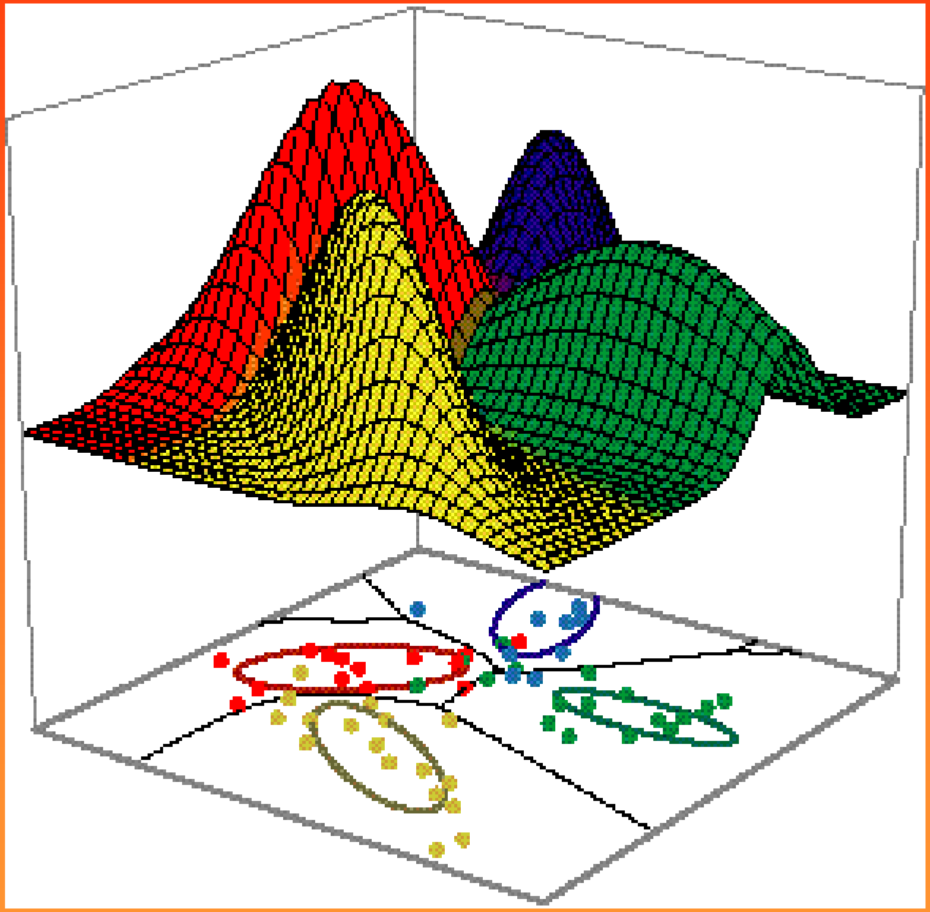


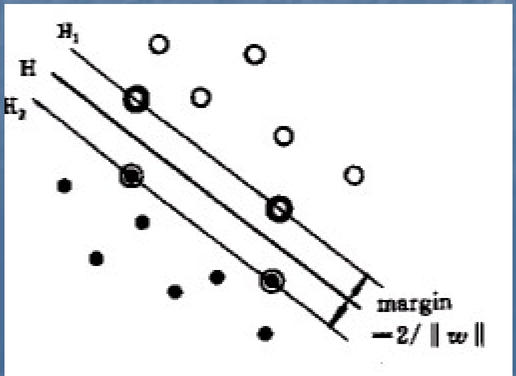
Pattern Classification

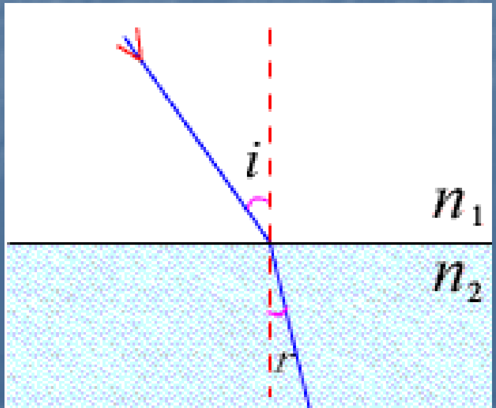
All materials in these slides were taken from
Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000
with the permission of the authors and the publisher

