

Announcements

- HW2 posted.
- Most of last two lecture was on the blackboard.

CSE190, Winter 2011

Non-Parametric Density Estimation

- Given a collection of n samples, estimate the probability density.
 - Parzen Windows
 - K-th nearest neighbor
- Main ideas:
 1. As number of samples (n) approaches infinity, estimated density should approach true density
 2. Approximated density should be “reasonable” for finite n .

CSE190, Winter 2011

3

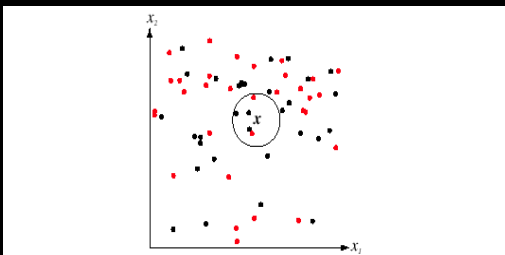


FIGURE 4.15. The k -nearest-neighbor query starts at the test point x and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point x would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

CSE190, Winter 2011

4

– The k – nearest-neighbor rule

- **Goal:** Classify x by assigning it the label most frequently represented among the k nearest samples and use a voting scheme

CSE190, Winter 2011

Pattern Classification, Ch4 (Part 1)

5

The nearest –neighbor rule

- Let $D_n = \{x_1, x_2, \dots, x_n\}$ be a set of n labeled prototypes
- Let $x' \in D_n$ be the closest prototype to a test point x then the nearest-neighbor rule for classifying x is to assign it the label associated with x'
- The nearest-neighbor rule leads to an error rate greater than the minimum possible: the Bayes rate
- If the number of prototype is large (unlimited), the error rate of the nearest-neighbor classifier is never worse than twice the Bayes rate (it can be demonstrated!)
- If $n \rightarrow \infty$, it is always possible to find x' sufficiently close so that:

$$P(\omega_j | x') \approx P(\omega_j | x)$$

CSE190, Winter 2011

6

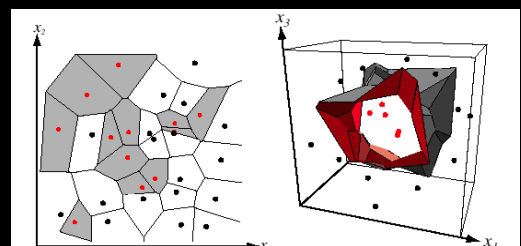


FIGURE 4.13. In two dimensions, the nearest-neighbor algorithm leads to a partitioning of the input space into Voronoi cells, each labeled by the category of the training point it contains. In three dimensions, the cells are three-dimensional, and the decision boundary resembles the surface of a crystal. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

CSE190, Winter 2011

Pattern Classification, Ch4 (Part 1)

Whitening Transform

See blackboard

CSE190, Winter 2011

Linear Discriminant Functions (Sections 5.1-5.2)

CSE190, Winter 2011

Linear Discriminant Functions (Sections 5.1-5.2)

•

Linear discriminant functions and decisions surfaces

10

- Definition

It is a function that is a linear combination of the components of x

$$g(x) = w^T x + w_0 \quad (1)$$

where w is the weight vector and w_0 the bias

- A two-category classifier with a discriminant function of the form (1) uses the following rule:
Decide ω_1 if $g(x) > 0$ and ω_2 if $g(x) < 0$
 $\Leftrightarrow$ Decide ω_1 if $w^T x > -w_0$ and ω_2 otherwise
If $g(x) = 0 \Rightarrow x$ is assigned to either class

Pattern Classification, Ch4 (Part 1)

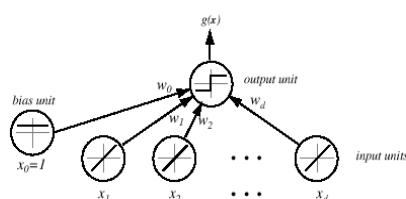


FIGURE 5.1. A simple linear classifier having d input units, each corresponding to the values of the components of an input vector. Each input feature value x_i is multiplied by its corresponding weight w_i ; the effective input at the output unit is the sum of all these products, $\sum w_i x_i$. We show in each unit its effective input-output function. Thus each of the d input units is linear, emitting exactly the value of its corresponding feature value. The single bias unit always emits the constant value 1.0. The single output unit emits a +1 if $w^T x + w_0 > 0$ or a -1 otherwise. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Pattern Classification, Ch4 (Part 1)

- The equation $g(x) = 0$ defines the **decision surface** that separates points assigned to the category ω_1 from points assigned to the category ω_2
- When $g(x)$ is linear, the decision surface is a hyperplane
- Algebraic measure of the distance from x to the hyperplane (interesting result!)

