

Discriminative Classifiers I

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Pattern Recognition and Machine Learning



Lecture 04 Discriminative Classifiers I

Chizhi Chris ZHANG

zhangchizhi@ciomp.ac.cn

Advanced Computing and Digital Technology Research Center

University of Chinese Academy of Sciences

Autumn 2026

Lecture plan and time allocation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Topic	Minutes
Linear classification	45
Fisher projection	45
Nonlinear representations	45

A score, a projection and a feature map

Interpret a boundary; learn a useful projection; change the representation.

The numerical steps are shown on the slides.



Generative and discriminative models target different objects

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The previous lecture obtained decisions from class distributions.

Tomato measurements. Color and firmness can be modelled by class distributions or used directly in a pack/hold score.

$$P(1 \mid x) = 1 / (1 + e^{-g(x)})$$



Bayes can yield the same linear decision form

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A shared-covariance Gaussian model produces a linear log posterior ratio. A discriminative learner can instead fit the same score parameters directly.

$$g(x) = \log \frac{p(x | 1)\pi_1}{p(x | 2)\pi_2} = w^T x + b,$$

$$w = \Sigma^{-1}(\mu_1 - \mu_2), \quad b = -\frac{1}{2}\mu_1^T \Sigma^{-1} \mu_1 + \frac{1}{2}\mu_2^T \Sigma^{-1} \mu_2 + \log \frac{\pi_1}{\pi_2}.$$

$$g(x) > 0 \quad g(x) = 0$$



One score links features, boundary, and class label

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Boundary. $x_1 - .8x_2 - 3.5 = 0$. **Rule.** Sign turns a continuous score into a label. Specify a tie rule at zero. The magnitude is not automatically a probability; multiplying every parameter by 10 leaves all labels unchanged.

$$g(x) = x_1 - .8x_2 - 3.5 = w^T x + b, \quad w = (1, -.8)^T.$$

Input	Score	Decision
(7.2, 4.1)	.42	Pack it immediately, +1
(5.4, 6.0)	-2.90	Suspended. -1
(6.8, 3.5)	.50	Pack it immediately, +1

$$x_1 - .8x_2 - 3.5 = 0$$

Slope and intercept describe the same boundary

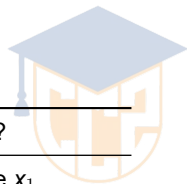
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

In d dimensions, $w^T x + b = 0$ is a hyperplane. If $w_2 = 0$, do not divide by zero: the boundary is vertical when $w_1 \neq 0$. A line equation must be paired with a class-side rule.

$$g(x) = w_1 x_1 + w_2 x_2 + b$$

$$x_2 = -\frac{w_1}{w_2} x_1 - \frac{b}{w_2}.$$

$$x_2 = 1.25x_1 - 4.375.$$



Boundary point	Score check	Which side packs?
(3.5, 0)	$3.5 - 0 - 3.5 = 0$	Fixed x_2 , Increase x_1
(7.5, 5)	$7.5 - 4 - 3.5 = 0$	Fixed x_1 , Reduce x_2

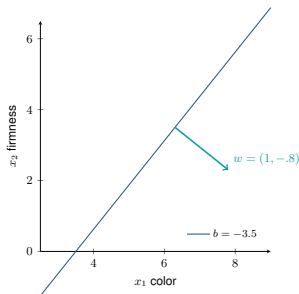
$$w^T x + b = 0$$

The normal vector crosses the boundary

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

$$w^T x_a + b = w^T x_b + b = 0 \quad \Rightarrow \quad w^T (x_a - x_b) = 0.$$

$$(1, -0.8)^T (4, 5) = 4 - 4 = 0.$$



Thus w is perpendicular to every movement along the hyperplane.

$$u = w / \|w\|$$

The normal vector crosses the boundary perpendicularly. Moving along it changes the score. Movement along the boundary has zero inner product with the normal.