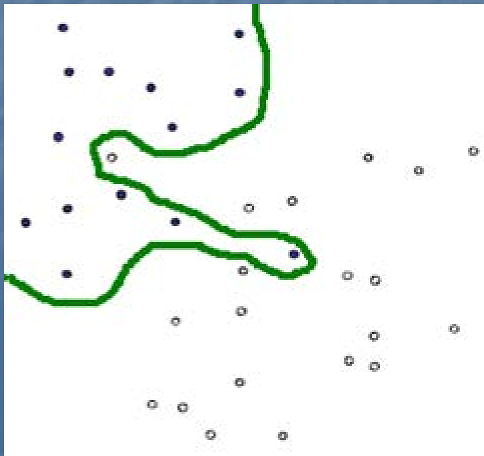
5.4 The Two-Category Linearly Separable Case

- Linearly separable

n samples $y_1, y_2, \ldots y_n$ belong to ω_1, ω_2, if there exits a linear discriminant function $g(\mathrm{x}) = \mathrm{a}^t \mathrm{y}$ that classifies all of them correctly, the samples are said to be linearly separable. Weight vector a is called a separating vector or solution vector







- Normalization

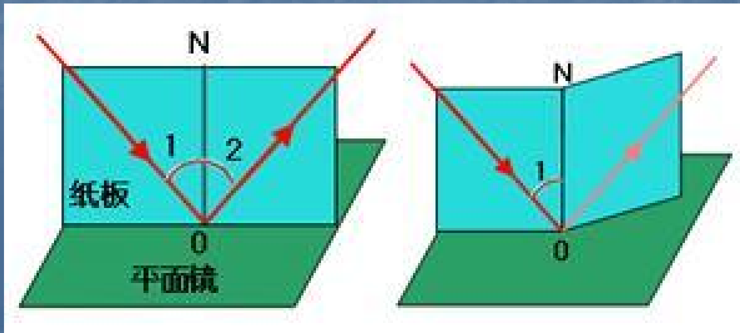
$$for\ y_i \in \omega_1, a^t y_i > 0.$$
$$for\ y_i \in \omega_2, a^t y_i < 0.$$

Multiplying all samples labeled with ω_2 by -1 is called normalization

with this normalization we only need to look for a weight vector a such that $a^t y_i > 0.$

- Solution Region
- Margin $a^t y_i \geq b > 0$

the distance between old boundaries and new boundaries is $b / \|y_i\|$



- The problem of finding a linear discriminant function will be formulated as a problem of minimizing a criterion function

- **Gradient Descent Procedures**

  define a criterion function $J(\mathrm{a})$ that is minimized if a is a solution vector. This can often be solved by a gradient descent procedure.

$$\mathrm{a}(\mathrm{k}+1) = \mathrm{a}(\mathrm{k}) - \eta(\mathrm{k})\nabla J(\mathrm{a}(\mathrm{k}))$$

η is a positive scale factor or learning rate that sets the step size

Using the Taylor extension, we have

$$\eta(k) = \frac{\|\nabla J\|^2}{\nabla J^t H \nabla J}$$

H is Hessian matrix, $\partial^2 J / \partial a_i \partial a_j$

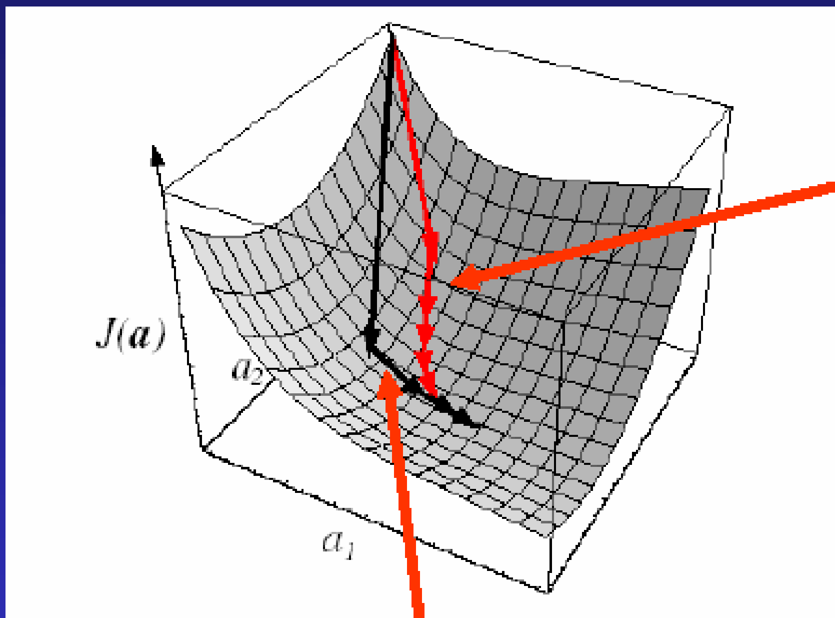
- Another Algorithm
- :Newton Descent Algorithm

