Recuperación de Información Multimedia

Repaso Machine Learning

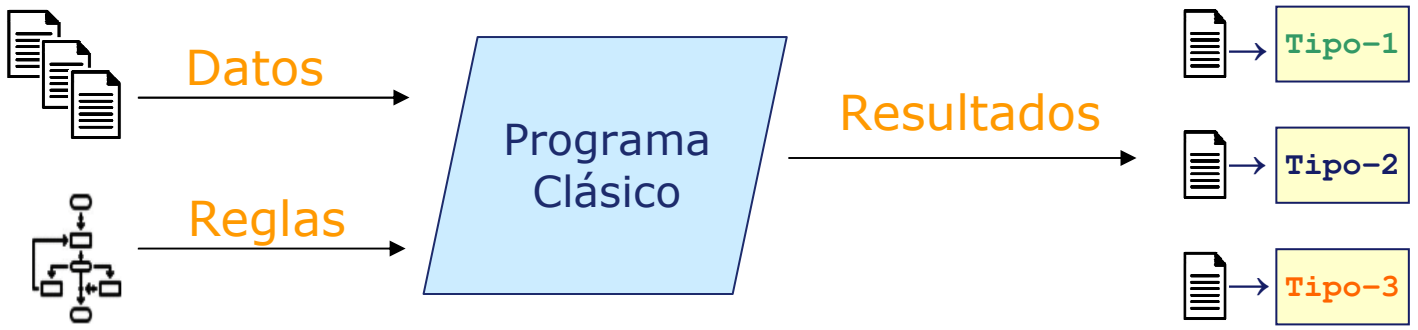
CC5213 – Recuperación de Información Multimedia
Departamento de Ciencias de la Computación
Universidad de Chile
Juan Manuel Barrios – https://juan.cl/mir/ – 2020

Inteligencia Artificial

- Se refiere a los esfuerzos por “automatizar tareas intelectuales normalmente realizadas por humanos”
- Enfoques:
  - Neurobiología
  - Symbolic AI (lógica+reglas)
  - Machine Learning (datos+estadística)

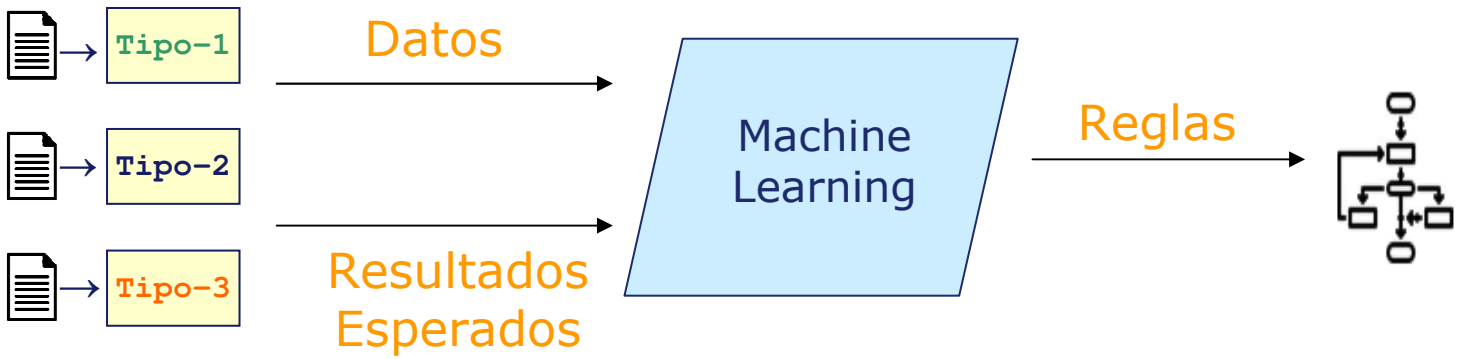
Programación Clásica

- Aplicar reglas sobre datos para generar una salida deseada



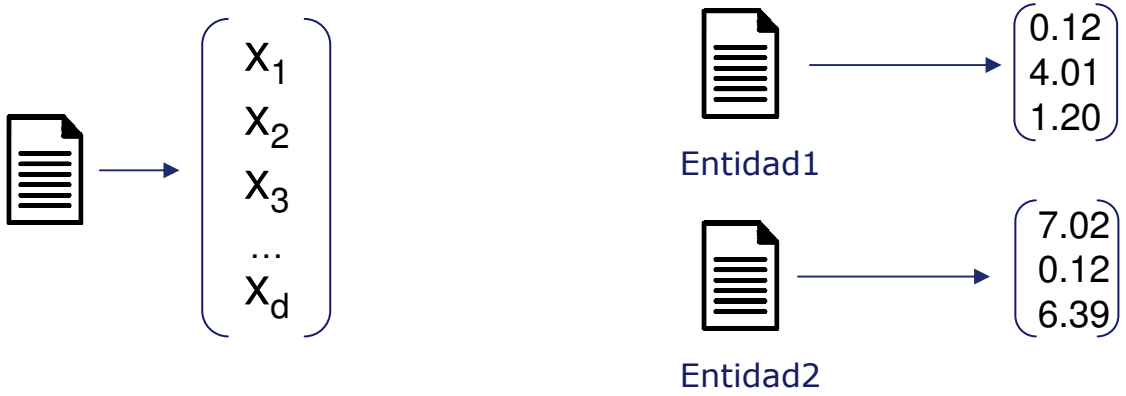
Machine Learning

- Aprender o descubrir las reglas que permiten producir el resultado deseado



Representación del Contenido

- Se tiene un conjunto de **entidades** que se desea analizar. Ej: clientes, productos, e-mails, páginas web, fotos, canciones, etc.
- Cada **entidad** se debe representar por sus datos, usualmente modelados como un vector de dimensión fija (vector en R^n)
  - Feature Vector, Vector Característico, Descriptor



Machine Learning

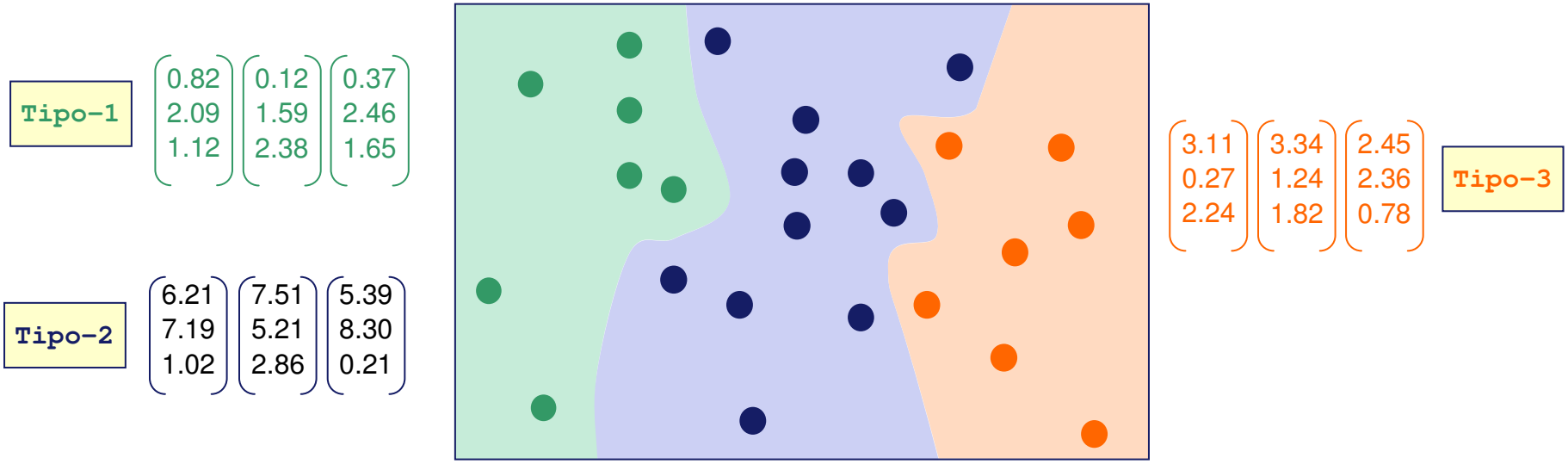
- Se tienen entidades (representadas por sus descriptores) y se desea encontrar reglas o patrones
- Métodos Supervisados:
  - Se tienen datos etiquetados (entidades con la respuesta correcta)
  - Tareas comunes: Asignar etiquetas (clasificación, caso discreto). Generar uno o más números (regresión, caso continuo).
- Métodos No Supervisados:
  - No hay información extra (a parte de los descriptores)
  - Tarea común: Encontrar grupos de descriptores similares e interpretar esos grupos (clustering).