Signed distance removes arbitrary score scaling

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The distance is geometric in the chosen feature units. It is not automatically a calibrated class probability or a measure invariant to arbitrary feature rescaling.

$$d(x) = \frac{w^T x + b}{\|w\|}.$$

$$\|w\| = \sqrt{1 + .8^2} = \sqrt{1.64}$$

$$d((7.2, 4.1)) = .42 / \sqrt{1.64} \approx .328.$$



Changing units requires changing the coefficients

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Example. With $\mu = (6, 5)$ and $\mathbf{s} = (2, 1)$, the sorting model becomes

$$x_2 = 4.1 \quad x'_2 = 41$$

$$x_1 - .8x_2 - 3.5 = x_1 - .08x'_2 - 3.5.$$

$$\mathbf{z}_j = (x_j - \mu_j) / s_j$$

$$g(\mathbf{x}) = \sum_j (w_j s_j) \mathbf{z}_j + \left(b + \sum_j w_j \mu_j \right).$$

$$\mu = (6, 5) \quad \mathbf{s} = (2, 1)$$

$$g(\mathbf{z}) = 2\mathbf{z}_1 - .8\mathbf{z}_2 - 1.5.$$



Augmentation unifies bias; separability defines capacity

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Linear separability. For labels $y_i \in \{-1, +1\}$, the training set is separable if some parameters satisfy

$$\tilde{x} = (x_1, \dots, x_d, 1)^T, \quad \tilde{w} = (w_1, \dots, w_d, b)^T, \quad g(x) = \tilde{w}^T \tilde{x}.$$

$y_i \in \{-1, +1\}$

$$y_i(w^T x_i + b) > 0 \quad \text{for every } i.$$



XOR shows a limitation of one affine boundary

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Fixed corners and labels

x_1	x_2	y
-1	-1	+1
-1	+1	-1
+1	-1	-1
+1	+1	+1

Use $g = w_1x_1 + w_2x_2 + b$ and require $yg > 0$.
Reversing all labels does not change nonseparability.

Write and add the inequalities

$$\begin{aligned} -w_1 - w_2 + b &> 0, \\ w_1 + w_2 + b &> 0 \end{aligned} \Rightarrow 2b > 0.$$

$$\begin{aligned} -w_1 + w_2 + b &< 0, \\ w_1 - w_2 + b &< 0 \end{aligned} \Rightarrow 2b < 0.$$

The same b cannot satisfy both. More optimization cannot create a missing affine separator.

Three ways to combine linear scores into a multiclass rule

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Different issues. Independent yes/no rules can overlap or leave gaps. Pairwise votes can form a cycle. Argmax always has a largest score, but tied maxima still need a rule.

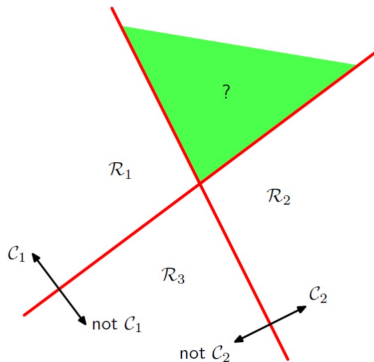
$$\hat{y} = \arg \max_k g_k(x)$$

Construction	Models for C classes	How predictions combine
One versus rest	C	Independent yes/no results can overlap or leave gaps.
One versus one	$C(C - 1)/2$	Pairwise votes can cycle or tie.
Common class scores	C scores	Take the largest comparable score; resolve tied maxima.

One-versus-rest can leave ambiguous regions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Read the arrows. Each red line separates one class from its rest. The green region is marked “?” because independent binary answers do not supply one agreed class there.

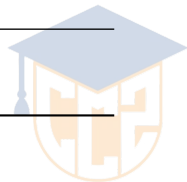


Pairwise votes can agree or cycle

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Why a cycle is possible. Three separately trained pairwise functions need not be differences of one shared set of class scores. Their preferences can therefore be inconsistent.