Using the Taylor extension and let $\frac{\partial J(a(k+1))}{\partial a(k+1)} = 0$, we can get

$$\mathrm{a}(\mathrm{k}+1) = \mathrm{a}(\mathrm{k}) - \mathrm{H}^{-1}\nabla J(\mathrm{a}(\mathrm{k}))$$

This the so-called Newton Descent Algorithm

- Newton's algorithm vs the simple gradient decent algorithm



Simple gradient descent method

Newton's second order method
Has greater improvement per step even
When using optimal learning rates for both
Method.
However added computational
Burden of inverting the Hessian matrix.

5.5 Minimizing The Perception Criterion Function

$$J_p(\mathrm{a}) = \sum_{\mathrm{y}\in \mathrm{Y}} (-a^t \mathrm{y}), \text{ where } \mathrm{Y}(\mathrm{a}) \text{ is the set of samples misclassified by the Batch Perception Algorithm}$$

- The Perception Criterion Function of the Batch Perception Algorithm

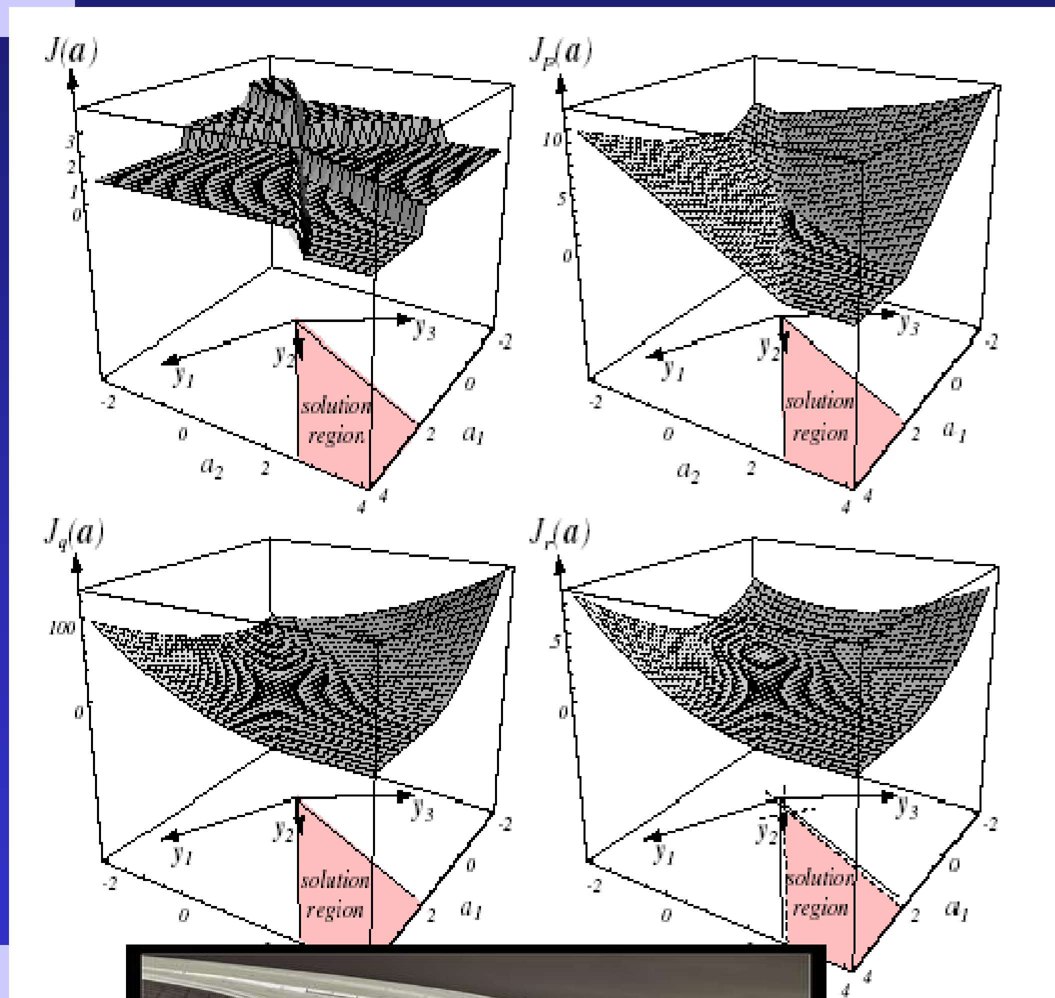
$$\nabla J_p = \sum_{\mathrm{y}\in \mathrm{Y}} -\mathrm{y}$$

