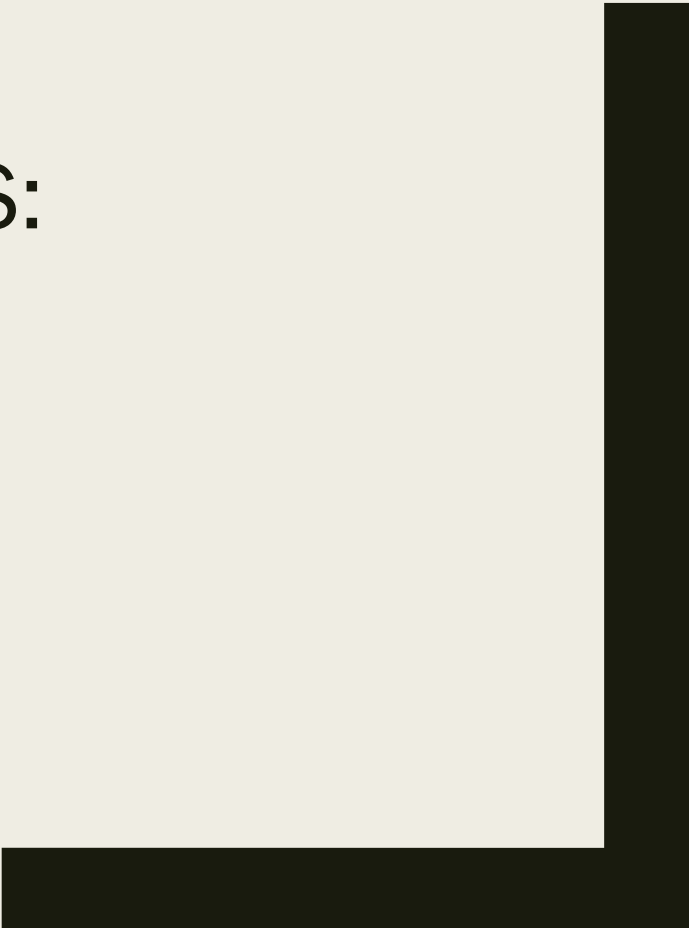




ANALÍTICA AVANZADA DE DATOS:

KNN

A. Alejandra Sánchez Manilla
asanchezm.q@gmail.com



Clasificador k-NN

- El clasificador k-NN (k-Nearest Neighbour), es un clasificador basado en instancias.
- A menudo a estos clasificadores se les denomina perezosos (lazy) debido a la característica de no contar como tal con una etapa de entrenamiento y dejar todo el procesamiento para la etapa de clasificación.

Clasificador k-NN

- Inicialmente los patrones del conjunto de entrenamiento son almacenados. Cuando se presenta un patrón de prueba, se utiliza una función de distancia para determinar cual patrón(es) del conjunto de entrenamiento es los vecinos más cercano del patrón de prueba a clasificar.
- Generalmente la función de distancia usada es la euclidiana.
- La clase es asignada de acuerdo con la clase más frecuente entre los k vecinos más cercanos.

Clasificador k-NN

1. Sea \tilde{x} un patrón de prueba a clasificar.
2. Calcular la distancia entre \tilde{x} y todos los patrones del conjunto de entrenamiento.
3. Se seleccionan los k patrones cuyas distancias sean las menores.
4. Verificar cual es la clase más frecuente entre los k patrones seleccionados y asignar dicha clase a \tilde{x} .

Nota: la k se selecciona con un valor impar para evitar empates.

Clasificadores basados en métricas

Clasificador k-NN

```
for each  $\tilde{x} \in P$  do
  for  $i=1:N_E$ 
     $d(i) = distancia(\tilde{x}, x^i)$  donde  $x^i \in E$ 
  end
   $d\_minima = \min(d)$ ;
  Asignar  $\tilde{x}$  a la clase que corresponda con la distancia mínima.
  Verificar que  $\tilde{x}$  haya sido asignado a la clase correcta
end
```

