Nearest Neighbor Classification

nanopoulos@ismll.de

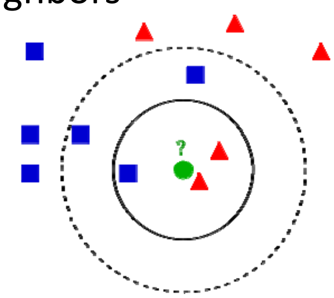
1

Outline

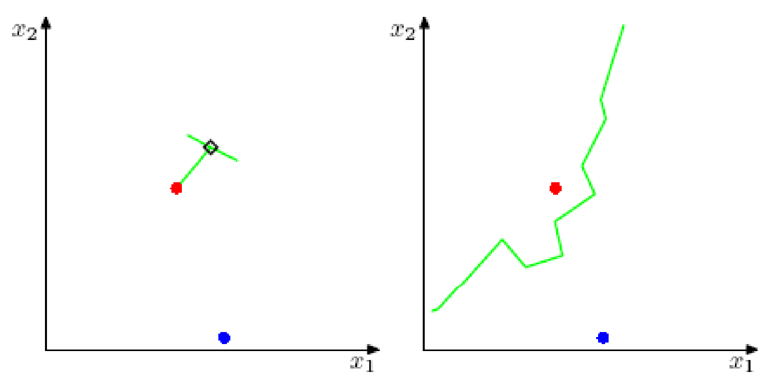
- Nearest neighbor classification
- Distance measures
- Error of 1-NN classification
- Dimensionality curse

k-nearest neighbors classification

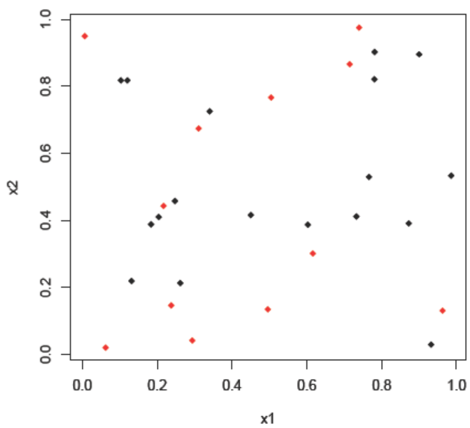
- An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its k nearest neighbors



