Clasificación Bayesiana con Técnicas no paramétricas : k-NN

FaMAF

2019

Clasificación Bayesiana con Técnicas no paramétricas

- La mayoría de las densidades paramétricas son unimodales (tienen un único máximo local), mientras que muchos problemas prácticos envuelven densidades multi-modales.
- La estimación paramétrica de la mezcla de gaussianas precisa muchos recursos, tanto en datos como en tiempo de cómputo.
- Los procedimientos no paramétricos pueden ser usados con distribuciones arbitrarias y sin la hipótesis de que las densidades son conocidas.

Clasificación Bayesiana: Técnicas no paramétricas

Enfoques de estimación de densidad no paramétrica en el problema de clasificación:

- Estimar funciones de verosimilitud$p\left(\mathbf{x}|\omega_j\right)$.
- Estimar directamente probabilidades posteriores $p\left(\omega_j|\mathbf{x}\right)$.

Cuando la densidad es desconocida, la estimación paramétrica intenta ajustar una forma global a la muestra de entrenamiento, como vimos con la mezcla de Gaussianas.

En cambio la estimación no paramétrica estima localmente valores para la densidad.

Estimación densidad no paramétrica

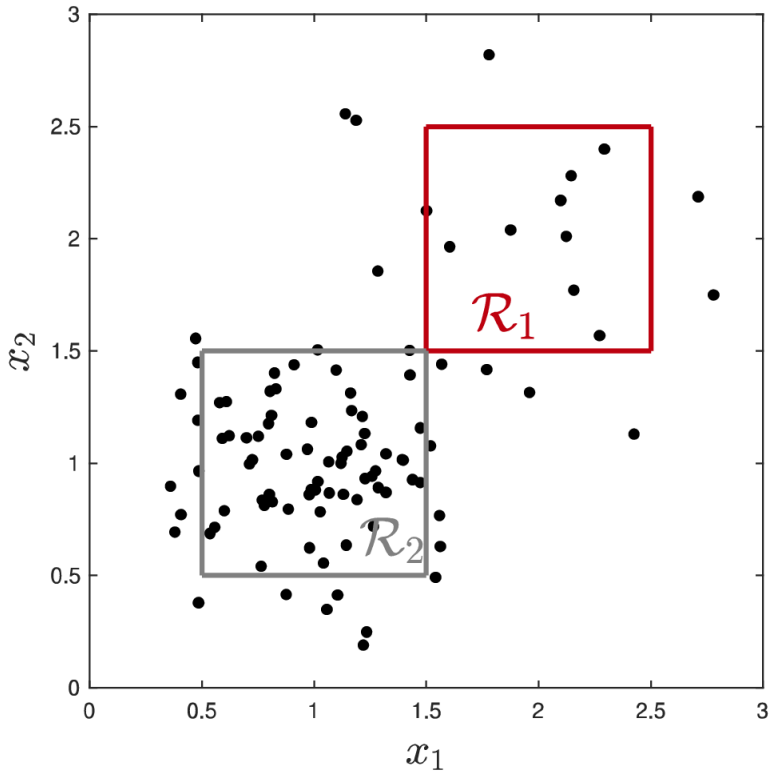
Idea Básica de estimación de densidad $p(\mathbf{x})$:

- Estimamos $p(\mathbf{x})$ con la probabilidad P de que un patrón $\mathbf{x}_i$ caiga en una región $\mathcal{R}$ con volumen 1 que contiene a $\mathbf{x}$.

$$P = \int_{\Re} p\left(\mathbf{x}'\right) d\mathbf{x}' \tag{1}$$

- Intuitivamente, un estimador de P es la fracción de muestras que caen en $\mathcal{R}$.
- P es una version suavizada o promediada de la función de densidad $p(\mathbf{x})$

Estimación densidad no paramétrica



Se tienen $n = 100$ muestras en un espacio bidimensional. En la región $\mathcal{R}_1$ caen $k = 9$ muestras:$\hat{p}(\hat{\mathbf{x}}) = 0.09$. En la región $\mathcal{R}_2$ caen $k = 60$ muestras: $\hat{p}(\hat{\mathbf{x}}) = 0.6$. Ambas regiones tienen área 1.

Estimación densidad no paramétrica

- Si tenemos una muestra de tamaño n, $\mathbf{x}_1, \ldots, \mathbf{x}_n$ tomadas i.i.d. de una distribución $p(\mathbf{x})$.
- Ley binomial: probabilidad de que k de estas n muestras caigan en una región arbitraria $\mathcal{R}$ es:

$$P(N = k) = \left(\begin{array}{c} n \\ k \end{array} \right) P^k (1 - P)^{n-k} \tag{2}$$

- Estimador de máxima verosimilitud para P:

$$\hat{P} = \frac{k}{n} \tag{3}$$

  es un buen estimador de la probabilidad P.

Estimación densidad no paramétrica

- Si p es continua y si la región $\mathcal{R}$ es tan pequeña que p no varía mucho en ella podemos escribir

$$\int_{\mathcal{R}} p\left(\mathbf{x}'\right) d\mathbf{x}' \cong p(\hat{\mathbf{x}}) \int_{\mathcal{R}} d\mathbf{x}' = p(\hat{\mathbf{x}})V \tag{4}$$

  donde $\hat{\mathbf{x}}$ es el punto medio de $\mathcal{R}$ y V es el volumen de $\mathcal{R}$.

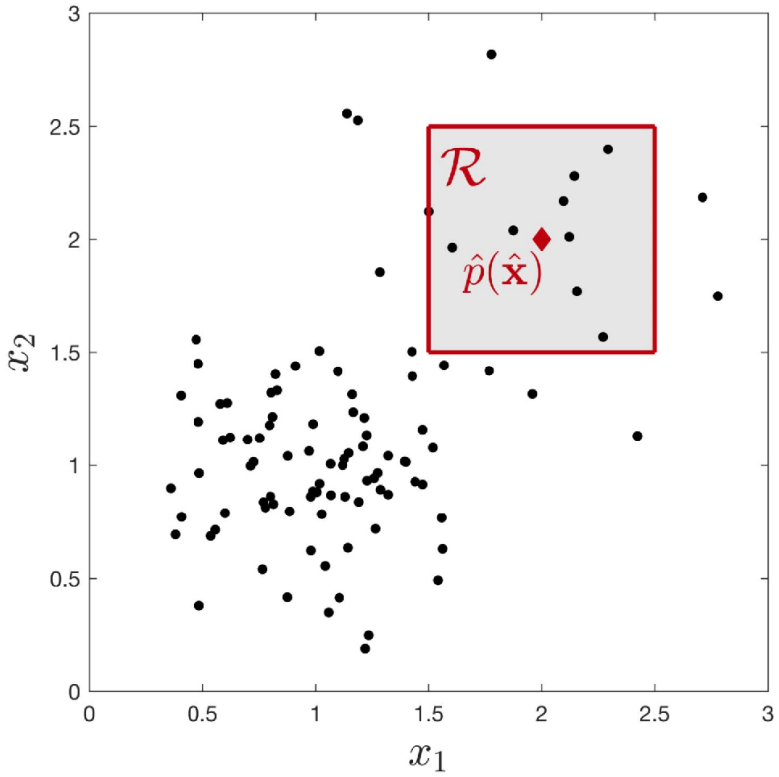
- Por lo cual

$$p(\hat{\mathbf{x}}) \cong \frac{k/n}{V} = \hat{p}(\hat{\mathbf{x}}) \tag{5}$$

  donde $\hat{\mathbf{x}}$ es el punto medio de $\mathcal{R}$ y V el volumen de $\mathcal{R}$.