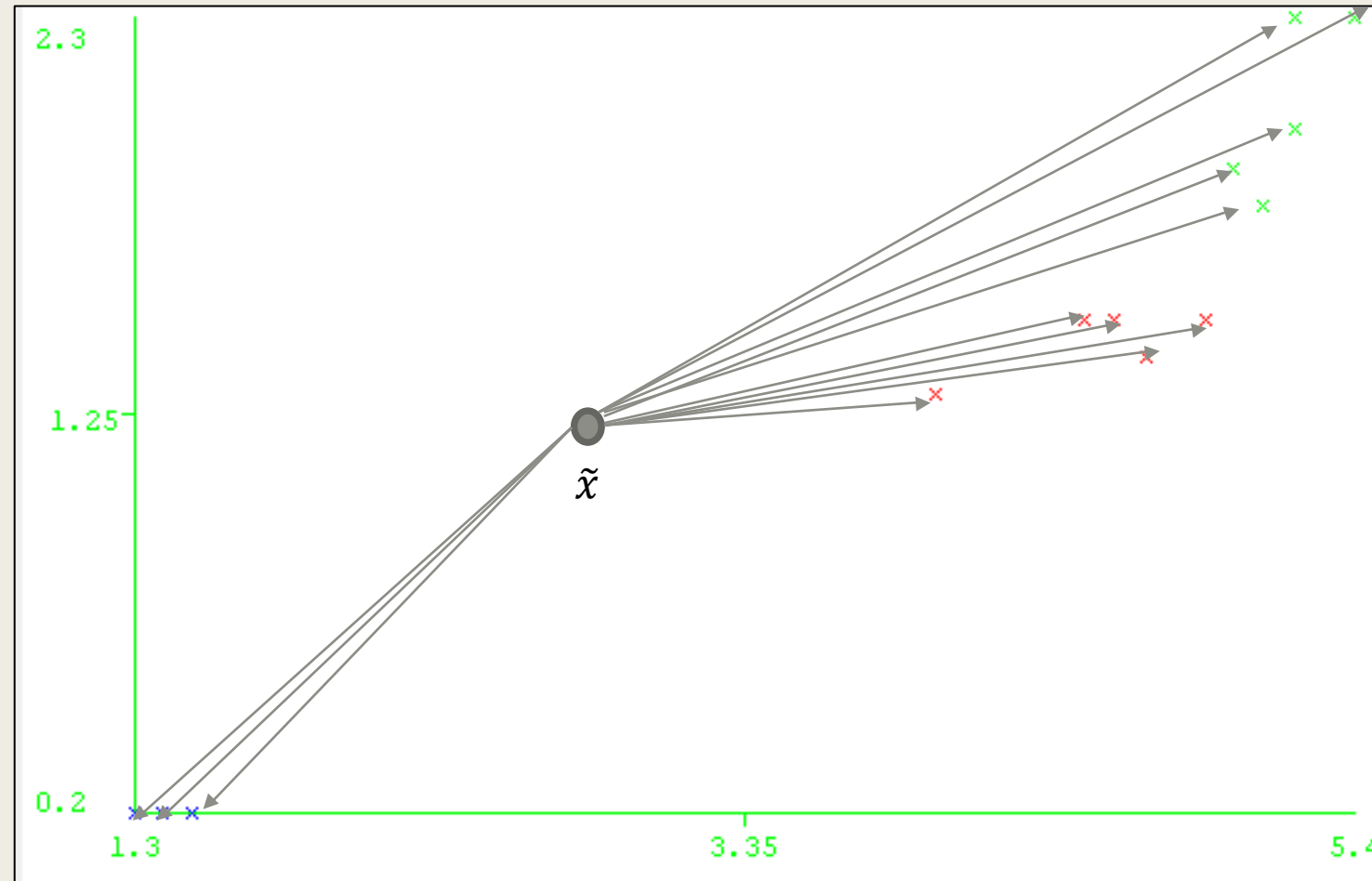
Clasificadores basados en métricas

Clasificador 1-NN (ejemplo)

Petallength	Petalwidth	Class
1.4	0.2	1
1.4	0.2	1
1.3	0.2	1
1.5	0.2	1
1.4	0.2	1
4.7	1.4	2
4.5	1.5	2
4.9	1.5	2
4	1.3	2
4.6	1.5	2
5.2	2.3	3
5	1.9	3
5.2	2	3
5.4	2.3	3
5.1	1.8	3

Clasificadores basados en métricas

Clasificador 1-NN (ejemplo)



Clasificadores basados en métricas

Clasificador 1-NN (ejemplo)

Vamos a considerar el último patrón de la tabla para clasificarlo, es decir $\tilde{x} = (5.1, 1.8)$

Distancia	Clase
4.0311	1
4.0311	1
4.1231	1
3.9395	1
4.0311	1
0.5657	2
0.6708	2
0.3606	2
1.2083	2
0.5831	2
0.5099	3
0.1414	3
0.2236	3
0.5831	3



EJEMPLOS DE MÉTRICAS



Ejemplos de métricas

- Para clasificar patrones, necesitan ser comparados entre sí y contra un estándar. Cuando un nuevo patrón se presenta y es necesario clasificarlo, la proximidad de este patrón a los patrones en el conjunto de entrenamiento debe ser encontrada.
- Las métricas de distancia son usada para encontrar la similitud entre patrones. Los patrones que son más similares deberían estar más cerca.