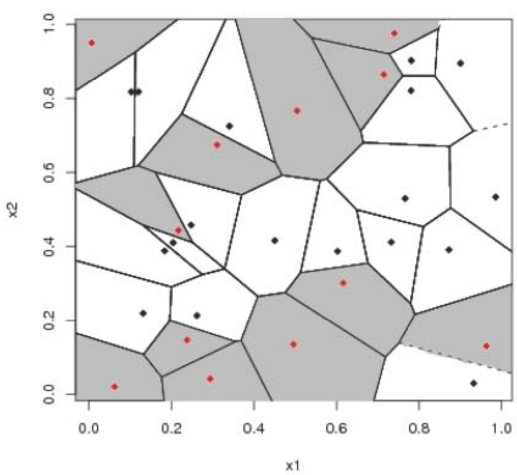
Decision boundaries



Voronoi diagram



Voronoi diagram



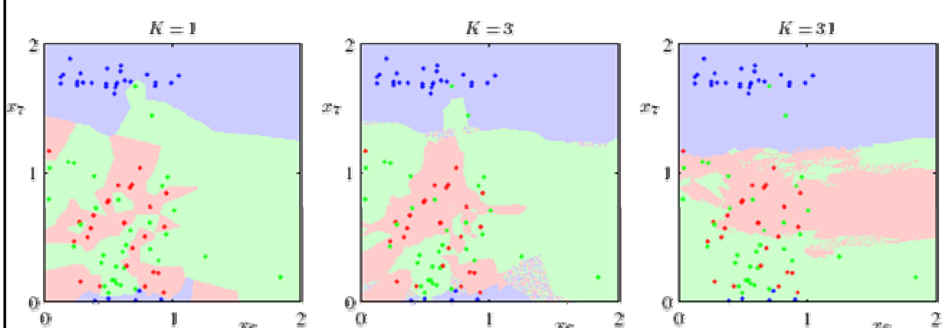
Characteristics of k-NN classification

- Amongst the simplest of all machine learning algorithms
- k is a positive integer, typically small
- If k = 1, then the object is simply assigned to the class of its nearest neighbor
- The training phase of the algorithm consists only of storing the feature vectors and class labels of the training samples (lazy classifier)

7

How to select k?

- Larger values of k reduce the effect of noise, but make boundaries between classes less distinct



8

Outline

- Nearest neighbor classification
- Distance measures
- Error of 1-NN classification
- Dimensionality curse

Distance measures

Let d be a **distance measure** (also called **metric**) on a set $\mathcal{X}$, i.e.,

$$d : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}_0^+$$

with

1. d is **positiv definite**: $d(x, y) \geq 0$ and $d(x, y) = 0 \Leftrightarrow x = y$
2. d is **symmetric**: $d(x, y) = d(y, x)$
3. d is **subadditive**: $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality)

(for all $x, y, z \in \mathcal{X}$.)

Example: **Euclidean metric** on $\mathcal{X} := \mathbb{R}^n$:

$$d(x, y) := (\sum_{i=1}^{n} (x_i - y_i)^2)^{\frac{1}{2}}$$

10

Minkowski metric

Minkowski Metric / L_p **metric** on $\mathcal{X} := \mathbb{R}^n$:

$$d(x, y) := (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

with $p \in \mathbb{R}, p \geq 1$.

$p = 1$ **(taxicab distance; Manhattan distance):**

$$d(x, y) := \sum_{i=1}^{n} |x_i - y_i|$$

$p = 2$ **(euclidean distance):**

$$d(x, y) := (\sum_{i=1}^{n} (x_i - y_i)^2)^{\frac{1}{2}}$$

$p = \infty$ **(maximum distance; Chebyshev distance):**

$$d(x, y) := \max_{i=1}^{n} |x_i - y_i|$$

11

Example

Example:

$$x := \begin{pmatrix} 1 \\ 3 \\ 4 \end{pmatrix}, \quad y := \begin{pmatrix} 2 \\ 4 \\ 1 \end{pmatrix}$$

$$d_{L_1}(x, y) = |1 - 2| + |3 - 4| + |4 - 1| = 1 + 1 + 3 = 5$$

$$d_{L_2}(x, y) = \sqrt{(1-2)^2 + (3-4)^2 + (4-1)^2} = \sqrt{1 + 1 + 9} = \sqrt{11} \approx 3.32$$

$$d_{L_\infty}(x, y) = \max\{|1 - 2|, |3 - 4|, |4 - 1|\} = \max\{1, 1, 3\} = 3$$

12

Distances for sets

For set-valued variables (which values are subsets of a set A) the **Hamming distance** often is used:

$$d(x, y) := |(x \setminus y) \cup (y \setminus x)| = |\{a \in A \,|\, I(a \in x) \neq I(a \in y)\}|$$

(the number of elements contained in only one of the two sets).

Example:

$$d(\{a, e, p, l\}, \{a, b, n\}) = 5, \quad d(\{a, e, p, l\}, \{a, e, g, n, o, r\}) = 6$$

Also often used is the similarity measure **Jaccard coefficient**:

$$\text{sim}(x, y) := \frac{|x \cap y|}{|x \cup y|}$$

Example:

$$\text{sim}(\{a, e, p, l\}, \{a, b, n\}) = \frac{1}{6}, \quad \text{sim}(\{a, e, p, l\}, \{a, e, g, n, o, r\}) = \frac{2}{8}$$

Distances for strings

edit distance / **Levenshtein distance**:

$d(x, y) :=$ minimal number of deletions, insertions or substitions to transform x in y

Examples:

$$d(\text{man}, \text{men}) = 1$$
$$d(\text{house}, \text{spouse}) = 2$$

$$d(\text{order}, \text{express order}) = 8$$

Outline

- Nearest neighbor classification
- Distance measures
- Error of 1-NN classification
- Dimensionality curse

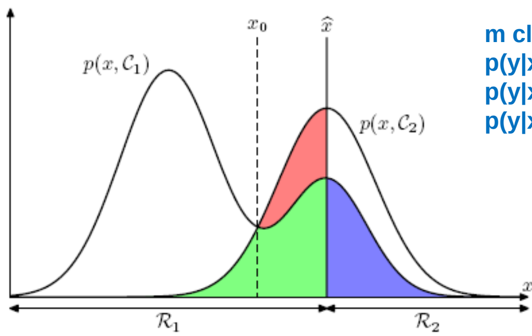
Theorem for 1-NN classification

- **Theorem**: For sufficiently large training set size n, the error rate of the 1-NN classifier is less than twice the Bayes error rate
- Guarantees for error!

16

Bayes error rate

- The error prob is minimized if each x is assigned to the class $y^*(x) := \text{argmax}_{y \in \mathcal{Y}}\, p(y \mid x)$



m classes
p(y|x) known (optimal case)
p(y|x) = 1 => p(error) = 0
p(y|x) = 1/m => p(error) = (m-1)/m

Proving the theorem for 1-NN

$$E^* = \int_x p(x)[1 - \max_i p(i|x)]$$

Expected Bayes (optimal) error

Let *x*' = 1-NN *x*. For each *x* the error of 1-NN about class i is:
p(*i*|*x*) [1-p(*i*|*x*')]

x disagrees with *x*'

if *n* → ∞ => *p*(*i*|*x*) = *p*(*i*|*x*')

Critical assumption

Expected 1-NN error for each *x*:

$$\sum_{i=1}^{m} p(i|x)[1 - p(i|x)]$$

Proving the theorem for 1-NN

Expected 1-NN error for each *x*:

$$\sum_{i=1}^{m} p(i|x)[1 - p(i|x)]$$

Expected Bayes error for all *x*:

$$E^* = \int_x p(x)[1 - \max_i p(i|x)]$$

We need to show that:

$$\sum_{i=1}^{m} p(i|x)[1 - p(i|x)] \le 2[1 - \max_i p(i|x)]$$

19

Proving the theorem for 1-NN

$\max_i p(i|x) = r$ Attained when for *i* = *j*

$$\sum_{i=1}^{m} p(i|x)[1 - p(i|x)] = r(1-r) + \sum_{i \neq j} p(i|x)[1 - p(i|x)]$$ Left hand

$$2[1 - \max_i p(i|x)] = 2(1-r)$$ Right hand

We need to show that:

$$r(1-r) + \sum_{i \neq j} p(i|x)[1 - p(i|x)] \le 2(1-r)$$

20

Proving the theorem for 1-NN

$\sum_{i \neq j} p(i|x)[1 - p(i|x)]$ is maximum when all p(i|x) are equal for all $i \neq j$

For m classes this means that all $i \neq j$ p(i|x) = (1-r)/(m-1)

$$\sum_{i \neq j} p(i|x)[1 - p(i|x)] = (m-1)\frac{1-r}{m-1}\frac{m-1-(1-r)}{m-1}$$

$$r(1-r) + \sum_{i \neq j} p(i|x)[1 - p(i|x)] = r(1-r) + (m-1)\frac{1-r}{m-1}\frac{m-1-(1-r)}{m-1}$$

$$= r(1-r) + (1-r)\frac{m+r-2}{m-1}$$

Proving the theorem for 1-NN

We need to show that

$$r(1-r) + (1-r)\frac{m+r-2}{m-1} \leq 2(1-r)$$

This holds because:

$r \leq 1$

$m - 2 + r < m - 1$

QED

Implications of the theorem

- with a large enough training set, no classifier can do better than half the error rate of a 1NN classifier $E^* \geq E/2$
- Estimate a lower bound for the Bayes error rate by measuring the error rate of a 1NN classifier, then dividing by two
- True regardless of which distance metric is used
- Be careful: For finite sample sizes, not true!