$$a(k+1) = a(k) + \eta(k) \sum_{\mathrm{y}\in \mathrm{Y}} \mathrm{y}$$

Comparison of Four Criterion functions

No of misclassified samples:
Piecewise constant,
unacceptable

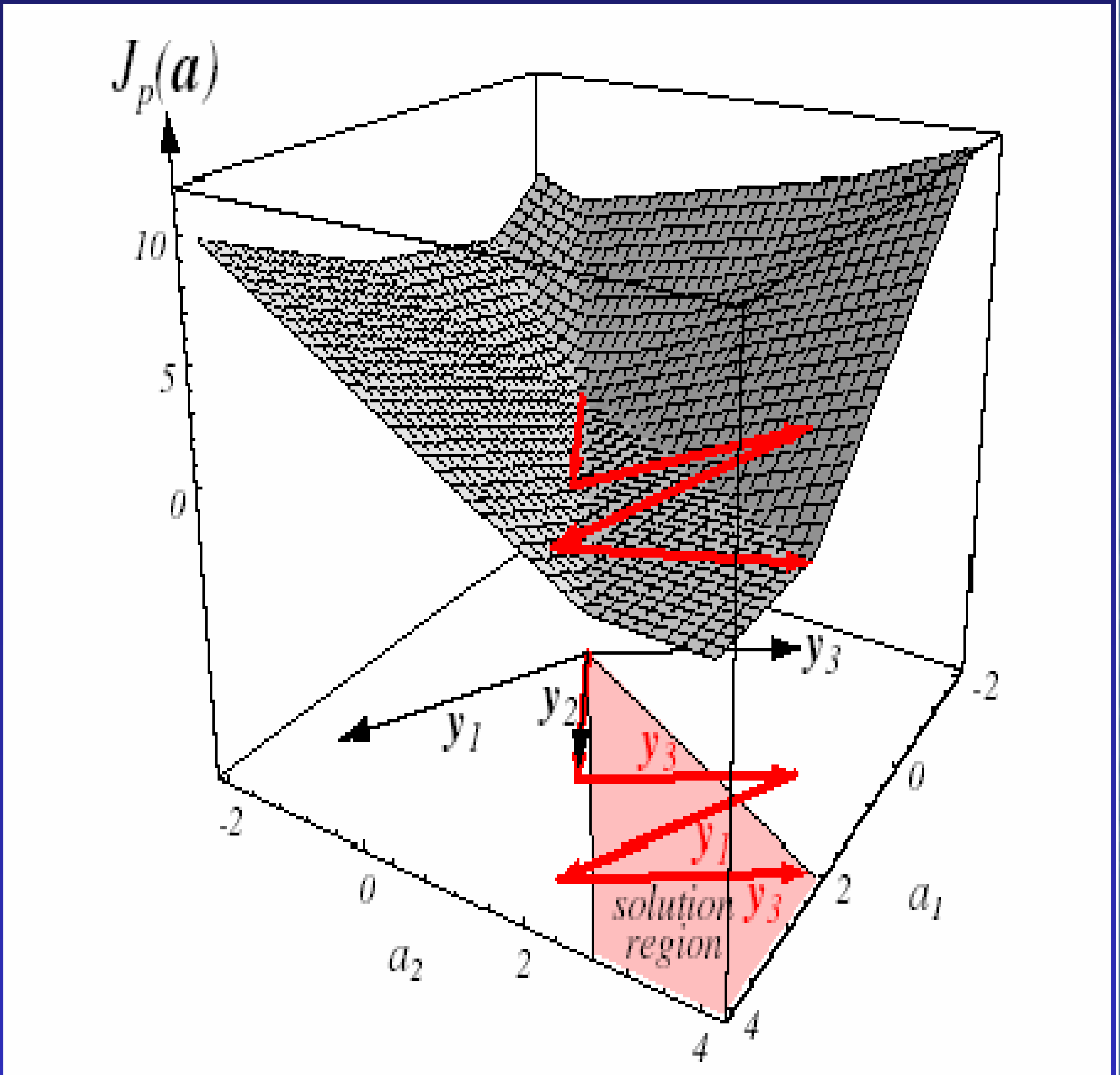
Perceptron criterion:
Piecewise linear,
acceptable for
gradient descent