Clasificadores

- Se tiene un conjunto clases o etiquetas $C=\{C_1,..,C_n\}$
- Se tiene uno o más descriptores de ejemplo para cada clase C_i
- Se desea asignar la etiqueta que le corresponde a entidades nuevas (ej. etiquetar tipos de correos, etiquetar tipos de clientes, etc.)
- La etapa de entrenamiento consiste en encontrar el patrón de cada clase
- La etapa de clasificación consiste en etiquetar descriptores nuevos

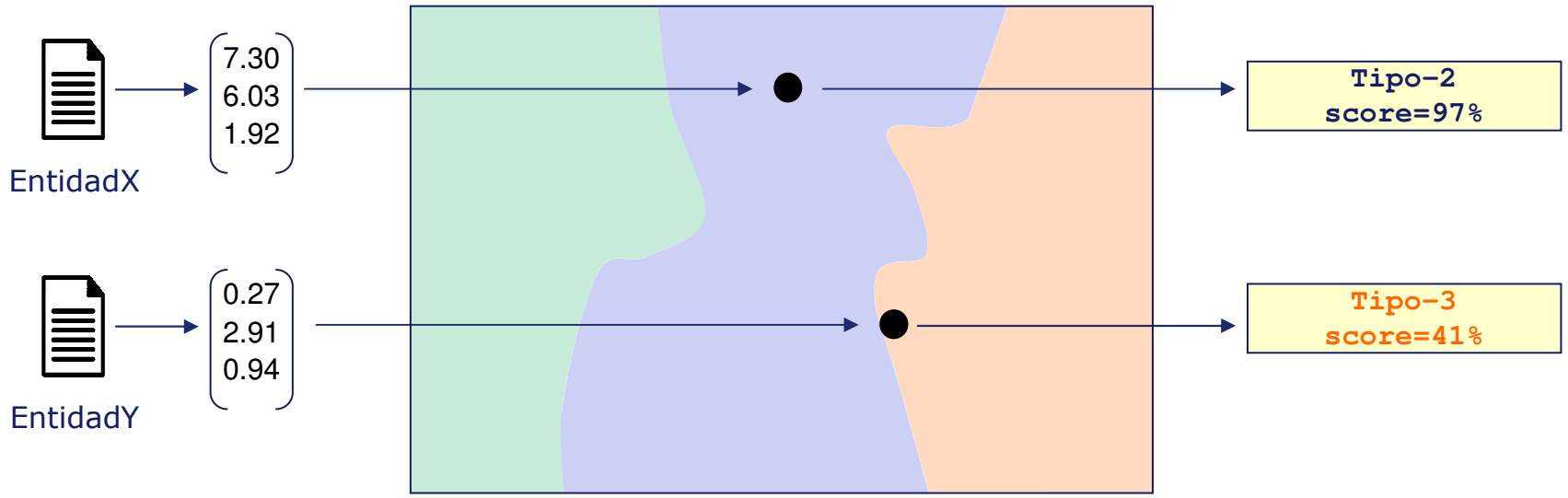
Entrenamiento de Clasificador

- Se tienen descriptores de las entidades y sus etiquetas
- Se desea determinar una función *F* que asocie cada descriptor con su etiqueta $F: R^n \rightarrow \{\text{etiqueta}_1, \ldots, \text{etiqueta}_n\}$
  - *F* busca regiones de R^n con descriptores de una misma etiqueta
  - *F* determina las fronteras de separación entre esas regiones



Uso de Clasificador

- Clasificación de una **entidad nueva**:
  - Obtener o calcular el descriptor que corresponde a esa entidad
  - Utilizar el clasificador ya entrenado para determinar la **región** a la que pertenece el descriptor
  - Obtener la **etiqueta** asociada a esa zona junto con un **valor de confianza** de que sea la etiqueta correcta



Entrenamiento y Underfit

- Los clasificadores se ajustan (entrenan) automáticamente mediante un algoritmo que usa todos los datos de entrenamiento
- Los clasificadores usualmente tienen además parámetros de mayor nivel (hiperparámetros) que se deben ajustar manualmente
  - Se debe probar varias veces con distintos hiperparámetros hasta lograr que el entrenamiento logre un buen resultado de clasificación
- **Underfit** ocurre cuando no se logra un buen resultado de clasificación sobre los datos de entrenamiento
  - Los datos son muy difíciles de agrupar en regiones
  - Se deben ajustar los hiperparámetros para aumentar el “poder expresivo” del clasificador

Datos de Producción

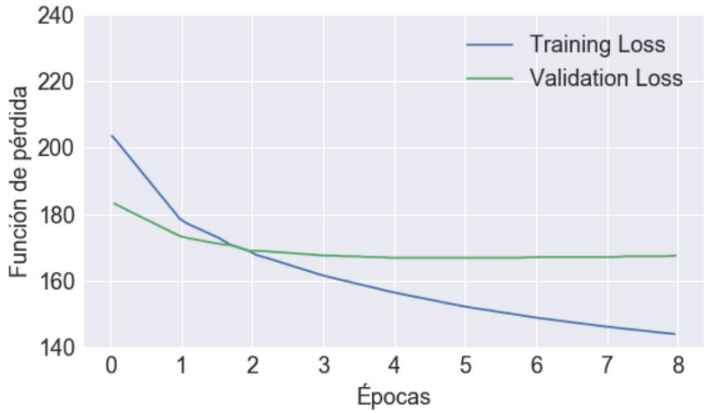
- Usualmente se desea entrenar un clasificador para etiquetar datos nuevos o desconocidos (producción)
  - Los datos usados para el entrenamiento deben ser similares a los que después se usarán en producción
- **Overfit (1)** ocurre cuando se obtiene una gran diferencia en los resultados de clasificación entre los datos de **entrenamiento** y los de **producción**
  - El clasificador funciona bien con datos conocidos, pero obtiene malos resultados en producción al clasificar datos nuevos
  - **Problema grave**, el proyecto tiene altas probabilidades de fracasar
  - Para evitar esto se separa una fracción de los datos etiquetados para estimar el resultado que se obtendrá con datos futuros (conjunto de test)
  - Notar que si el conjunto de test influye en el modelo final entonces hay riesgo de Overfit (1)