Pairwise outputs	Number of votes (A, B, C)	Predicted
A beats B; C beats A; C beats B	(1, 0, 2)	C
A beats B; B beats C; C beats A	(1, 1, 1)	Tie



Argmax compares all classes on one score scale

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

At (3, 5), two scores are positive, so strict OvR would be ambiguous; argmax still selects the larger score.

$$g_1 = -x_1 + x_2, \quad g_2 = x_1 + x_2 - 5, \quad g_3 = -x_2 + 1.$$

Input	g_1	g_2	g_3	Argmax
(6, 5)	-1	6	-4	Class 2
(3, 5)	2	3	-4	Class 2

$$g_c(x) = g_k(x) \implies (w_c - w_k)^T x + (b_c - b_k) = 0.$$

Worked review: preserve the rule, then interpret the score

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Completed calculation. Score .42, positive label, distance .328 in the original feature geometry. Equivalent new firmness coefficient: $-.08$. Positive parameter scaling changes raw scores but preserves the boundary and labels. A probability requires a specified probability model, not merely a score.

$$g = x_1 - .8x_2 - 3.5 \quad x = (7.2, 4.1)$$

$$x'_2 = 10x_2$$

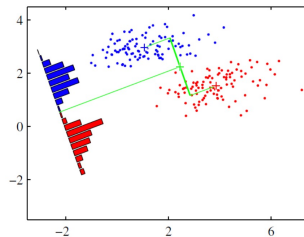
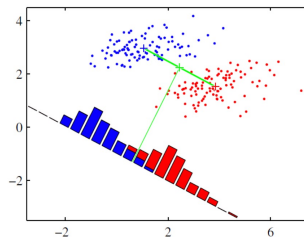
Check	Calculation or change	Supported conclusion
Given point	$7.2 - .8(4.1) - 3.5 = .42$	Positive label.
Distance	$.42 / \sqrt{1.64} \approx .328$	Distance in the original feature geometry.
Feature unit	$x'_2 = 10x_2$; coefficient $-.08$	The physical rule is unchanged.
Parameter scale	Multiply both w, b by 5	Scores multiply by 5; boundary and labels stay fixed.

Projection can keep or destroy class separation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A dot product assigns each point one coordinate along a chosen direction.

Compare the projections. Both panels show the same blue/red observations. The projection axis changes. Histograms along the axis show how the two classes overlap after projection.



A dot product gives a one-dimensional coordinate

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For unit w , $z = \|x\| \cos \theta$ is the signed coordinate along the projection axis. The projected vector in the original space is zw .

$$z = w^T x.$$

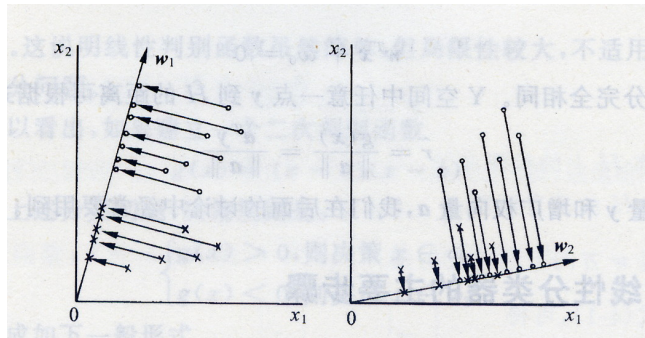
$$z = \|x\| \cos \theta$$

Point	$w = (1, 0)^T$	$w = (0, 1)^T$	$w = (1, 1)^T / \sqrt{2}$
(2, 0)	2	0	$\sqrt{2}$
(2, 2)	2	2	$2\sqrt{2}$

Compare overlap after two projections

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Compare the overlap of the projected class marks along the two candidate directions.



Projected means and scatter follow directly from the dot product

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Scatter is a sum, not an average covariance.
Center samples by their own class means.
Centering everything by the grand mean would mix between-class separation into the denominator we intend to keep small.

$$m_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i, \quad S_k = \sum_{i:y_i=k} (x_i - m_k)(x_i - m_k)^T.$$

$$S_W = S_1 + S_2, \quad S_B = (m_1 - m_2)(m_1 - m_2)^T.$$

Project a class mean using the same dot product as a sample. Substituting the projected deviations gives the within-class scatter $w^T S_k w$.