Squared error:
Useful when patterns
are not linearly separable

Squared Error
with margin

- Perceptron Criterion as function of weights: Demo



Criterion function plotted as a function of weights a1 and a2,
Starts at origin,
Sequence is y2,y3, y1, y3
Second update by y3 takes
solution farther than first update

- Single-Sample Correction

$$a(1) \qquad \text{arbitrary}$$
$$\mathrm{a}(\mathrm{k}+1) = \mathrm{a}(\mathrm{k}) + \eta(k)\mathrm{y}^{\mathrm{k}}$$

- Interpretation of Single-Sample Correction

Single-Sample correction and batch perception algorithm are two algorithms of perception machine.

Either of them can be used !

- Some Direct Generalizations
  - Variable-Increment Perceptron with Margin

$$a(1) \qquad \text{arbitrary}$$

$$\mathrm{a}(\mathrm{k}+1)=\mathrm{a}(\mathrm{k})+\eta(k)\mathrm{y}^{\mathrm{k}} \quad \mathrm{k}\geq 1 \;\; \text{where } \mathrm{a}^{\mathrm{t}}(k)\mathrm{y}^{\mathrm{k}}\leq b \text{ for all k}$$

  - Algorithm Convergence
  - Batch Variable Increment Perception

$$a(1) \qquad \text{arbitrary}$$

$$\mathrm{a}(\mathrm{k}+1)=\mathrm{a}(\mathrm{k})+\eta(k)\sum_{\mathrm{y}\in \mathrm{Y}_{\mathrm{k}}}\mathrm{y}$$

5.6 Relaxation Procedures

- Decent Algorithm
  - criterion function: $J_q(a) = \sum_{y \in Y} (a^t y)^2$
  - two problems of this criterion function
  - criterion function:

$$J_r(a) = \frac{1}{2} \sum_{y \in Y} \frac{(a^t y - b)^2}{\|y\|^2}$$

where $Y(a)$ is a set of samples for which $a^t y \le b$

- Batch Relaxation with Margin
- Single-Sample Relaxation with margin
- Geometrical interpretation of Single-Sample Relaxation with margin algorithm

$$\frac{b-a^t y^k}{\|y^k\|^2} y^k = \frac{b-a^t y^k}{\|y^k\|} \times \frac{y^k}{\|y^k\|} = r(k) \times \frac{y^k}{\|y^k\|}$$

r(k) is the distance from a(k) to the hyperplane $a^t y^k = b$

- From Eq.35 we obtain

$$a^t(k+1)y^k - b = (1-\eta)(a^t(k)y^k - b)$$

5.7 Nonseparable Behavior

- Error-correction procedure
- For nonseparable data the corrections in an error-correction procedure can never cease.
- By averaging the weight vector produced by the correction rule, we can reduce the risk of obtaining a bad solution.
- Some heuristic methods are used in the error-correction rules. The goal is to obtain acceptable performance on nonseparable problems while preserving the ability to find a separating vector on separable problems.
- Usually we let $\eta(k)$ approach zero as k approaches infinity