Datos de Test

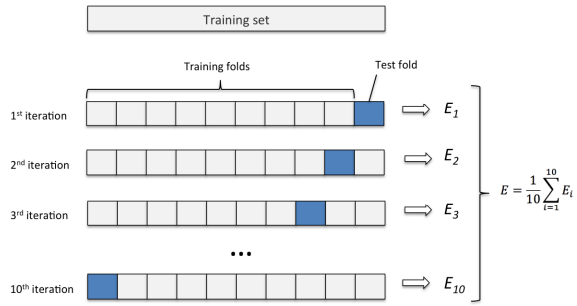
- Los datos de test son una muestra de los datos de producción
  - Permiten estimar el resultado que obtendrá el clasificador al instalarlo en producción y prevenir un Overfit (1)
- **Overfit (2)** ocurre cuando se obtiene una gran diferencia en los resultados de clasificación entre los datos de **entrenamiento** y los de **test**
  - El algoritmo de entrenamiento está memorizando los datos de entrenamiento y no encuentra un patrón general en los datos
  - Si se prueban muchos clasificadores (ajustar hiperparámetros) y se escoge el que maximiza el resultado en test entonces se arriesga un Overfit (1)
  - Separar de los datos de entrenamiento un subconjunto llamado **validación**
  - El conjunto de validación se usa en la fase de entrenamiento de un clasificador, por tanto influye en el modelo final

Datos de Validación

- El conjunto de validación permite probar distintos hiperparámetros y escoger el que logre los que logren los mejores resultados
  - Se considera parte de los datos usados para entrenar
- **Overfit (3)** ocurre cuando hay una gran diferencia en resultados de clasificación entre los datos de **entrenamiento** y los de **validación**
  - Detener el entrenamiento por épocas: cuando el resultado en entrenamiento mejora y en validación comienza a empeorar



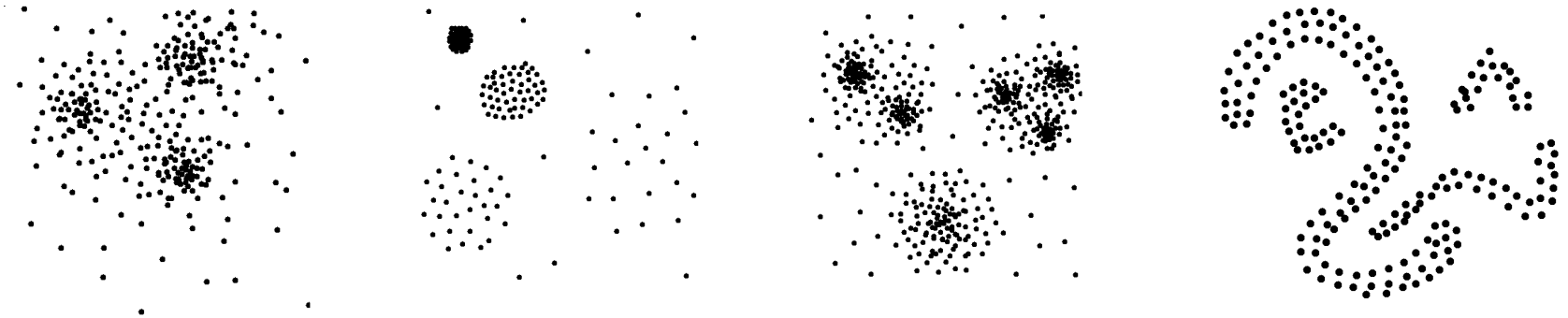
Validación Cruzada



- Si existen muy pocos datos etiquetados se hace difícil separar un conjunto de test adecuado
  - Una división fija puede producir un conjunto de test pequeño que **no representa bien los datos futuros** de producción
- **Validación cruzada (cross validation)** consiste en realizar múltiples experimentos separando entrenamiento/test en forma aleatoria
- **K-Fold** divide los datos en K subconjuntos y realiza K veces: entrenar con (K-1) subconjuntos y calcular el score de clasificación sobre el subconjunto restante
  - El promedio de los K scores corresponde al score de test (i.e. un estimador del score que logrará en datos futuros)
    - Si se escoge el modelo de máximo score se necesitará un test extra
- **Validación anidada (nested validation)** al entrenar con los (K-1) conjuntos se usa un nuevo **K'-Fold** (K' puede ser distinto de K) para separar conjuntos de entrenamiento/validación y se escoge el modelo de máximo score en validación

Clustering

- Definir un conjunto finito de clusters (grupos, categorías o clases) de la colección de datos tales que:
  - Objetos en el *mismo* cluster deben ser lo más similares posible
  - Objetos en *diferentes* clusters deben ser lo más disímiles posible



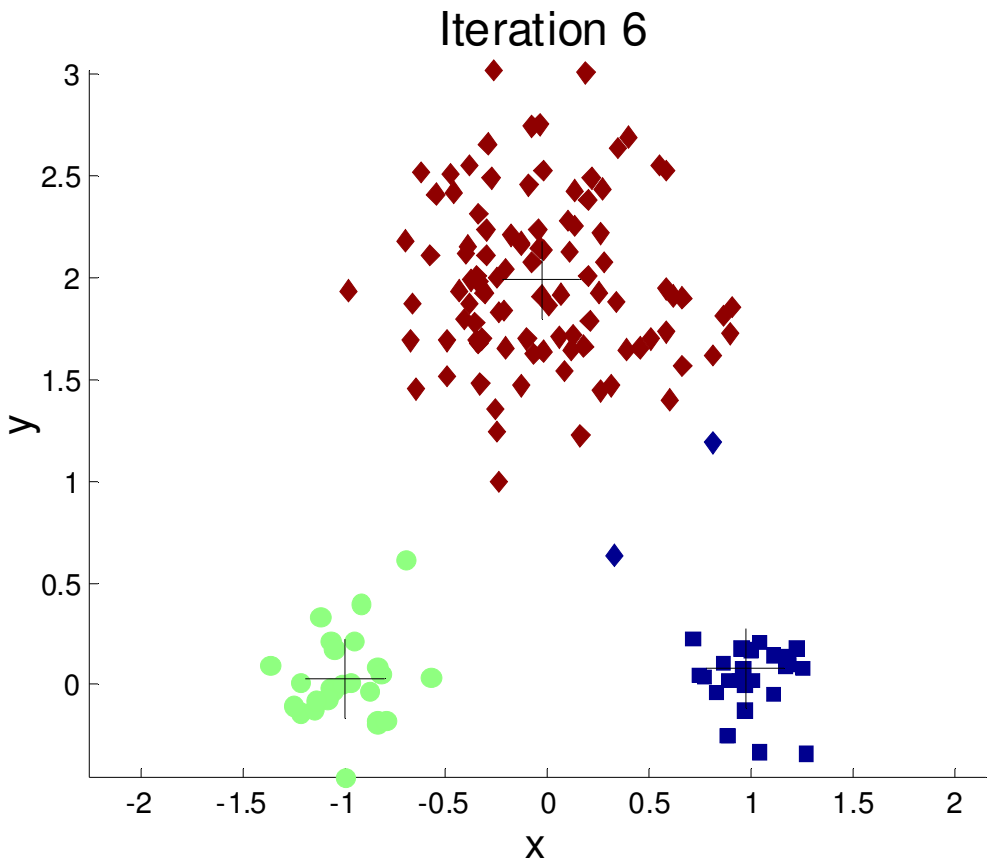
- **Propiedades de los cluster:**
  - clusters pueden tener tamaños, formas y densidades diferentes
  - clusters pueden formar una jerarquía
  - clusters pueden traslaparse o ser disjuntos

K-means

- Cada cluster se asocia con un centroide (punto central)
- Cada punto es asignado al cluster con el centroide más cercano
- El número de clusters K debe ser especificado

