Méthodes à base de voisinage

- Idée
  - trouver des points d'apprentissage similaires au point de test
  - faire "voter" ces "voisins"
- Deux stratégies
  - nombre de voisins fixe $\rightarrow$ k-plus proches voisins (k-PPV)
  - voisinage fixe $\rightarrow$ fenêtres de Parzen

Méthodes à base de voisinage

- Terminologie/notation
  - données d'entraînement: $D_n = \big((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\big)$
  - observation: $\mathbf{x}_i \in \mathbb{R}^d$
  - étiquette/classe: $y_i \in \{-1, 1\}$
  - fonction discriminante: $g : \mathbb{R}^d \mapsto \mathbb{R}$, souvent $g : \mathbb{R}^d \mapsto [-1, 1]$
  - fonction de classification/classifieur: $f : \mathbb{R}^d \mapsto \{-1, 1\}$
  - fonction discriminante $\rightarrow$ classifieur:

$$f(\mathbf{x}) = \begin{cases} 1, \text{ si } g(\mathbf{x}) \geq 0 \\ -1, \text{ si } g(\mathbf{x}) < 0 \end{cases}$$

Méthodes à base de voisinage

- Vote des voisins formellement:

$$g(\mathbf{x}) = \frac{1}{n} \sum_{\mathbf{x}_i \in V(\mathbf{x})} y_i$$

  - k-PPV:

    $V(\mathbf{x})$ est l'ensemble des k points plus proches à $\mathbf{x}$ dans D_n

  - Parzen (avec paramètre h):

$$V(\mathbf{x}) = \{\mathbf{x}_i : d(\mathbf{x}_i, \mathbf{x}) < h\}$$

Méthodes à base de voisinage

- Erreur d'entraînement (risque empirique)

$$\widehat{R}(f, D_n) = \widehat{R}(f) = \frac{1}{n} \sum_{i=1}^{n} I_{\{f(\mathbf{x}_i) \neq y_i\}}$$

  - fonction indicatrice: $I_{\{A\}} = \begin{cases} 1, \text{ si } A \text{ est vrai} \\ 0, \text{ sinon} \end{cases}$

- Comment choisir k ou h?
  - minimiser $\widehat{R}(f)$?
  - $k = 1$, $h \rightarrow 0$

Méthodes à base de voisinage

- But: généralisation!
  - k ou h petit: les "électeurs" sont proches (donc fiables) mais pas nombreux (donc le vote est bruité)
  - k ou h grand: les "électeurs" sont nombreux (donc les fluctuations statistiques sont lissées) mais loin (donc moins fiables)
- Comment mesurer la généralisation?
  - sur un ensemble de test: $D'_m = \left((\mathbf{x}'_1, y'_1), \ldots, (\mathbf{x}'_m, y'_m)\right)$
- Erreur de test

$$\widehat{R}(f, D'_m) = \widehat{R}'(f) = \frac{1}{m}\sum_{i=1}^{m} I_{\{f(\mathbf{x}'_i) \neq y'_i\}}$$

Méthodes à base de voisinage

- Courbes d'apprentissage
  - erreurs d'entraînement et de test en terme du paramètre de complexité/capacité
- Fléau de la dimensionnalité
  - les espaces de haute dimension sont presque vides: on a besoin de $O(c^d)$ points pour la même densité
  - les voisins plus proches sont loin
  - les méthodes à base de voisinage "global" s'écroulent

Fenêtres de Parzen

- Vote des voisins formellement:

$$g(\mathbf{x}) = \frac{1}{n}\sum_{d(\mathbf{x}_i,\mathbf{x})<h} y_i = \frac{1}{n}\sum_{i=1}^{n} I_{\{d(\mathbf{x}_i,\mathbf{x})<h\}} y_i = \frac{1}{n}\sum_{i=1}^{n} I_{\left\{\frac{d(\mathbf{x}_i,\mathbf{x})}{h}<1\right\}} y_i$$

  - remplacer $I_{\{\cdot\}}$ par une fonction "lisse":

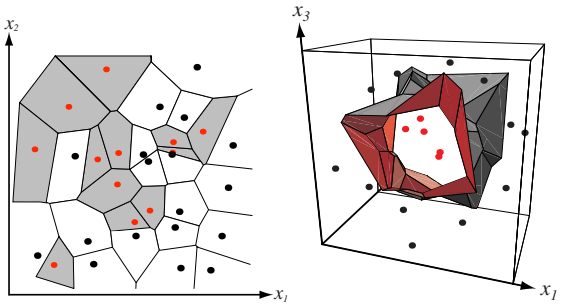
$$g(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} \phi\left(\frac{d(\mathbf{x}_i,\mathbf{x})}{h}\right) y_i$$