- El estimador $\hat{p}(\hat{\mathbf{x}})$ depende del volumen V de la region considerada. Si $V = 1$ el estimador es la fracción de muestras que caen en V.

Estimación densidad no paramétrica



Densidad $p(.)$ se asume constante en todo $\mathcal{R}$, $p(\mathbf{x}) \equiv \hat{p}(\hat{\mathbf{x}}) \approx \frac{1}{V}\frac{k}{n}$ donde $\hat{\mathbf{x}}$ es el punto medio y V el volumen de $\mathcal{R}$.

Estimación densidad no paramétrica

- La fracción $k/(nV)$ es un valor promediado de $p(\mathbf{x})$.
- $p(\mathbf{x})$ se obtiene exacta solo si V se acerca a cero.
- Si n es un número fijo, puede no haber muestras en $\mathcal{R}$, por lo cual

$$\lim_{V \to 0, \mathrm{k}=0} \hat{p}(\mathbf{x}) = 0 \tag{6}$$

- Si alguna o mas muestras coinciden con $\mathbf{x}$, el estimador diverge

$$\lim_{V \to 0, k \neq 0} \hat{p}(\mathbf{x}) = \infty \tag{7}$$

Clasificación Bayesiana: Técnicas no paramétricas

- El volumen V necesita ir a cero o no se podria usar esta estimación.
- Sin embargo, en la práctica V no puede ser muy chico pues el número de muestras es siempre limitado.
- Se tiene que aceptar variabilidad en el radio k/n y un promedio en la densidad $p(\mathbf{x})$.

Clasificación Bayesiana: Técnicas no paramétricas

- Teóricamente, con infinitas muestras, para estimar la densidad en $\mathbf{x}$, se forma una sucesión de regiones $\mathcal{R}_1, \mathcal{R}_2, \ldots$ que contienen $\mathbf{x}$: la primera con una muestra, la segunda con dos, etc...
- Si V_n es el volumen de $\mathcal{R}_n$, k_n el número de muestras en $\mathcal{R}_n$ y $\hat{p_n}(\mathbf{x})$ el n-esimo estimador de $p(\mathbf{x})$.

$$\hat{p_n}(\mathbf{x}) = \frac{1}{V_n}\frac{k_n}{n} \tag{8}$$

- Hay tres condiciones necesarias si para asegurar la convergencia $\hat{p_n}(\mathbf{x}) \to p(\mathbf{x})$ cuando $n \to \infty$
  - $\lim_{n\to\infty} V = 0$
  - $\lim_{n\to\infty} k = \infty$
  - $\lim_{n\to\infty} k/n = 0$, n y k suficientemente grandes y V suficientemente pequeño.

Clasificación Bayesiana: Técnicas no paramétricas

- La primera condición
  - $\lim_{n\to\infty} V = 0$

  asegura que el radio P/V converge a $p(\mathbf{x})$.
- La segunda condición
  - $\lim_{n\to\infty} k = \infty$

  asegura que la frecuencia de observación converge en probabilidad a P
- La tercera condición
  - $\lim_{n\to\infty} k/n = 0,$

  asegura que si bien hay infinitas muestras en la ventana, solo son una parte pequeña del total de muestras.

Métodos de estimación

Hay dos métodos principales para obtener estas condiciones :

a) Determinar ventana mediante una función como $V_n = 1/\sqrt{n}$ de tal forma que el k_n resultante se comporte bien y se cumpla

$$\hat{p}_n(\mathbf{x}) \underset{n\to\infty}{\to} p(\mathbf{x}) \tag{9}$$

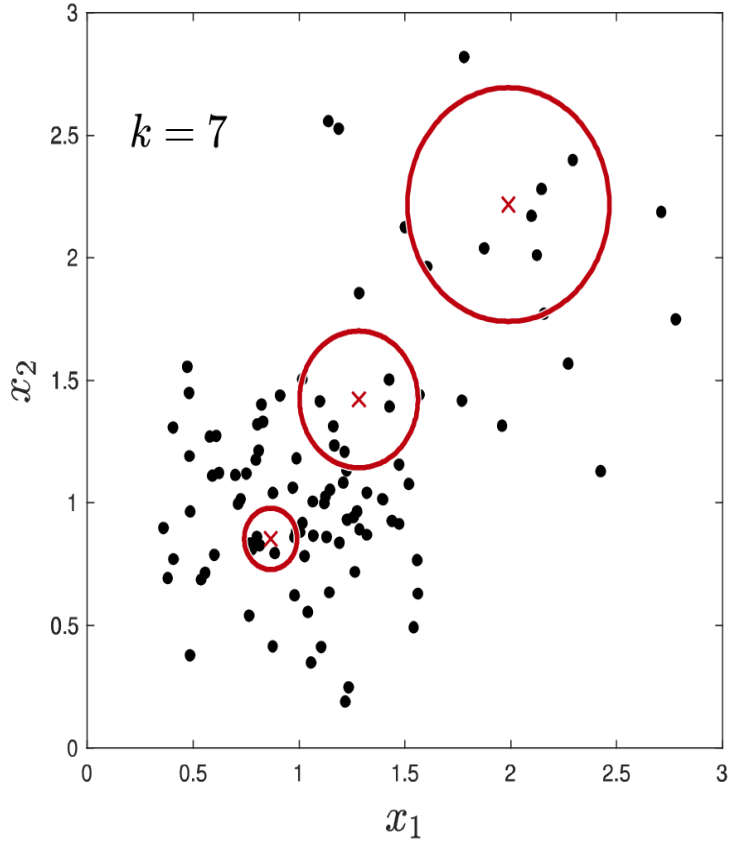
Este método es llamado “método de estimación de la ventana de Parzen”

b) Especificar k_n como una función de n, como $k_n = \sqrt{n}$; el volumen V_n crece hasta que engloba k_n vecinos de $\mathbf{x}$. Este método se llama “estimacion por los vecinos mas cercanos”.

Métodos de estimación: Vecinos mas cercanos

- En la estimación por k-vecinos más cercanos (kNN) el número de puntos k es fijo y el tamaño del volumen alrededor de $\hat{\mathbf{x}}$ se ajusta para incluir exactamente k puntos.