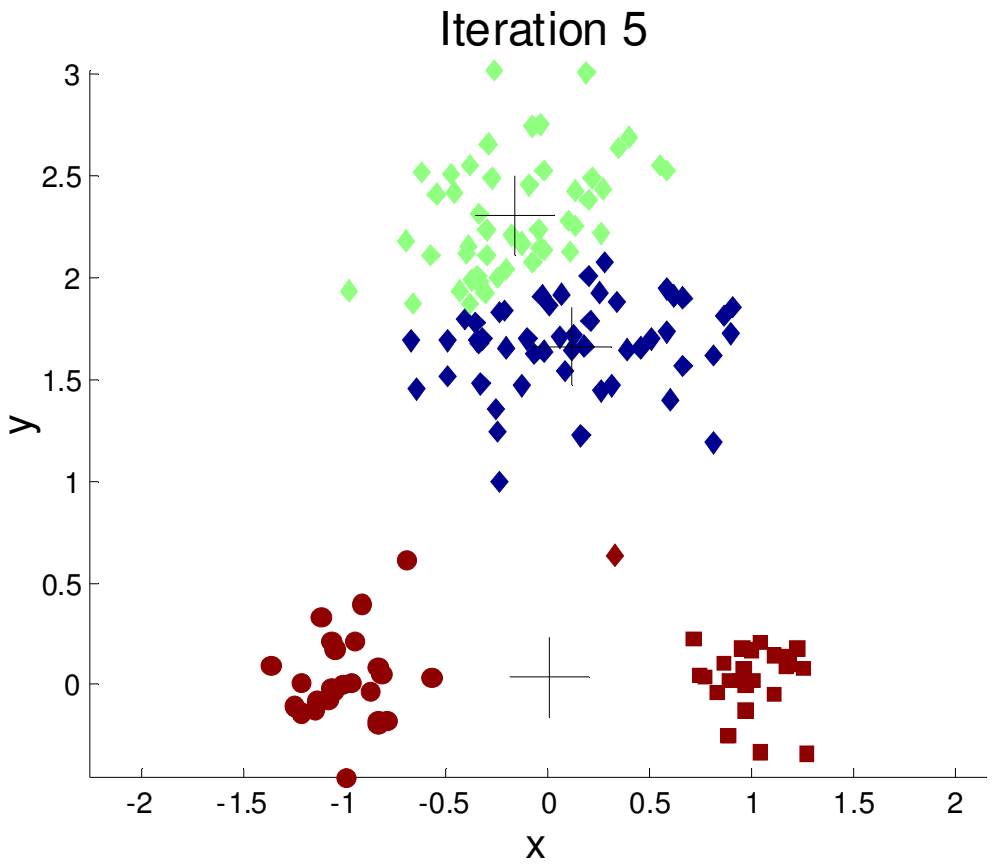
1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

K-means Ejemplo 1



Fuente: http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap8_basic_cluster_analysis.ppt

K-means Ejemplo 2



Fuente: http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap8_basic_cluster_analysis.ppt

K-means

- Ventajas
  - Relativamente eficiente: O(**t k n**), donde: **n** es el número de vectores, **t** número de iteraciones (**t** << **n**), **k** es número de centroides
  - Fácil de implementar, paralelismo
- Desventajas
  - Muchas veces termina en un óptimo local
  - Utilizable sólo cuando el promedio está definido (no se puede utilizar en espacios métricos generales)
  - Enfocado en identificar clusters circulares
  - Necesita especificar **k** (número de clusters) como parámetro
  - No es robusto a ruido, outliers, densidades distintas

Otros algoritmos de Clustering

- Los algoritmos jerárquicos son muy buenos pero muy costosos
  - Requieren calcular una matriz de distancia entre todos los vectores del dataset, costo $O(n^2)$ en memoria, y el costo en tiempo del clustering es usualmente $O(n^3)$
- Algoritmos basados en densidades (como DBSCAN) no funcionan bien con vectores de alta dimensión
- En el caso de Multimedia Information Retrieval casi siempre se usa K-means principalmente porque es el único algoritmo factible de usar

Clasificadores

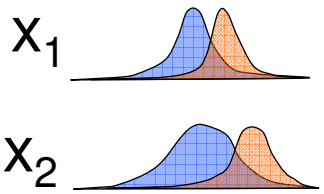
- Los distintos clasificadores se diferencian en el enfoque usado para procesar descriptores y crear regiones
- Ejemplos de clasificadores:
  - Regresión Logística
  - Clasificador Bayesiano
  - Árbol de Decisión
  - Clasificador K-NN
  - Máquina de vectores de soporte (SVM)
  - Redes Neuronales Artificiales

Clasificador Bayesiano

- Con datos de entrenamiento se obtienen las distribuciones de probabilidad de los descriptores de cada clase
- Para una entidad nueva, se calcula su descriptor y se calcula la clase más probable utilizando el Teorema de Bayes:

$$P(\text{Clase} \mid \vec{x}) = \frac{P(\vec{x} \mid \text{Clase})\ P(\text{Clase})}{P(\vec{x})}$$

Entrenamiento:



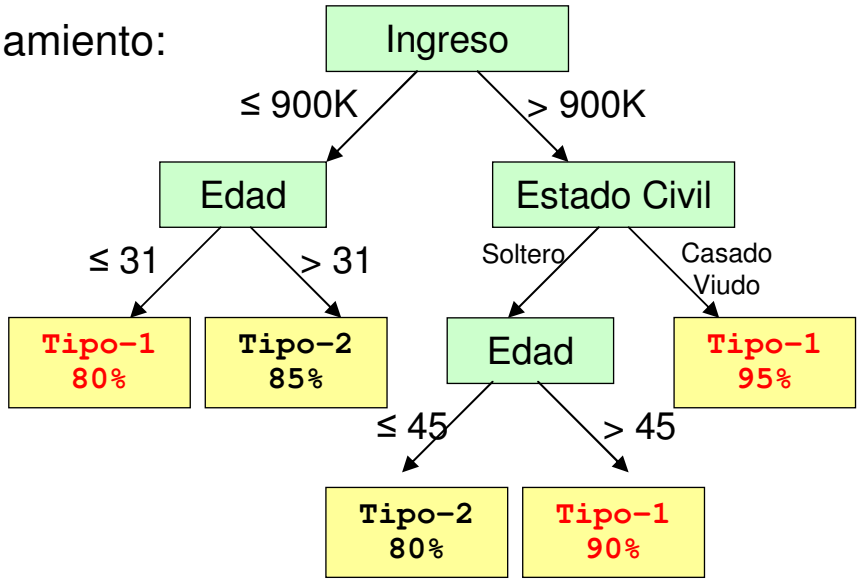
$$\vec{x} = (152, 350\text{K})$$

$$P(\texttt{Tipo1} \mid \vec{x}) = P(\vec{x} \mid \texttt{Tipo1})\ P(\texttt{Tipo1})$$
$$P(\texttt{Tipo2} \mid \vec{x}) = P(\vec{x} \mid \texttt{Tipo2})\ P(\texttt{Tipo2})$$

Árbol de Decisión

- Usar los descriptores de entrenamiento para determinar una secuencia de atributos y rangos que mejor separen los descriptores en cada clase
- Random Forest: Se calculan varios árboles sobre los mismos datos

Entrenamiento:



