP(3)

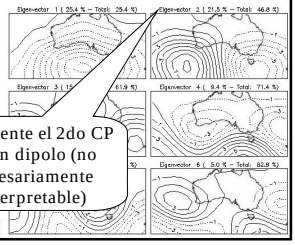
- Las proyecciones OH_k se denominan *scores*
- Hay n scores por cada fila de la tabla
- La gracia está en que…
- Se demuestra que los $\underline{e}_i$ son los Vectores Propios de C, matriz de Covarianza
- Los Valores Propios son proporcionales a la varianza de los scores
  - ➤VP pequeños ←→*scores* pequeños←→se desprecian
- Las series de los *scores* son no-correlacionadas

ACP(4)

- Las Componentes Principales son los $\underline{e}_i$
- También conocidas como *Empirical Orthogonal Functions* (EOF)
- Ampliamente utilizadas en Ciencias de la Tierra
- Suelen tener interpretación individual
- Pero tienen algunos problemillas…

Ej. Meteorológico

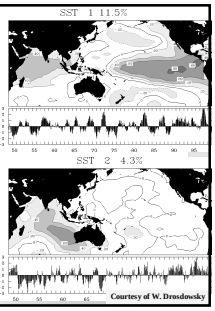
Fuente: Dr. Bertrand Timbal



Típicamente el 2do CP es un dipolo (no necesariamente interpretable)

Rotación de CP

- Facilitate physical interpretation
- Review by Richman (1986) and by Jolliffe (1989, 2002)
- New set of variable: RPCs
- Varimax is a very classic rotation technique (many others)



First two rotated PCAs of Indian/Pacific SSTAs using data from Jan 1949 to Dec 1991.

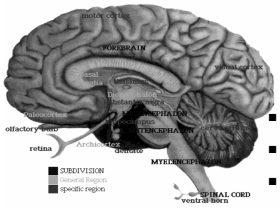
Algo más que sólo Estadística…

- Presentaremos ahora algo de Redes Neuronales Artificiales
- Será mencionado en Detección de outliers y en Regresión

¿Qué es una red neuronal?

- Es un modelo matemático que tiene un vago parecido con las neuronas biológicas
- La neurona es la unidad básica. En ella se distinguen las conexiones sinápticas, las dendritas (muchas), y el axón (único)
- Muchas neuronas fuertemente conectadas forman una red

Cerebro humano…



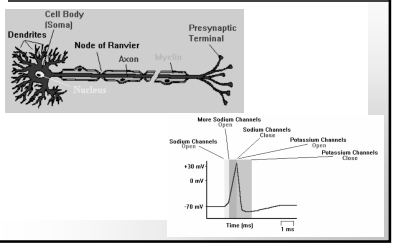
- 10^{11} Neuronas (procesadores)
- Poder desconocido
- 1000 – 10000 conexiones por neurona

Fuente: Dr. Juan José Flores Romero

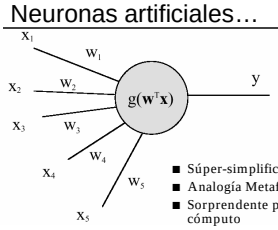
¿Cómo funciona una neurona?

- A través de las (muchas) dendritas llega la información al núcleo de la neurona
- Estimulado por esta información, se produce un efecto transmitido vía el axón
- Las conexiones sinápticas vinculan al axón con otras dendritas de otras neuronas, formando así la red

Neuronas biológicas…



Neuronas artificiales…

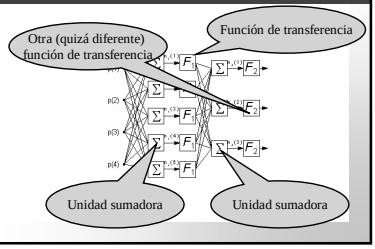


- Súper-simplificación
- Analogía Metafórica
- Sorprendente poder de cómputo

Las redes neuronales artificiales

- Simulan muy crudamente sólo algunos aspectos sustanciales de las biológicas
- La topología se modela satisfactoriamente
- Las conexiones sinápticas se modelan con coeficientes de ponderación
- La relación causa-efecto del núcleo es simulada con una función cualquiera

Esquema de una red neuronal

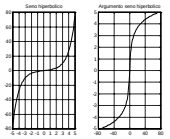


Función de transferencia (ejs.)

- Logsig

$$output_j = \frac{1}{1+\exp\left[-\sum_i a_{ij} * input_i\right]}$$

- Senh



Algunos aspectos interesantes

- Dependiendo de la aplicación, se eligen diferentes arquitecturas de redes
- Las redes pueden utilizarse para predecir un número (output continuo), identificar una letra (output categorizado), etc.
- Toda red requiere de un "entrenamiento"
- Si la función de transferencia es no lineal, la red también lo será

Entrenamiento...

- Función objetivo (caso de regresión):

$$EMC(w) = \frac{1}{N}\sum_{i=0}^{N}(a_i - RN_i(w))^2$$

- Son los conocidos mínimos cuadrados (no lineales…)

Algunos términos…

- Aprendizaje
- Algoritmo de entrenamiento
- *Training set/Test set*
- Generalización
- *Overfitting*

Curso intensivo de Krigeado

- Es un método de Interpolación
- Lo hemos citado y lo citaremos en:
  - Imputación de ausencias (obvio…)
  - Detección de errores
  - Estimación de sensibilidad de modelos
- Base estadística
- Incorporado en algunos GIS (¿malamente? ¿parcialmente?...)