- En áreas de densidad baja el volumen será grande, contrariamente en áreas de densidad alta el volumen será pequeño.



Métodos de estimación: kNN

La estimación de densidad knn se expresa como:

$$\hat{p}(\hat{\mathbf{x}}) = \frac{k}{nV} \tag{10}$$

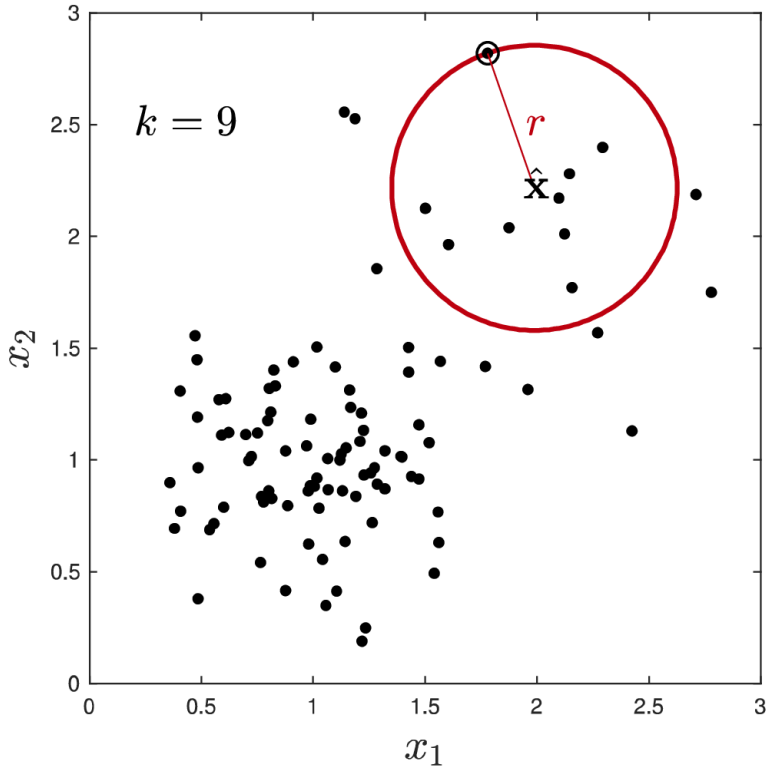
donde V es el volumen de una hiperesfera alrededor de $\hat{\mathbf{x}}$:

$$V = V_0 \cdot r^d \tag{11}$$

donde r es el radio y V_0 es el volumen de la hiperesfera unitaria:

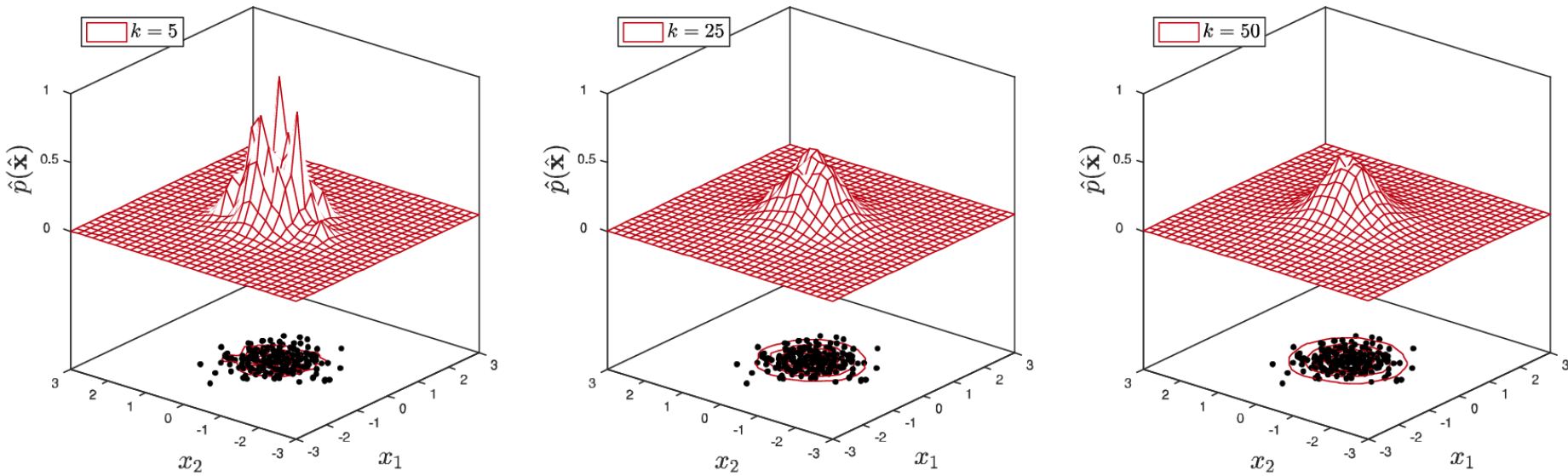
$$V_0 = \begin{cases} \pi^{d/2}/(d/2)! & \text{si } d \text{ es par} \\ 2^d \pi^{(d-1)/2} \left(\frac{d-1}{2}\right)!/d! & \text{si } d \text{ es impar} \end{cases} \tag{12}$$

k-vecinos mas cercanos



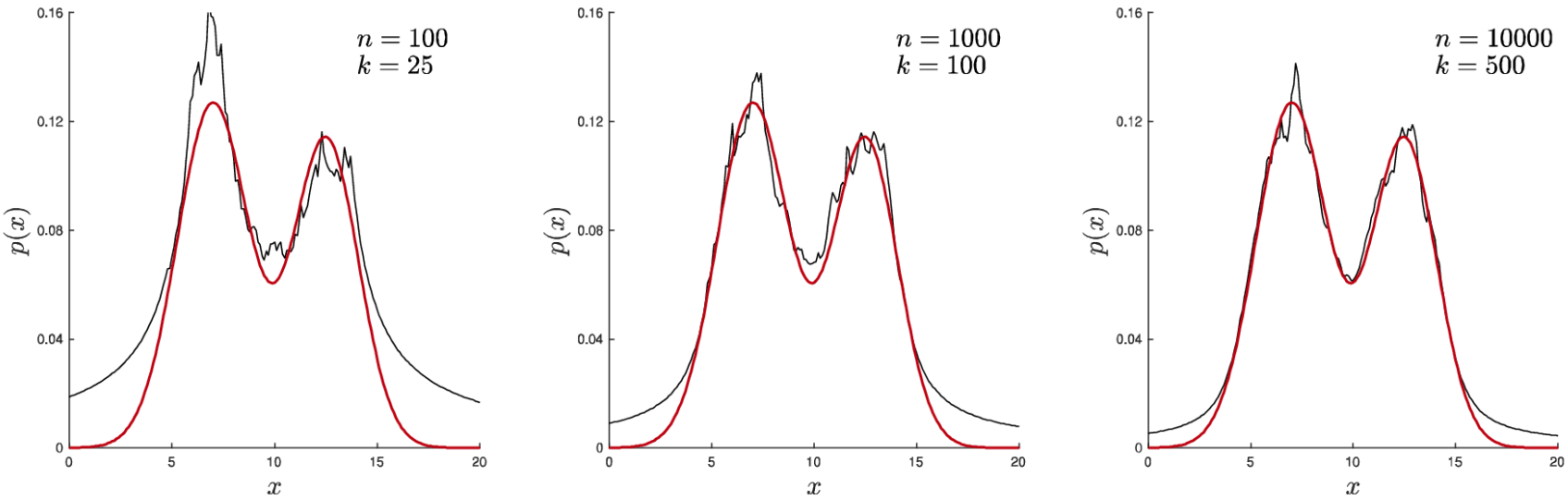
El radio r de la hiperesfera corresponde con la distancia entre $\hat{\mathbf{x}}$ y la muestra $\mathbf{x}$ más alejada entre los k vecinos de $\hat{\mathbf{x}}$.