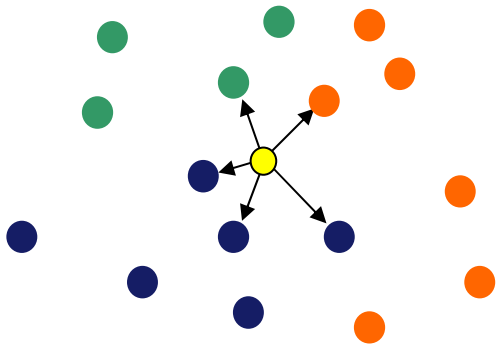
$$\vec{x} = (38, 950\text{K}, \texttt{Soltero})$$

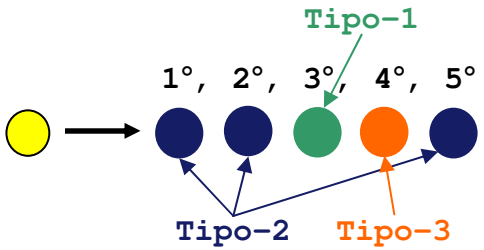
Tipo-2
score=80%

Clasificador k-NN

- El entrenamiento consiste en indexar vectores (Linear Scan, R-tree, kd-tree, kmeans tree, LSH, etc.)
- Para clasificar un descriptor nuevo, se buscan los k descriptores de entrenamiento más cercanos
- Se realiza una votación entre las k etiquetas de entrenamiento encontradas y se decide la más votada

Entrenamiento:





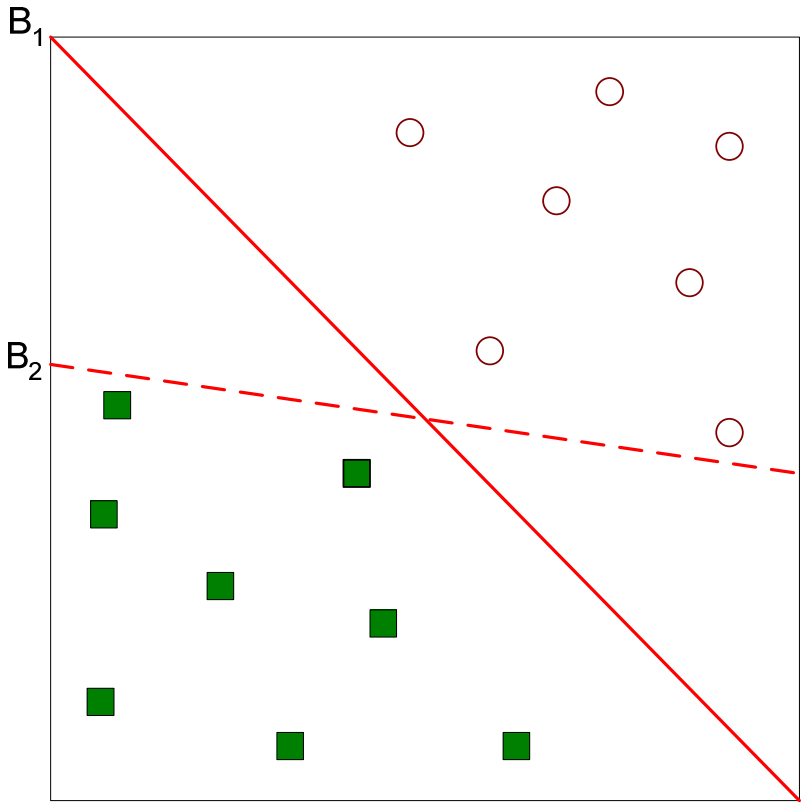
Resultado: Tipo-2 score=60%

Clasificador k-NN

- Métodos de votación de etiquetas:
  - El voto del *i*-ésimo NN vale *k*+1-*i*
  - El voto del *i*-ésimo NN vale 1/i
  - El voto del NN a distancia *d* vale 1/*d*
- No generaliza a regiones (Lazy Learning)
- Basta un ejemplo por clase
- Si se usa un índice dinámico (ej: R-tree) se pueden actualizar los datos de entrenamiento online
- Difícil escalar con muchos datos de entrenamiento
- Es posible explicar cada respuesta (mostrar los datos de entrenamiento que causaron una respuesta)

Support Vector Machine (SVM)

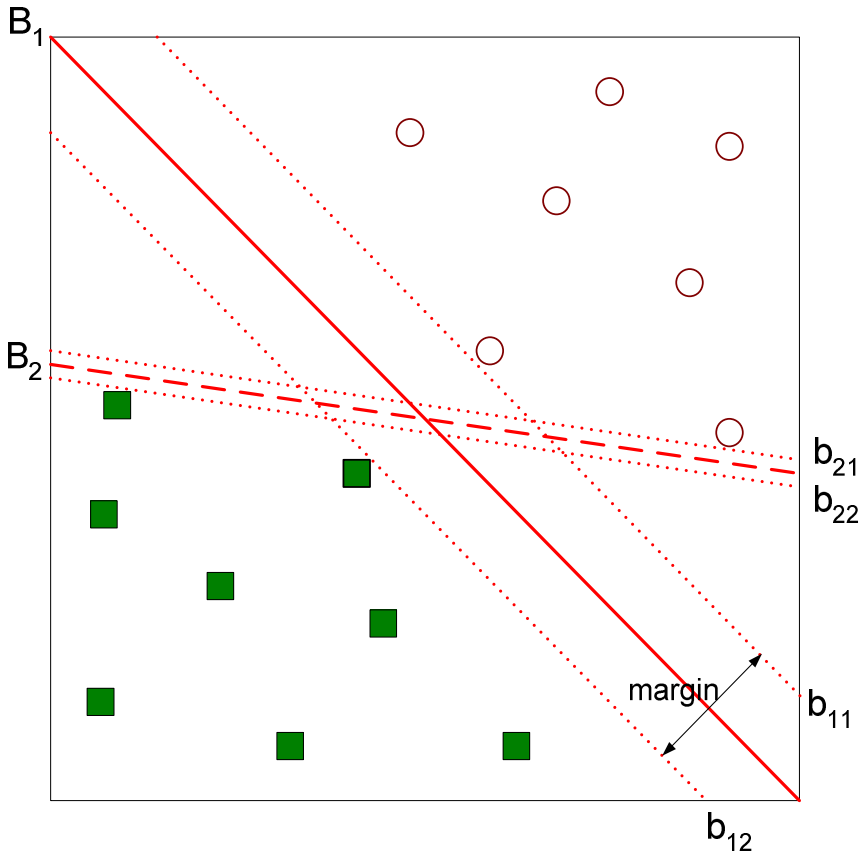
- Clasificador lineal de vectores en el espacio
- Se basa en calcular un hiperplano que separa las dos clases
- Dado un conjuntos de datos de entrenamiento, encontrar la recta que separa las dos clases



Fuente: http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap5_alternative_classification.ppt

SVM

- Existen muchos planos que pueden separar las dos clases
- Elegir el plano que maximiza el “margen”