$$\tilde{m}_k = \frac{1}{n_k} \sum_{i:y_i=k} w^T x_i = w^T m_k, \\ \tilde{S}_k = \sum_{i:y_i=k} (w^T x_i - w^T m_k)^2 = w^T S_k w,$$

$$(\tilde{m}_1 - \tilde{m}_2)^2 = [w^T (m_1 - m_2)]^2 = w^T S_B w.$$

$$m_k, w \in \mathbb{R}^d \quad S_k \in \mathbb{R}^{d \times d}$$

Fisher measures separation relative to within-class spread

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

Scaling the axis stretches both the center gap and spread, leaving their ratio unchanged. For a derivation, fix $w^T S_W w = 1$. For display, normalize $\|w\| = 1$.

$$J(w) = \frac{[w^T(m_1 - m_2)]^2}{w^T S_W w} = \frac{w^T S_B w}{w^T S_W w}.$$

$$J(cw) = J(w) \quad \text{for nonzero } c.$$

$$w^T S_W w = 1$$

The ratio is unchanged by a nonzero rescaling. During the derivation choose $w^T S_W w = 1$; this generally cannot simultaneously impose $\|w\| = 1$.



A noisy mean-difference direction can be inferior

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The first coordinate separates the means but varies greatly within classes. Fisher downweights that noisy direction without throwing it away completely.

$$\Delta = m_1 - m_2 = (2, 1)^T \quad S_W = \text{diag}(16, 1)$$

Direction w	$(w^T \Delta)^2$	$w^T S_W w$	$J(w)$
Mean-difference direction $(2, 1)$	25	65	.385
Vertical direction $(0, 1)$	1	1	1
Fisher's direction. $(1/8, 1)$	25/16	5/4	1.25



Derive the direction from a constrained objective

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Assumptions. Distinct class means and positive-definite S_W give the nonzero separating solution. The constant sets scale, not direction. Numerically, solve $S_W w = \Delta$; do not explicitly form the inverse.

$$w^T S_W w = 1$$

$$\mathcal{L}(w, \lambda) = w^T S_B w - \lambda(w^T S_W w - 1).$$

$$\nabla_w \mathcal{L} = 2S_B w - 2\lambda S_W w = 0 \quad \Rightarrow \quad S_B w = \lambda S_W w.$$

$$\Delta = m_1 - m_2 \quad S_B = \Delta \Delta^T$$

$$S_B w = \Delta(\Delta^T w), \quad w \propto S_W^{-1} \Delta.$$

$$S_W w = \Delta$$

$$\Delta(\Delta^T w) = \lambda S_W w \quad \Rightarrow \quad w = \frac{\Delta^T w}{\lambda} S_W^{-1} \Delta.$$



Interpret the inverse scatter as a directional correction

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The inverse scatter corrects the mean difference using within-class variation. First use the supplied diagonal example from the preceding comparison.

$$\Delta = (2, 1)^T, \quad S_W = \text{diag}(16, 1), \quad w = S_W^{-1} \Delta = (1/8, 1)^T.$$

Coordinate	Mean difference	Within-class scatter	Corrected coefficient
1	2	16	$2/16 = 1/8$
2	1	1	$1/1 = 1$



The first coordinate has the larger raw mean gap but much more within-class variation, so it receives less weight.

With correlated coordinates, S_W has off-diagonal entries. Solve the full system $S_W w = \Delta$ rather than dividing each coordinate independently. Next derive these statistics from a complete training table.

Training data for one complete Fisher calculation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

This is a new eight-point dataset, chosen to expose correlated within-class variation.

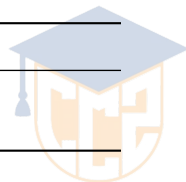
What the data suggest.

The second coordinate changes between classes; both classes share the same within-class shape because class 2 is a translation of class 1 by $(0, 2)$.