k-vecinos mas cercanos



Estimación de densidad con kNN para muestras en $\mathbb{R}^2$. Para valores pequeños de k, la densidad es ruidosa, y se suaviza a medida que aumenta su valor.

k-vecinos mas cercanos



Las condiciones $\lim_{n\to\infty} k = \infty$ y $\lim_{n\to\infty} k/n = 0$ son necesarias y suficientes para que la densidad estimada $\hat{p}(\hat{\mathbf{x}})$ converja a $p(x)$. La curva roja muestra la densidad verdadera $p(x)$.

Probabilidad a posteriori en kNN

- Supongamos tener n muestras etiquetadas, considerando un volumen V alrededor de $\mathbf{x}$ que engloba exactamente k muestras, de las cuales k_i resultan ser pertenecientes a ω_i.
- El estimador de la densidad conjunta $P_n(\omega_i|\mathbf{x})$ es

$$p_n(\mathbf{x}, \omega_i) = \frac{k_i/n}{V} \tag{13}$$

- Y la probabilidad posterior resulta:

$$P_n(\omega_i|\mathbf{x}) = \frac{p_n(\mathbf{x}, \omega_i)}{\sum_{j=1}^{c} p_n(\mathbf{x}, \omega_j)} = \frac{k_i}{k} \tag{14}$$

Clasificación en kNN

- La estimación de las densidades tiene un sesgo provocado por el suavizado introducido por el volumen V. Ya sea eligiendo $V_n = 1/\sqrt{n}$ o $k_n = \sqrt{n}$, la estimación es razonable cuando n es grande.
- La regla de decisión Bayesiana resulta

$$\text{Decidir } \omega_i \text{ si } P_n(\omega_i|\mathbf{x}) > P_n(\omega_j|\mathbf{x}) \text{ para todo } i \neq j \tag{15}$$

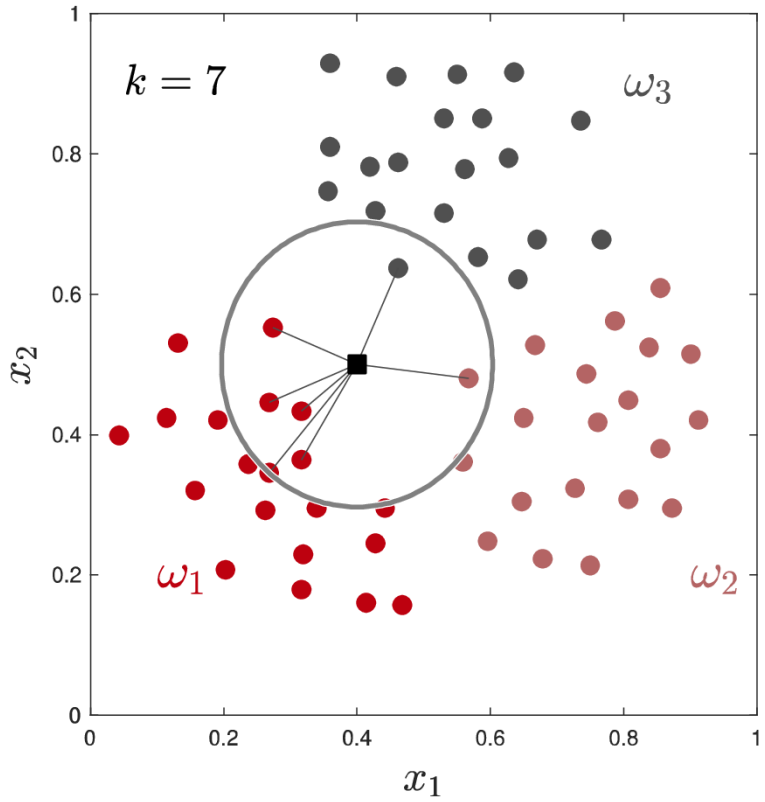
resulta

$$\text{Decidir } \omega_i \text{ si } k_i > k_j \text{ para todo } i \neq j \tag{16}$$

esto es, se elige la categoría mejor representada en la celda generada por k quien en teoría debería ser una función de n, lo cual aumenta la complejidad de la regla.

- Sin embargo, en el problema de clasificación, se pueden obtener resultados equivalentes utilizando el vecino mas cercano, esto es, $k = 1$, o k relativamente pequeño con respecto al tamaño de n.

Clasificación k-NN



$$p(\omega_1|\mathbf{x}_t) = 5/7 \quad p(\omega_2|\mathbf{x}_t) = 1/7 \quad p(\omega_3|\mathbf{x}_t) = 1/7$$

Algoritmo kNN

Algoritmo kNN: clasifica un patrón de prueba $\mathbf{x}_t$ asignándole la etiqueta de clase más frecuente entre sus k muestras de entrenamiento más cercanas

Input: $Z = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}, \mathbf{x}_t, k$