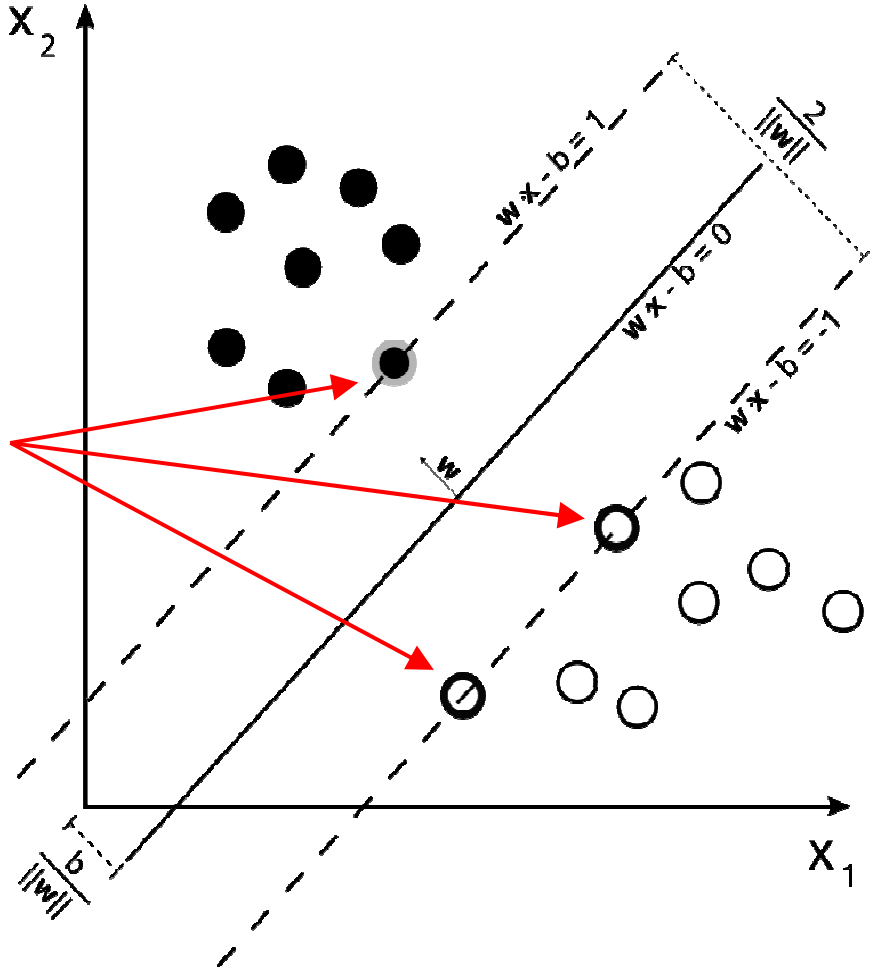
Fuente: http://www-users.cs.umn.edu/~kumar/dmbook/dmslides/chap5_alternative_classification.ppt

SVM

- Un hiperplano es definido como todos los puntos $\vec{x}$ del espacio que cumplen:

$$\vec{w} \cdot \vec{x} - b = 0$$

- Donde $\vec{w}$ es la normal y es la $b$istancia al plano.



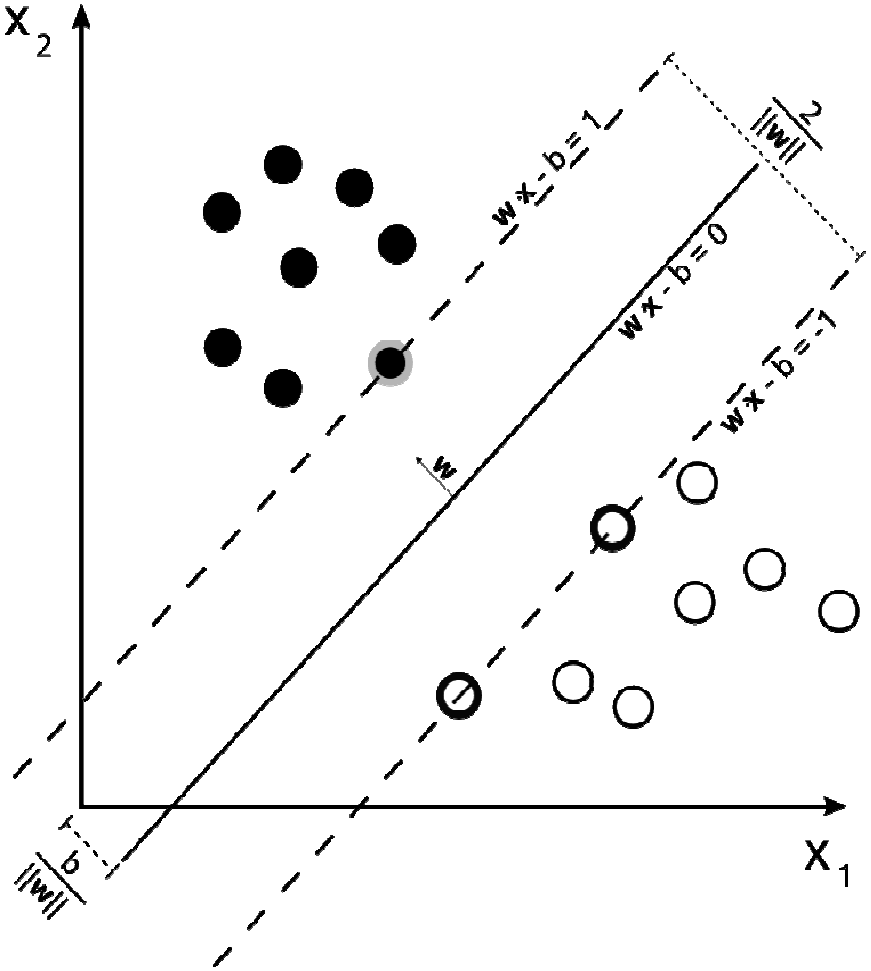
SVM

- Cuando un punto está en la frontera de la clase 1 entonces $\vec{w} \cdot \vec{x} - b$ vale 1, y si está en la otra frontera vale -1
- Se desea encontrar el plano con mayor distancia entre las dos fronteras (margen):

$$\min_{\vec{w},b} \frac{1}{2}||\vec{w}||^2$$

$$\text{sujeto a} \qquad c_i(\vec{w} \cdot \vec{x_i} - b) \geq 1$$

Donde x_i es el i-ésimo dato de entrenamiento que pertenece a la clase c_i (-1 o 1)



SVM: Soft Margin

- Las clases podrían no ser linealmente separables o pueden existir datos de entrenamiento incorrectos
- Soft Margin: asignar una penalización a los datos incorrectamente clasificados:

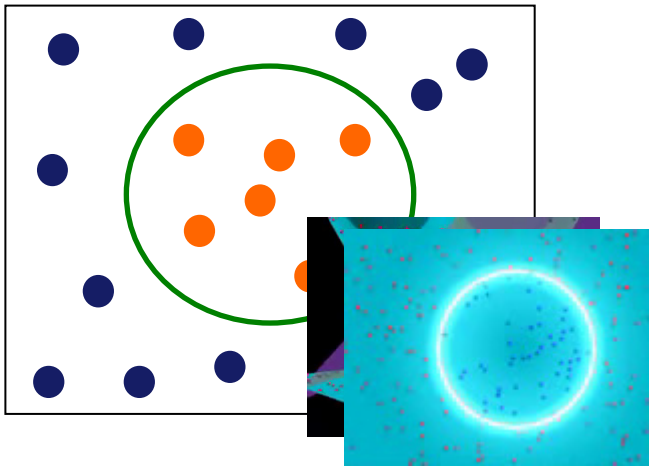
$$\min_{\vec{w},b} \ \frac{1}{2}||\vec{w}||^2 + C\sum_i \xi_i$$

$$\text{sujeto a} \qquad c_i(\vec{w}\cdot\vec{x_i} - b) \geq 1 - \xi_i$$

- Donde *C* es un parámetro y ξ_i es el grado de error permitido al dato x_i
- Al resolver este problema con multiplicadores de Lagrange la solución final sólo depende de *C* que queda como parámetro de clasificador