  - par exemple, gaussien standard $N(0,1)$:

$$\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$$

k-plus-proche-voisin

- Partition de Voronoi



k-plus-proche-voisin

- Complexité computationnelle
  - méthode naïve: $T(n,k,d) = O(nkd) = O(n^2 d)$
  - méthode de distances partielles:

$$d_r(\mathbf{a},\mathbf{b}) = \left(\sum_{i=1}^{r} (a_i - b_i)^2 \right)^{1/2}, r \leq d$$

  - méthodes d'arbre de recherche

k-plus-proche-voisin

- Complexité computationnelle
  - méthode de suppression/émondage (editing/pruning/condensing)

```
ÉMONDAGEDEPLUSPROCHEVOISIN(D_n)
1  construire le diagramme de Voronoi complet de D_n
2  pour j ← 1 à n faire
3      pour tout les voisins de Voronoi x' de x_i faire
4          si y_i ≠ y' alors
5              marquer x_i
6  pour j ← 0 à n faire
7      si x_i n'est pas marqué alors
8          supprimer x_i
```

  - $T(n,d) = O(d^3 n^{\lfloor d/2 \rfloor} \ln n)$

Métriques

- Propriétés d'une métrique
  - positivité: $d(\mathbf{a},\mathbf{b}) \geq 0$
  - réflexivité: $d(\mathbf{a},\mathbf{a}) = 0$
  - symétrie: $d(\mathbf{a},\mathbf{b}) = d(\mathbf{b},\mathbf{a})$
  - inégalité de triangle: $d(\mathbf{a},\mathbf{b}) + d(\mathbf{b},\mathbf{c}) \geq d(\mathbf{a},\mathbf{c})$

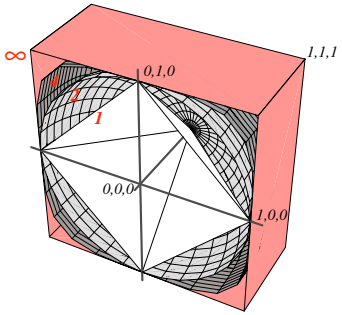
Métriques

- Exemples des métriques

euclidienne	L_2	$d(\mathbf{a},\mathbf{b}) = \left(\sum_{i=1}^{d} (a_i - b_i)^2 \right)^{1/2}$
Manhattan	L_1	$d(\mathbf{a},\mathbf{b}) = \sum_{i=1}^{d} \lvert a_i - b_i \rvert$
	L_∞	$d(\mathbf{a},\mathbf{b}) = \max_i \lvert a_i - b_i \rvert$
Minkowski	L_p	$d(\mathbf{a},\mathbf{b}) = \left(\sum_{i=1}^{d} \lvert a_i - b_i \rvert^p \right)^{1/p}$
Tanimoto	L_{Tanimoto}	$d(S_1,S_2) = \frac{\lvert S_1 \rvert + \lvert S_2 \rvert - 2\lvert S_1 \cap S_2 \rvert}{\lvert S_1 \rvert + \lvert S_2 \rvert - \lvert S_1 \cap S_2 \rvert}$

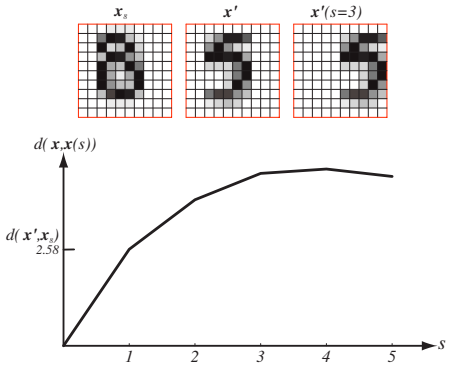
Métriques

- La métrique de Minkowski



Métriques

- Les limitations de la métrique euclidienne

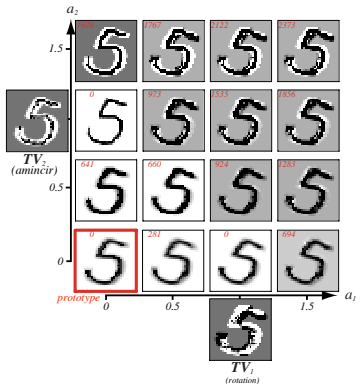


La distance tangente

- Capturer l'invariance de certaines transformations:

$$\mathbf{TV}_i = F_i(\mathbf{x}';a_i) - \mathbf{x}'$$

La distance tangente



La distance tangente

$$d_{tan}(\mathbf{x}', \mathbf{x}) = \min_{\mathbf{a}}[\|(\mathbf{x}' + \mathbf{Ta}) - \mathbf{x}\|]$$