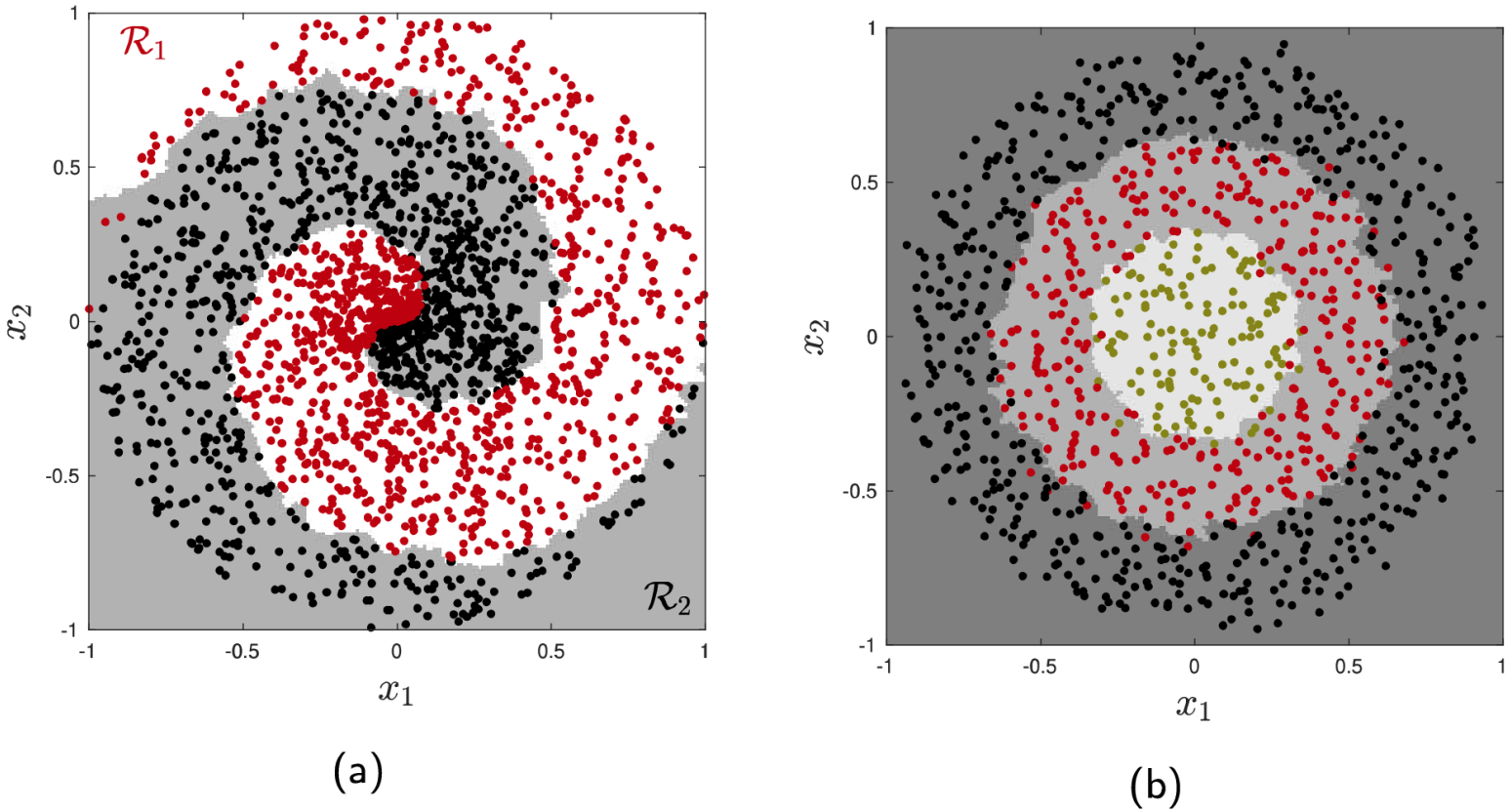
Output: y_t

- Computar $\|\mathbf{x}_t - \mathbf{x}_i\|_2$, la distancia entre $\mathbf{x}_t$ y cada patrón $(\mathbf{x}, y) \in Z$
- Seleccionar $Z_k \subset Z$, el conjunto de los k patrones más cercanos a $\mathbf{x}_t$
- **return** $y_t = \arg\max_{j=1,\ldots,c} \sum_{(\mathbf{x}_i, y_i) \in Z_k} \mathbf{1}_j (y_i)$

donde la función indicatriz

$$\mathbf{1}_j(y) = \begin{cases} 1 & y = j \\ 0 & y \neq j \end{cases}$$

Clasificación en kNN



(a) (b)

Regiones de decisión con kNN: (a) dos clases ($k = 3$) y (b) tres clases ($k = 5$). Para evitar empates, el valor de k no debe ser múltiplo del número total de clases.

k-vecinos mas cercanos

Input: $\boldsymbol{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ **con** $\mathbf{x}_i = [x_{i,1}, \ldots, x_{i,d}]^T$, $\mathbf{x}_t = [x_{t,1}, \ldots, x_{t,d}]^T$, k
Output: $\mathbf{s} = [s_1, \ldots, s_n]$

```
1  s_i ← 0,  ∀i = 1, ..., n
2  D_i ← 0,  ∀i = 1, ..., n
3  for (i = 1; i ≤ n; i++) do
4        sum ← 0
5        for (l = 1; l ≤ d; l++) do
6              sum ← sum + (x_{t,l} − x_{i,l})^2
7        end
8        D_i ← √sum
9  end
10 for (j = 1; j ≤ k; j++) do
11       D_mín ← D_1
12       q ← 1
13       for (i = 1; i ≤ n; i++) do
14             if (D_i < D_mín) ∧ (s_i = 0) then
15                   D_mín ← D_i
16                   q ← i
17             end
18       end
19       s_q ← 1
20 end
21 return s
```

La búsqueda k-vecinos más cercanos tiene una complejidad de O(nd+ kn)

Regla del vecino mas cercano, 1NN.

- $\mathcal{D}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ muestra de entrenamiento y $\mathbf{x}'$ el punto de la muestra mas cercano a $\mathbf{x}$.
- la regla del vecino mas cercano clasifica a $\mathbf{x}$ con la etiqueta θ' de $\mathbf{x}'$ la cual es una variable aleatoria tal que $P(\theta' = \omega_i) = P(\omega_i|\mathbf{x}')$.
- Cuando n es grande, $\mathbf{x}'$ va a estar cerca de $\mathbf{x}$ y las probabilidades $P(\omega_i|\mathbf{x}') \sim P(\omega_i|\mathbf{x})$, por lo cual la regla del vecino mas cercano reproduce las probabilidades de los estados naturales.

Regla del vecino mas cercano, 1NN.

- Si definimos $\omega_m(\mathbf{x})$ como la etiqueta que realiza

$$P(\omega_m|\mathbf{x}) = \max_i P(\omega_i|\mathbf{x})$$

es decir, la elección de Bayes, resulta

  - Si $P(\omega_m|\mathbf{x}) \sim 1$, entonces 1NN elije la selección de Bayes.
  - Si $P(\omega_m|\mathbf{x}) \sim 1/c$, esto es, todas las clases son igualmente probables, raramente coinciden 1NN y la selección de Bayes, pero el error que cometen ambas es $1 - 1/c$
  - NN es un procedimiento sub-optimo, usualmente tiene un error peor que el mínimo posible, la tasa de Bayes.

Regla del vecino mas cercano, 1NN.