Métrica de Minkowski:

$$d^m(x, y) = \left(\sum_{k=1}^d (x_k - y_k)^m \right)^{\frac{1}{m}}$$

- Cuando $m=1$ se le llama distancia Manhattan o distancia L_1
- Cuando $m=2$ se le llama distancia Euclidiana o distancia L_2

$$d^2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_d - y_d)^2}$$

Ejemplos de métricas

Métrica de Minkowski ponderada:

$$d^m(x, y) = \left(\sum_{k=1}^d w_k \times (x_k - y_k)^m \right)^{\frac{1}{m}}$$

Distancia Mahalanobis

$$d_M(x, y) = \sqrt{(x - y)^T S^{-1} (x - y)}$$

$$S = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}$$

$$S^{-1} = \frac{1}{\sigma_1^2 \sigma_2^2} \begin{bmatrix} \sigma_2^2 & 0 \\ 0 & \sigma_1^2 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix}$$

Distancia Mahalanobis

$$\begin{aligned}d_M(x, y) &= \sqrt{(x - y)^T S^{-1} (x - y)} \\&= \sqrt{[x_1 - y_1 \quad x_2 - y_2] \begin{bmatrix} \frac{1}{\sigma_1^2} & 0 \\ 0 & \frac{1}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}} \\&= \sqrt{\begin{bmatrix} \frac{x_1 - y_1}{\sigma_1^2} & \frac{x_2 - y_2}{\sigma_2^2} \end{bmatrix} \begin{bmatrix} x_1 - y_1 \\ x_2 - y_2 \end{bmatrix}} \\&= \sqrt{\frac{(x_1 - y_1)^2}{\sigma_1^2} + \frac{(x_2 - y_2)^2}{\sigma_2^2}}\end{aligned}$$

$$S = \frac{1}{m} \sum_{i=1}^m (x^{(i)} - \mu)(x^{(i)} - \mu)^T$$

Clasificadores basados en métricas

Distancia k-mediana

$$d(x, y) = k\text{-medianas}\{|x_1 - y_1|, |x_2 - y_2|, \dots, |x_n - y_n|\}$$

- Donde el operador $k\text{-medianas}$ regresa el $k\text{-esimo}$ valor del vector de diferencias ordenado.

Ejemplo:

- Sea $x = (50, 3, 100, 29, 62, 140)$ y $y = (55, 15, 80, 50, 70, 170)$ y $k = 3$.
- El vector de diferencias sería $(5, 12, 20, 21, 8, 30)$, entonces:

$$d(x, y) = k\text{-medianas}\{5, 8, 12, 20, 21, 30\}$$

$$d(x, y) = 12$$

Referencias

- [1] Leondes, C.T. (2018). *Image Processing and Pattern Recognition*. California: Academic Press.
- [2] Duda, R.O., Hart, P.E. & Stork, D.G. (2001). *Pattern Classification*. 2nd edition. Wiley-Interscience.
- [3] Marques de Sá, J:P. (2001). *Pattern Recognition: Concepts, Methods and Applications*. Berlin: Springer-Verlag.
- [4] Kuncheva, L. (2014). *Combining Pattern Classifiers: Methods and Algorithms*. 2nd edition. USA: Wiley.
- [5] Witten, I.H., Frank, E. & Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. USA: Elsevier.
- [6] Murty, N.M. & Devi, V.S. (2011). *Pattern Recognition: An Algorithmic Approach*. Springer.
- [7] Zaki, M.J. & Meira, W. (2014). *Data Mining and Analysis: Fundamental Concepts and Algorithms*. Cambridge University Press.
- [8] Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. & Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems With Applications*, 73, 220-239.

Tarea 4.1

1. Codificar el clasificador euclidiano al banco de datos Iris plant; el resubstitution error y los métodos de validación: leave one out, hold out (70-30) y 10-fold cross-validation.
2. Realizar el mismo procedimiento solicitado en el punto 1 para alguno de los bancos que eligieron en la Tarea 1.

Tarea 4.2

1. Aplicar el clasificador 1-NN al banco de datos Iris plant usando como método de validación Resubstitution error y el 10-fold cross-validation.
2. Aplicar el clasificador 3-NN al banco de datos Iris plant usando como método de validación 10-fold cross-validation.
3. Realizar lo solicitado en el punto 2 para el banco de datos elegido en la tarea 1.