Class	Points	Mean
1	$(1, 1), (2, 0), (2, 1), (3, 0)$	$(2, .5)^T$
2	$(1, 3), (2, 2), (2, 3), (3, 2)$	$(2, 2.5)^T$

$$\Delta = m_1 - m_2 = (0, -2)^T.$$

$$S_W W = \Delta$$



One centered row contributes an outer product

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Class 1 has mean $(2, 0.5)^T$. Its first row $(1, 1)^T$ has deviation $d = (-1, 0.5)^T$.
 $d = (-1, 0.5)^T$

$$dd^T = \begin{bmatrix} -1 \\ .5 \end{bmatrix} \begin{bmatrix} -1 & .5 \end{bmatrix} = \begin{bmatrix} 1 & -.5 \\ -.5 & .25 \end{bmatrix}.$$

$$S_1 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}.$$

$$S_2 = S_1 \quad S_W = 2S_1$$



Compute the class scatter from four centered rows

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The off-diagonal entry is negative because horizontal and vertical deviations tend to have opposite signs. Scatter sums squared deviations; it is not divided by 4 in the Fisher convention used here.

Worked calculation

Class-1 deviation	δ_1^2	$\delta_1 \delta_2$	δ_2^2
(-1, .5)	1	-.5	.25
(0, -.5)	0	0	.25
(0, .5)	0	0	.25
(1, -.5)	1	-.5	.25
Sum	2	-1	1

$$S_1 = S_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}, \quad S_W = \begin{bmatrix} 4 & -2 \\ -2 & 2 \end{bmatrix}.$$

Solve the Fisher system and compare the resulting ratio

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

$S_W(-1, -2)^T = (0, -2)^T = \Delta$. The direction is now supported by the same training rows.

$$S_W = \begin{bmatrix} 4 & -2 \\ -2 & 2 \end{bmatrix}, \quad \Delta = \mu_1 - \mu_2 = \begin{bmatrix} 0 \\ -2 \end{bmatrix}.$$

$$S_W w = \Delta$$

$$4w_1 - 2w_2 = 0 \Rightarrow w_2 = 2w_1.$$

$$-2w_1 + 2w_2 = -2 \Rightarrow -2w_1 + 4w_1 = -2$$

$$\Rightarrow w_1 = -1,$$

$$w_2 = -2.$$

$$S_W(-1, -2)^T = (0, -2)^T = \Delta$$

The Fisher direction uses the negative within-class correlation to reduce projected variation relative to the gap. It improves the ratio even though the class means differ only vertically.

$$w_2 = 2w_1 \quad w_1 = -1$$

Direction	Gap sq.	Within	J
Vertical direction (0, -1)	4	2	2
Fisher's direction. (-1, -2)	16	4	4

Projection needs a threshold and a class-side rule

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Equal-prior midpoint rule. Predict class 1 if $z > -5$, class 2 if $z < -5$. The boundary is

$$w = (-1, -2)^T$$

Class	Projected training values $z = w^T x$	Projected mean
1	-3, -2, -4, -3	-3
2	-7, -6, -8, -7	-7

$$t = \frac{-3 + (-7)}{2} = -5.$$

$$z > -5 \quad z < -5$$

$$x = (2, 1) \quad z = -4$$

$$x = (2, 2) \quad z = -6$$

$$-x_1 - 2x_2 = -5 \iff x_1 + 2x_2 = 5.$$

Priors and losses set the projected threshold

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Let $\mathbf{z} = \mathbf{w}^T \mathbf{x}$. Assume projected classes are Gaussian with means μ_1, μ_2 and a shared variance σ_z^2 . Priors are π_1, π_2 .

$$\log \frac{P(1 | \mathbf{z})}{P(2 | \mathbf{z})} = \frac{\mu_1 - \mu_2}{\sigma_z^2} \left(\mathbf{z} - \frac{\mu_1 + \mu_2}{2} \right) + \log \frac{\pi_1}{\pi_2}.$$

c_{12} means choosing 1 when truth is 2; c_{21} means choosing 2 when truth is 1. Correct decisions have zero cost.

$$c_{12}P(2 | \mathbf{z}) < c_{21}P(1 | \mathbf{z}) \iff \log \frac{P(1 | \mathbf{z})}{P(2 | \mathbf{z})} > \log \frac{c_{12}}{c_{21}}.$$

$$t = \frac{\mu_1 + \mu_2}{2} + \frac{\sigma_z^2}{\mu_1 - \mu_2} \log \frac{c_{12}\pi_2}{c_{21}\pi_1}.$$

Equal costs, $\mu_1 = -3$, $\mu_2 = -7$, $\sigma_z^2 = .5$, $\pi_1 = .8$: $t = -5 - \frac{1}{8} \log 4 \approx -5.173$. Since $\mu_1 > \mu_2$, choose 1 for $\mathbf{z} > t$.