- El error de Bayes es

$$P^* = \int P^*(e|\mathbf{x})p(\mathbf{x})d\mathbf{x} \quad \text{con} \quad P^*(e|\mathbf{x}) = 1 - P\left(\omega_m|\mathbf{x}\right)$$

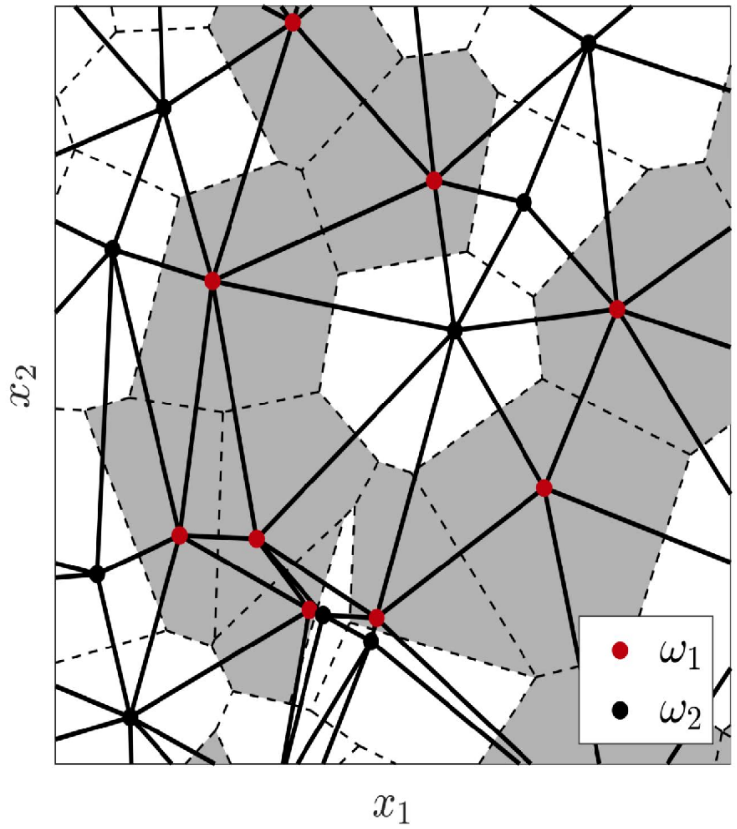
- Si $P_n(e)$ es el error de 1NN en una muestra de tamaño n, y $P = \lim_{n\to\infty} P_n(e)$ entonces

$$P^* \leq P \leq P^*\left(2 - \frac{c}{c-1}P^*\right) \tag{17}$$

Esto es, cuando se tienen infinitas muestras, el error no es peor que dos veces la tasa de Bayes.

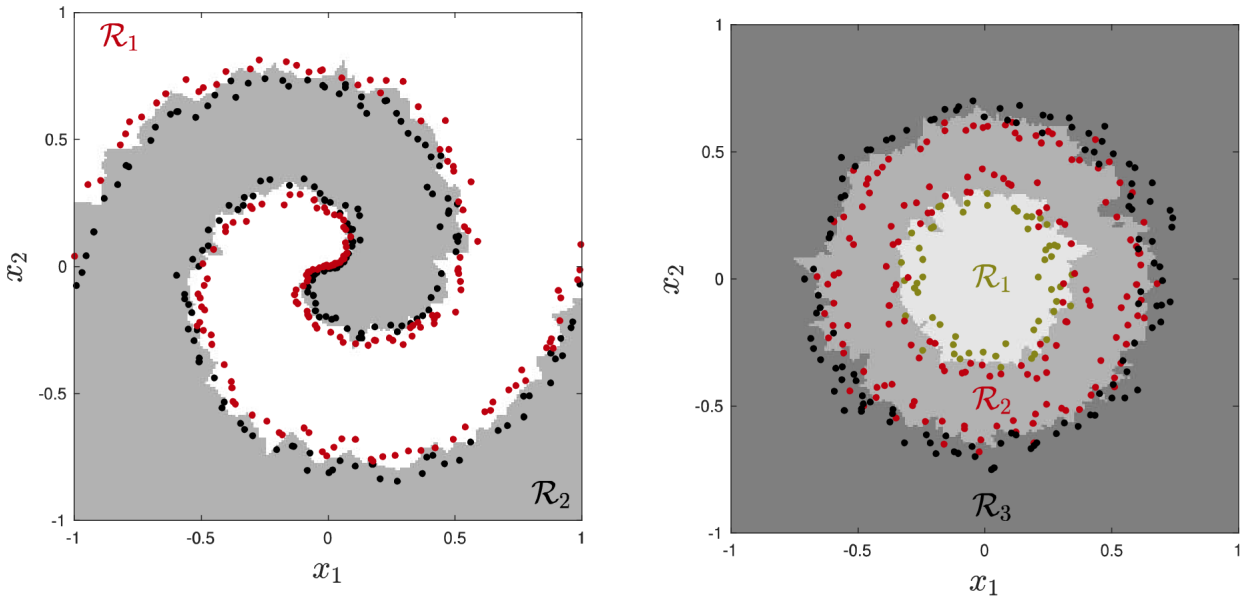
Regla del vecino mas cercano, 1NN.

- La partición de Voronoi tiene celdas definidas por todos los puntos con el mismo vecino mas cercano $\mathbf{x}'$.
- Si dos celdas son adyacentes, entonces sus nodos $\mathbf{x}', \mathbf{X}''$ están conectados con una arista.
- La triangulación de Delaunay es dual al diagrama de Voronoi.



Edición de patrones para reducir complejidad

Un nodo cuyos vecinos de Voronoi sean de la misma clase es eliminado de la muestra de entrenamiento, pues produce la misma clasificación.



(a) (b)

Regiones de decisión con kNN generadas con patrones editados con reducción del: (a) 82 % y (b) 68 %. Se preservan patrones que contribuyen a mantener las fronteras de decisión.

Edición de patrones: algoritmo

Input: $\boldsymbol{Z} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$
Output: $\boldsymbol{Z}'$

1 Computar la triangulación de Delaunay para obtener las aristas de todos los nodos en $\boldsymbol{Z}$
2 $j \leftarrow 0$
3 **do**
4 $\quad j \leftarrow j + 1$
5 $\quad$ Encontrar los nodos adyacentes a $\mathbf{x}_j$
6 $\quad$ **if** cualquier nodo adyacente es de diferente clase que $\mathbf{x}_j$ **then**
7 $\quad\quad$ Marcar $\mathbf{x}_j$
8 $\quad$ **end**
9 **until** $j = n$
10 Remover todos los puntos que no fueron marcados para obtener el conjunto editado $\boldsymbol{Z}'$