23

Outline

- Nearest neighbor classification
- Distance measures
- Error of 1-NN classification
- Dimensionality curse

Dimensionality curse

x is d-dimensional

For high d, it is hard to find meaningful nearest neighbors

Let's see why

d-dimensional hypersphere

Volume of hypersphere in d dimensions

$$V(B_1(r)) = 2r$$
$$V(B_2(r)) = \pi r^2$$
$$V(B_3(r)) = \frac{4}{3}\pi r^3$$

$$V(B_d(r)) = K_d r^d$$

$$K_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$$

$$\Gamma(\frac{d}{2}+1) = \begin{cases} (\frac{d}{2})! & \text{if } d \text{ is even} \\ \sqrt{\pi}\frac{d!!}{2^{\frac{n+1}{2}}} & \text{if } d \text{ is odd} \end{cases}$$

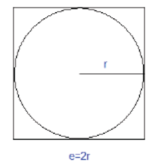
$$n!! = \begin{cases} 1 & n = 0, 1 \\ n(n-2)!! & n \geq 2 \end{cases}$$

Inscribe a Hypersphere Inside a Hypercube

$$\frac{V(B_d(r))}{V(H_d(2r))}$$

In 2 Dimensions

$$\frac{V(B_2(r))}{V(H_2(2r))} = \frac{\pi r^2}{4r^2} = \frac{\pi}{4} \approx 75\%$$



In 3 Dimensions

$$\frac{V(B_3(r))}{V(H_3(2r))} = \frac{\frac{4}{3}\pi r^3}{8r^3} = \frac{\pi}{6} \approx 50\%$$

In d Dimensions

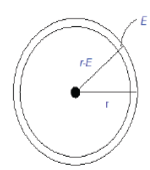
$$\lim_{d\to\infty} \frac{V(B_d(r))}{V(H_d(2r))} = \lim_{d\to\infty} \frac{K_d r^d}{2^d r^d} = \lim_{d\to\infty} \frac{K_d}{2^d} = \lim_{d\to\infty} \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1) * 2^d} = 0$$

In other words, a query for all results a certain distance from a given point will return no results as the number of dimensions approaches infinity.

How Many Points Lie in a Hypersphere?

In 2 dimensions

$$\frac{V(B_2(r-\varepsilon))}{V(B_2(r))} = \frac{\pi(r-\varepsilon)^2}{\pi(r)^2} = \frac{r^2 - 2r\varepsilon + \varepsilon^2}{r^2}$$



For a unit circle and $\varepsilon = 0.01$ the equation becomes:

$$1 - 0.02 + 0.01^2 = .9801 \approx 1$$

Ratio between $V(B_d(r-\varepsilon))$ to $V(B_d(r))$ for small ε

Generalized to d dimensions $\frac{\Delta V_d(r,\varepsilon)}{V(B_d(r))} = 1 - (1 - \frac{\varepsilon}{r})^d$ $\lim_{d\to\infty} \frac{\Delta V_d(r,\varepsilon)}{V(B_d(r))} = 1$

Thus, all results within a distance r of a point end up lying on the outer edge of the hypersphere $B_d(r)$ as $d \to \infty$.

Implications for k-NN classification

- As $d \rightarrow \infty$, the ratio of the nearest neighbor to the farthest neighbor from a given point approaches 1.
- This means that it becomes much more difficult to distinguish which point is nearest and which is farthest from a given point.