SVM: Kernel Trick

- SVM sólo puede separar clases linealmente, pero es común requerir separaciones no lineales
- Kernel Trick: modificar el espacio para llevarlo a una dimensión mayor. Reemplazar el producto punto $k(\vec{x}, \vec{y})$ por una función no lineal y luego usar un SVM lineal
- Kernels usuales:
  - Lineal $k(\vec{x}, \vec{y}) = \vec{x} \cdot \vec{y}$
  - RBF $k(\vec{x}, \vec{y}) = e^{\frac{-||\vec{x}-\vec{y}||^2}{std^2}}$
  - Polinomial $k(\vec{x}, \vec{y}) = (s(\vec{x} \cdot \vec{y}) + r)^d$



http://youtu.be/3liCbRZPrZA

SVM: Multiclase

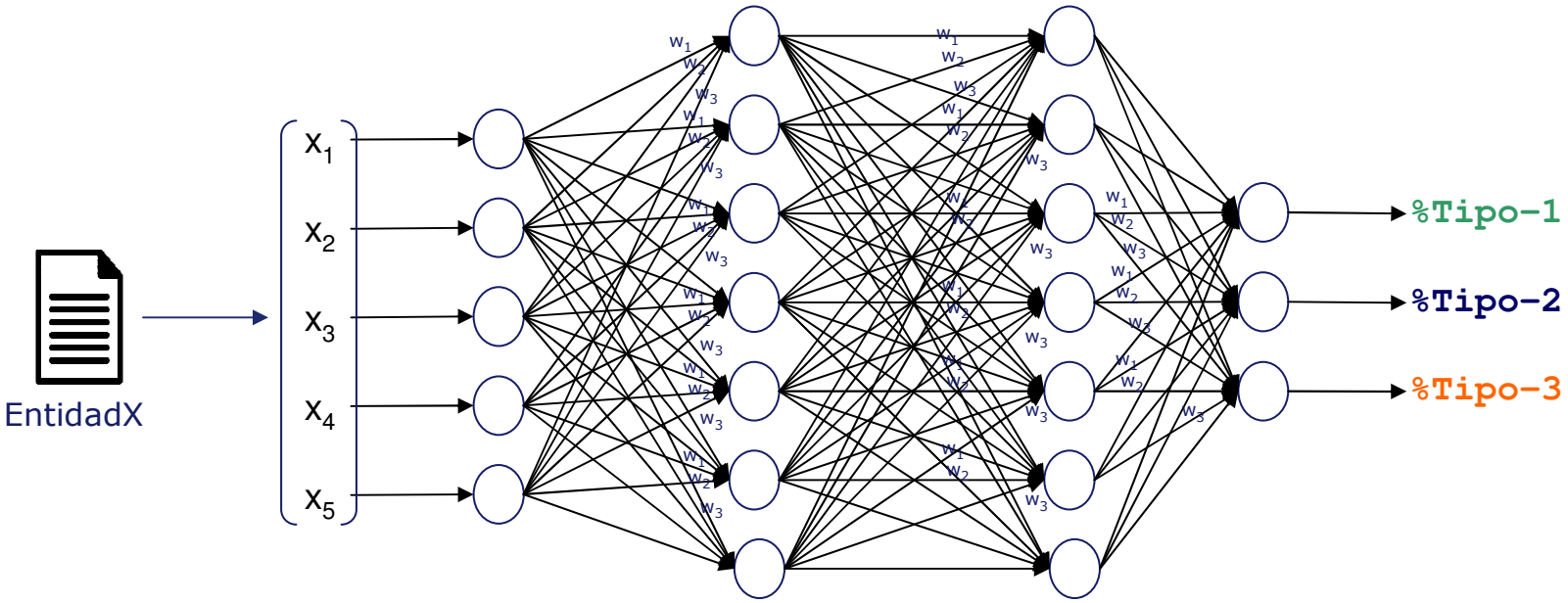
- SVM es un clasificador binario:

$$\vec{u} \in \omega_1 \Leftrightarrow h(\vec{u}) \geq 0$$
$$\vec{u} \in \omega_2 \Leftrightarrow h(\vec{u}) < 0$$

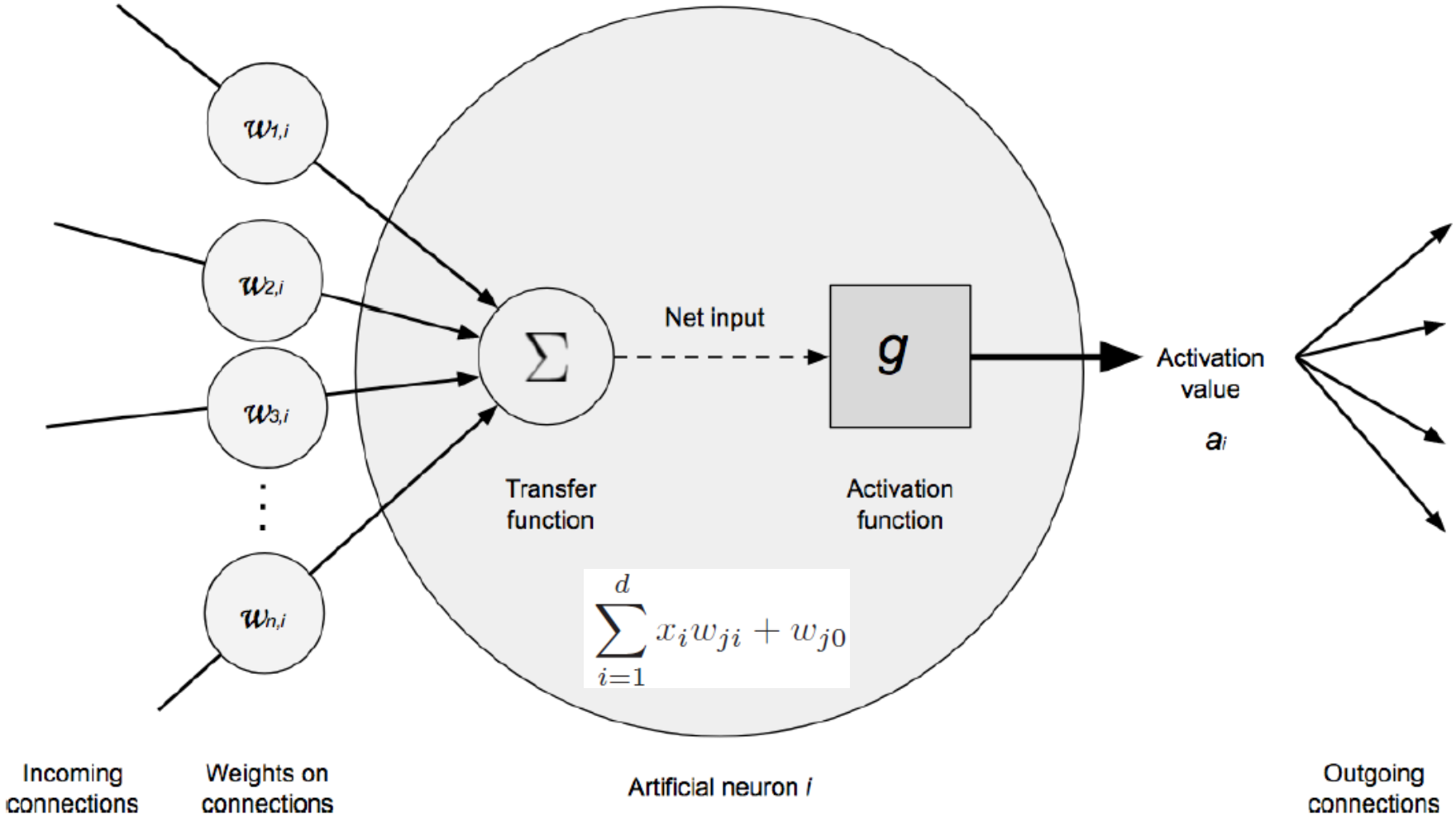
- “one-versus-all”: usar un SVM para cada clase, evaluar cada uno de ellos y elegir el de mayor confianza
- “one-versus-one”: entrenar un SVM para cada par de clases, evaluar en cada uno de ellos y elegir la clase con mayor votación