Métricas

- El algoritmo kNN depende de una métrica o función de distancia entre patrones.
- Una métrica debe satisfacer cuatro propiedades para todos los vectores $\mathbf{a}$, $\mathbf{b}$ y $\mathbf{c}$:
  - No negatividad: $D(\mathbf{a}, \mathbf{b}) \geq 0$
  - Reflexividad: $D(\mathbf{a}, \mathbf{b}) = 0$ si y sólo si $\mathbf{a} = \mathbf{b}$
  - Simetría: $D(\mathbf{a}, \mathbf{b}) = D(\mathbf{b}, \mathbf{a})$
  - Desigualdad triangular: $D(\mathbf{a}, \mathbf{b}) + D(\mathbf{b}, \mathbf{c}) \geq D(\mathbf{a}, \mathbf{c})$

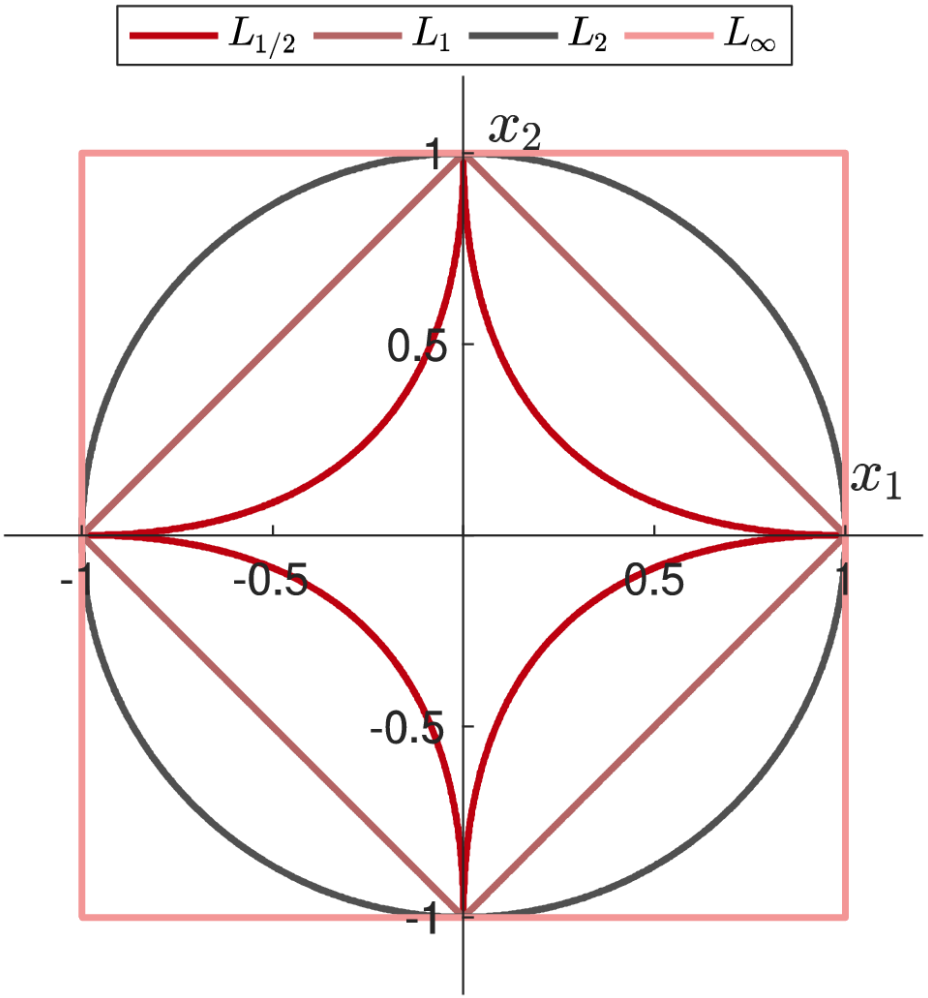
Métricas

- Una clase general de métricas para patrones d-dimensionales es la métrica Minkowski (o norma L_q):

$$L_q(\mathbf{a}, \mathbf{b}) = \left(\sum_{j=1}^{d} |a_j - b_j|^q \right)^{1/q} \tag{18}$$

- Casos particulares: L_1, distancia Manhattan o city block; L_2, distancia Euclidiana; y L_∞, distancia Chebyshev
- Para la búsqueda de los k vecinos más cercanos en espacios de alta dimensionalidad se recomiendan utilizar métricas fraccionales, esto es, con $q \in (0, 1)$.

Métricas



Curvas que consisten de puntos a distancia 1.0 a partir del origen.

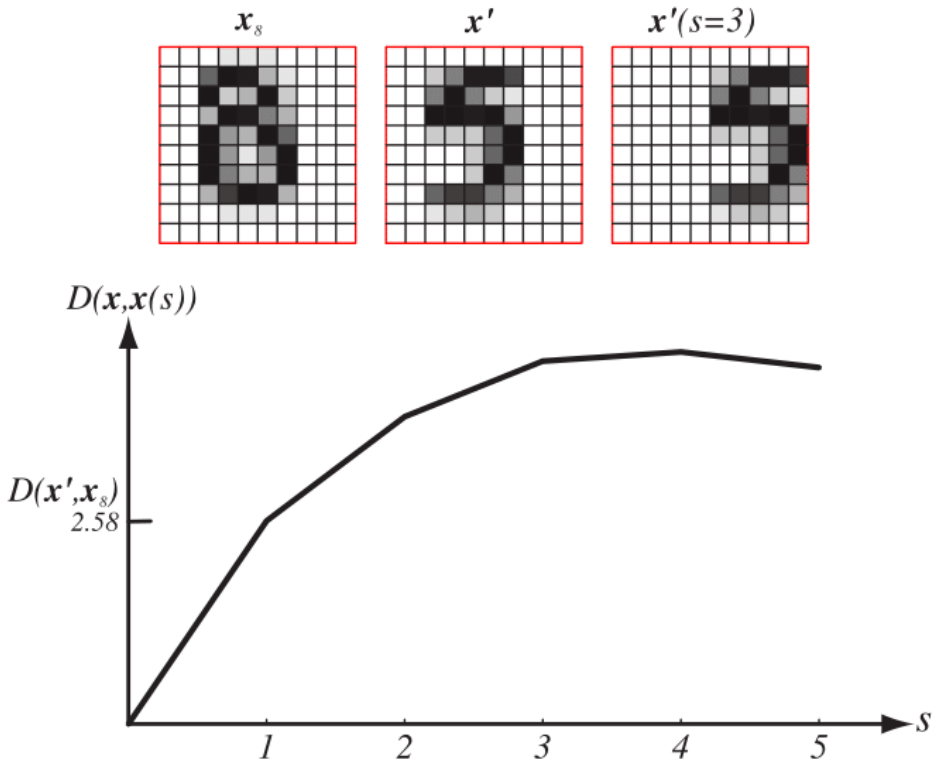
Métricas

- k-vecinos mas cercanos precisan métricas invariantes, las cuales no son universales
- Cada problema tiene transformaciones asociadas, a las cuales la medida debe ser invariante
- Transformaciones comunes son:
  - rotaciones
  - traslaciones
  - distorsiones
  - escalado

Soluciones

- Se pueden procesar los datos para ponerlos en pose
- Esto es muy costoso
- En imágenes, se realiza cuando hay que fusionar datos para determinar cambios o generar tres dimensiones
- Para clasificación, es indispensable trabajar con una distancia invariante

Problemas con la distancia euclídea



Patron x' es una imagen de 100 pixels (10x10) de un número 5. El patrón x_8 presenta el número 8 y el patrón $x'(s = 3)$ es el patrón x' movido 3 pixeles horizontalmente hacia la derecha. Distancia euclídea entre x' y x_8 es menor que entre x' y $x'(s = 3)$.