Pattern Classification, Ch4 (Part 1)

13

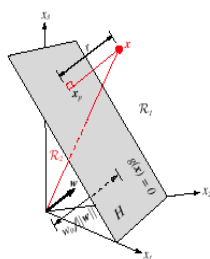


FIGURE 5.2. The linear decision boundary H , where $g(x) = w^T x + w_0 = 0$, separates the feature space into two half-spaces $\mathcal{R}_1$ (where $g(x) > 0$) and $\mathcal{R}_2$ (where $g(x) < 0$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Pattern Classification, Ch4 (Part 1)

14

$$x = x_p + \frac{r \cdot w}{\|w\|} \quad (\text{since } w \text{ is colinear with } x - x_p \text{ and } \frac{w}{\|w\|} = 1)$$

$$\text{since } g(x) = 0 \text{ and } w^T \cdot w = \|w\|^2$$

$$\text{therefore } r = \frac{g(x)}{\|w\|}$$

$$\text{in particular } d(0, H) = \frac{w_0}{\|w\|}$$

- In conclusion, a linear discriminant function divides the feature space by a hyperplane decision surface
- The orientation of the surface is determined by the normal vector w and the location of the surface is determined by the bias

Pattern Classification, Ch4 (Part 1)

15

• The multi-category case

- We define c linear discriminant functions

$$g_i(x) = w_i^T x + w_{i0} \quad i = 1, \dots, c$$

and assign x to ω_i if $g_i(x) > g_j(x) \forall j \neq i$; in case of ties, the classification is undefined

- In this case, the classifier is a "linear machine"
- A linear machine divides the feature space into c decision regions, with $g_i(x)$ being the largest discriminant if x is in the region $\mathcal{R}_i$
- For a two contiguous regions $\mathcal{R}_i$ and $\mathcal{R}_j$, the boundary that separates them is a portion of hyperplane H_{ij} defined by:

$$g_i(x) = g_j(x)$$

$$\Leftrightarrow (w_i - w_j)^T x + (w_{i0} - w_{j0}) = 0$$
- $w_i - w_j$ is normal to H_{ij} and

$$d(x, H_{ij}) = \frac{g_i - g_j}{\|w_i - w_j\|}$$

Pattern Classification, Ch4 (Part 1)

16

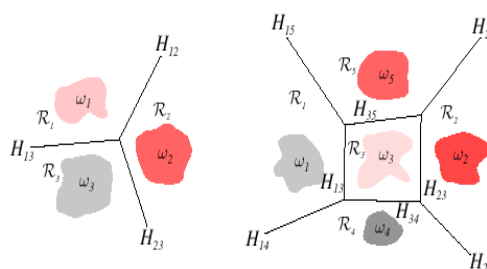


FIGURE 5.4. Decision boundaries produced by a linear machine for a three-class problem and a five-class problem. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

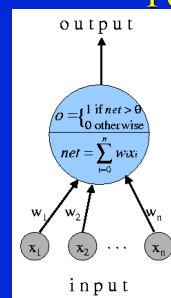
Pattern Classification, Ch4 (Part 1)

17

- It is easy to show that the decision regions for a linear machine are convex, this restriction limits the flexibility and accuracy of the classifier

Pattern Classification, Ch4 (Part 1)

Perceptron



Linear, threshold units

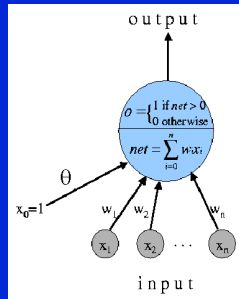
x_i : inputs

w_i : weights

θ : threshold

$$o(x_1, \dots, x_n) = \begin{cases} 1 & \text{if } w_1 x_1 + \dots + w_n x_n > \theta \\ -1 & \text{otherwise.} \end{cases}$$

Perceptron



The threshold can be easily forced to 0 by introducing an additional weight input $w_0 = \theta$

$$o(x_1, \dots, x_n) = \begin{cases} 1 & \text{if } w_0 + w_1x_1 + \dots + w_nx_n > 0 \\ -1 & \text{otherwise.} \end{cases}$$

How powerful is a perceptron?

Threshold = 0

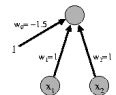
Inverter

input x_1	output
0	1
1	0



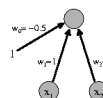
Boolean AND

input x_1	input x_2	output
0	0	0
0	1	0
1	0	0
1	1	1



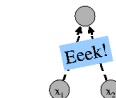
Boolean OR

input x_1	input x_2	output
0	0	0
0	1	1
1	0	1
1	1	1



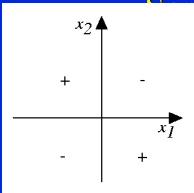
Boolean XOR

input x_1	input x_2	output
0	0	0
0	1	1
1	0	1
1	1	0

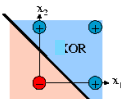


Concept Space & Linear

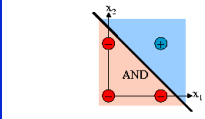
Separability



Linear Separability



Linear Separability



Linear Separability