Red Neuronal Artificial

- Redes Neuronales Artificiales (MLP)
  - La función de clasificación se modela como operaciones en capas
  - Capa de **entrada**, capas **escondidas** y capa de **salida**
  - El **entrenamiento** ajusta los **pesos internos** que logran **convertir** cada **entrada** en la **salida** deseada



MLP: Perceptrón

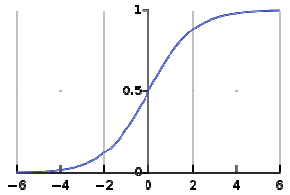


MLP: Parámetros

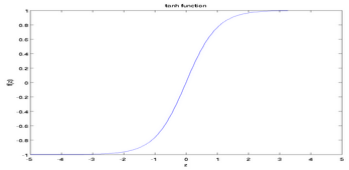
- Parámetros: pesos de las conexiones entre neuronas
- Función objetivo o de pérdida (loss): error entre la salida generada por la red y la salida deseada
- Hiperparámetros: arquitectura de red, número de capas, neuronas, función de activación, etc.
- Entrenamiento por Back Propagation. Ver videos:
  - Parte 1: https://www.youtube.com/watch?v=aircAruvnKk
  - Parte 2: https://www.youtube.com/watch?v=IHZwWFHWa-w
  - Parte 3: https://www.youtube.com/watch?v=Ilg3gGewQ5U

MLP: Función de Activación

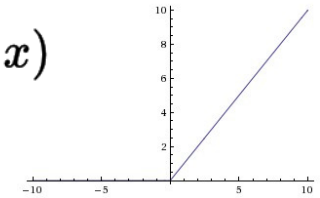
- Logistic (Sigmoid) $f(x) = \frac{1}{1 + e^{-x}}$



- Tanh $tanh(x) = \frac{2}{1+e^{-2x}} - 1$



- Rectified Linear Unit (ReLU) $f(x) = \max(0, x)$



- Leaky ReLU $f(x) = \begin{cases} x & \text{if } x > 0 \\ 0.01x & \text{otherwise} \end{cases}$

- Parametric ReLU
  - Reemplazar la constante 0.01 por un parámetro

MLP: Entrenamiento

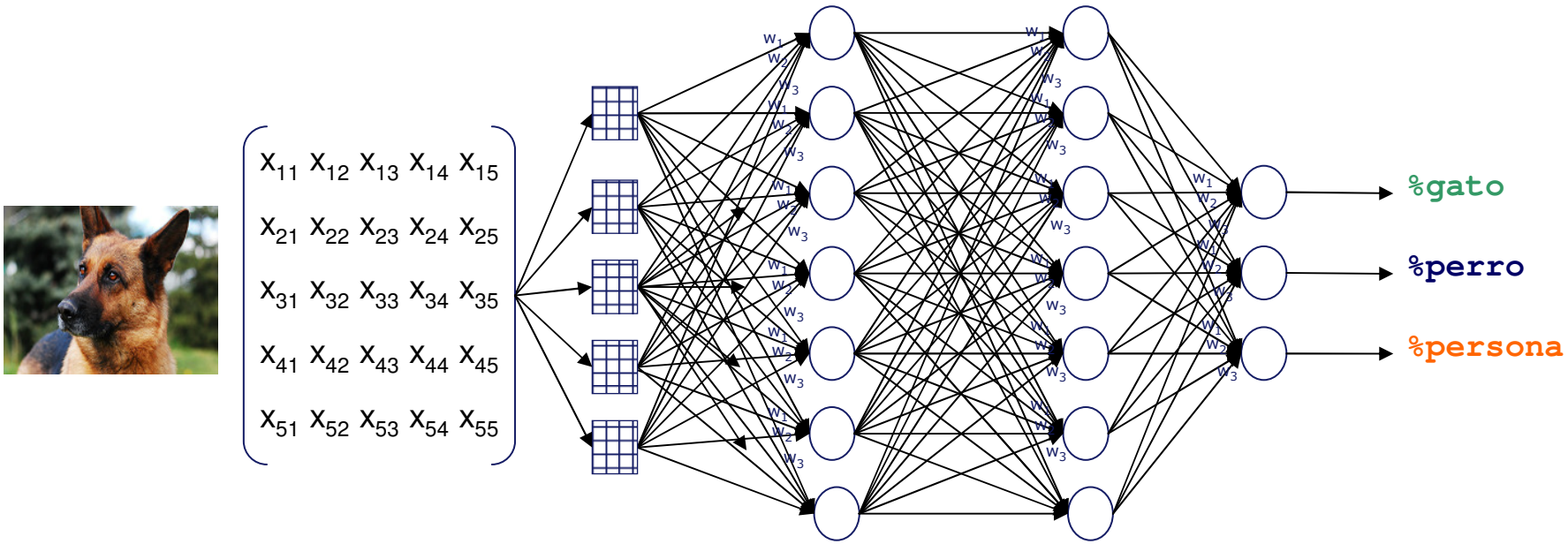
- Mini-batches: Entrenar por subconjuntos de vectores
- Stochastic Gradient Descent: El cálculo se realiza por grupos aleatorios (Stochastic=Random)
- Learning Rate: Ponderación a la modificación de pesos
- Dropout: En cada ciclo de entrenamiento se desactiva una fracción de las conexiones (usualmente 20-50%). Evita overfit al no permitir conexiones fuertes entre perceptrones
- Regularization: A la función de costo se le agrega el valor numérico del peso. Preferir varios pesos pequeños a un peso grande.
- Data Augmentation: Generar más datos de entrenamiento con pequeñas transformaciones o perturbaciones
- Transfer Learning: Comenzar el entrenamiento con valores obtenidos de otros datasets (incluso otros dominios)

Deep Learning

- Se refiere al uso de redes neuronales “muy grandes” o “profundas”
- Se diferencia de uso de redes neuronales “tradicionales” (MLP) por:
  - Incluir el cálculo del descriptor de contenido en la misma red
    - La entrada de la red es el dato multimedia mismo
  - Las red contiene gran cantidad de parámetros (neuronas, capas, arquitecturas complejas)
    - Entrenamiento requiere gran poder computacional

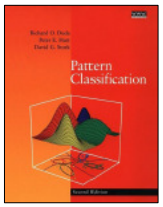
Red Neuronal Convolucional

- Red neuronal que contiene operadores de **convolución**
  - Forma eficiente cuando se buscan patrones con localidad espacial
  - El resultado de una convolución es una imagen de (casi) el mismo tamaño
  - Los valores del filtro son parámetros a entrenar
- El operador de pooling reduce el tamaño de una imagen
- El cálculo del descriptor es parte del entrenamiento

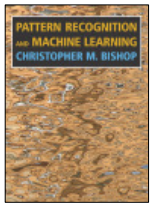


Bibliografía

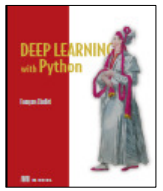
- **Pattern Classification. Second Edition**. Duda, Hart, Stork. 2001
  - Cap 6



- **Pattern Recognition and Machine Learning**. Bishop. 2006
  - Cap 5



- **Deep Learning with Python**. Chollet. 2018



https://xkcd.com/1838/