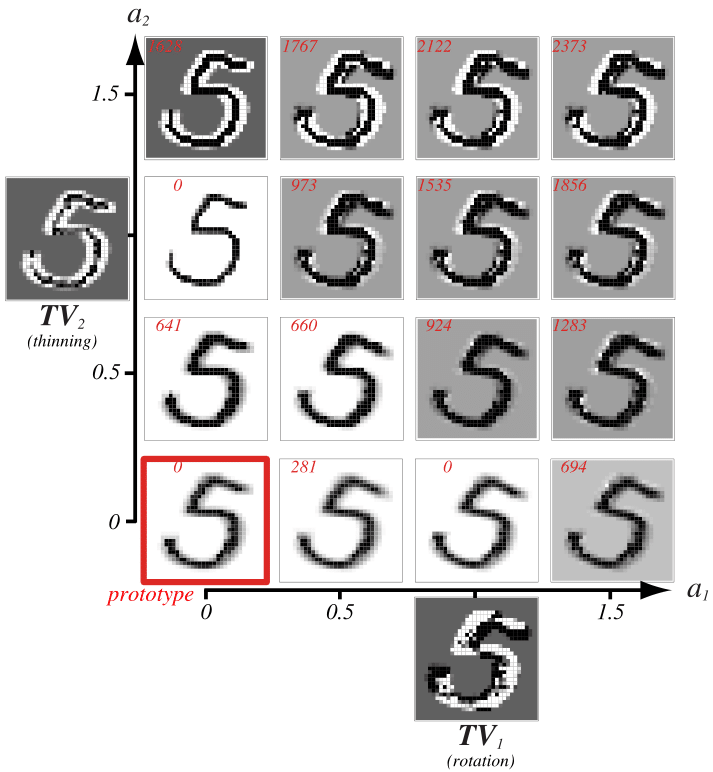
Clasificador basado en distancia tangente

- Se supone que se conocen r transformaciones para las cuales se tiene que obtener invariancia
- Para cada vector de entrenamiento $\mathbf{x}'$, se calcula $\mathcal{F}_i\left(\mathbf{x}'; \alpha_i\right)$ la transformación de x' (para cada transformación diferente)
- Se construye el vector tangente TV de cada transformación

$$\mathbf{TV}_i = \mathcal{F}_i\left(\mathbf{x}'; \alpha_i\right) - \mathbf{x}' \tag{19}$$

- El subespacio de los r vectores tangentes que pasan por $\mathbf{x}$ representa una aproximación lineal de la combinación de las transformaciones.

Distancia tangente: ejemplo



El prototipo 5 se transforma usando rotación y adelgazamiento. Cada una de las 16 imágenes son combinaciones lineales de los dos vectores tangentes.

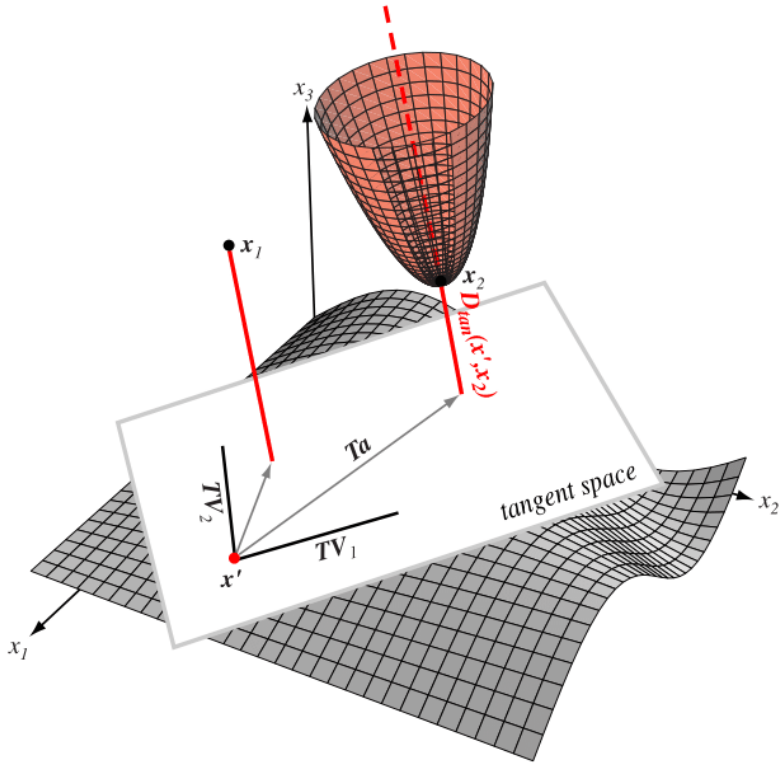
Clasificador basado en distancia tangente

- Cada punto en el subespacio generado por los r vectores tangentes que pasan por $\mathbf{x}'$ representa una aproximacion lineal a la combinacion completa de las transformaciones.
- En la clasificacion de un punto $\mathbf{x}$, se busca el punto del espacio tangente que es mas cercano, la aproximacion lineal ideal.
- La distancia tangente entre $\mathbf{x}$ y el prototipo $\mathbf{x}'$ involucra la matriz de$\mathbf{T}$ de los r vectores tangentes en $\mathbf{x}'$ y se define como

$$D_{tan}\left(\mathbf{x}', \mathbf{x}\right) = \min_{\mathbf{a}} \left[\|(\mathbf{x}' + \mathbf{T}\mathbf{a}) - \mathbf{x}\|\right] \tag{20}$$

- En la clasificacion mediante k-NN, se encuentra la distancia tangente optimizando el valor de $\mathbf{a}$ requerido lo cual es simple dado que la distancia euclidea es una funcion cuadratica en $\mathbf{a}$ y puede usarse descenso por el gradiente o algún método matricial.

Distancia tangente



La distancia euclidea entre el prototipo $\mathbf{x}'$ y $\mathbf{x_1}$ es menor que la distancia entre el prototipo y $\mathbf{x_2}$, sin embargo con la distancia tangente es al revez. La distancia euclidea entre $\mathbf{x_2}$ y el espacio tangente es una funcion cuadratica en $\mathbf{a}$ y se muestra en rosa