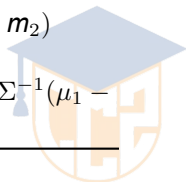


Fisher and Gaussian LDA share a direction under conditions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

If the pooled within-class covariance is proportional to S_W , the estimated directions agree up to scale. Gaussian LDA additionally supplies class probabilities under its distributional model.

Method	What is assumed/optimized	What is obtained
Fisher projection	Maximize mean gap relative to within scatter	Direction $S_W^{-1}(m_1 - m_2)$
Gaussian LDA	Gaussian classes with one shared covariance	Log-odds direction $\Sigma^{-1}(\mu_1 - \mu_2)$ and intercept



Regularization limits amplification of weak directions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For two classes, $\text{rank}(S_W) \leq \min(d, n - 2)$. Duplicated features or too few independent observations can make the system singular.

A separate singular example

Add λI with $\lambda = 1$

$$S_W = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \Delta = (1, 0)^T.$$

$$\det S_W = 1(1) - 1(1) = 0.$$

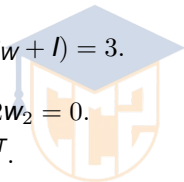
Both equations have the same left side but different right sides. There is no solution to $S_W w = \Delta$ here.

In general, solve $(S_W + \lambda I)w_\lambda = m_1 - m_2$, $\lambda > 0$. Regularization changes the solution; validation must judge that tradeoff.

$$S_W + I = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, \quad \det(S_W + I) = 3.$$

$$2w_1 + w_2 = 1, \quad w_1 + 2w_2 = 0.$$

$$w = (2/3, -1/3)^T.$$



Fisher calculations and the role of scale

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Explain why multiplying w and the threshold by 10 changes no labels.

$$m_1 = (0, 0)^T \quad m_2 = (2, 1)^T$$

$$S_W = \text{diag}(8, 2) \quad S_W w = m_1 - m_2$$

$$x = (1, 0)^T \quad w = (-1/4, -1/2)^T$$

$$t = -1/2 \quad z = -1/4 > t$$

$$w^T m_1 = 0, \quad w^T m_2 = -1, \quad t = (0 - 1)/2 = -.5.$$

$$z = w^T (1, 0) = -.25 > -.5.$$

$$w' = 10w: \quad z' = -2.5, \quad t' = 10t = -5, \quad z' > t'.$$

The label is unchanged only when the threshold scales with the direction.



A feature map changes the coordinates of the score

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

One affine score provides one straight boundary in its input coordinates.

Key idea

x is a measured scalar. A single affine score $w x + b$ crosses zero at most once.

$$\phi(x) = (1, x, x^2)^T \quad \beta = (ab, -a - b, 1)^T$$

$$\beta^T \phi(x) = ab - (a + b)x + x^2 = (x - a)(x - b).$$

$$a < x < b$$



A fixed feature map makes nonlinear structure accessible

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Linear **in parameters** does not imply linear **in the original input**. The feature map is chosen before fitting w . Its interpretation depends on the measurement definition and preprocessing.

$$\phi(x) = (\phi_1(x), \dots, \phi_M(x))^T, \quad g(x) = w^T \phi(x) + b.$$

Original observation	Added feature	New structure represented
One measurement x	x^2	Two threshold crossings
Two symbols (x_1, x_2)	$x_1 x_2$	Sign agreement/XOR pattern
Color and firmness	Interaction term	Effect of color depends on firmness

A quadratic score isolates a middle interval

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Desired rule. Class 2 for $1 < x < 3$; class 1 outside that interval.

$$1 < x < 3$$

$$g(x) = (x - 1)(x - 3) = x^2 - 4x + 3.$$

$$\phi(x) = (x^2, x, 1)^T, \quad w = (1, -4, 3)^T.$$

Input	Score	Class
0	3	1
1.5	-.75	2
2	-1	2
3.5	1.25	1

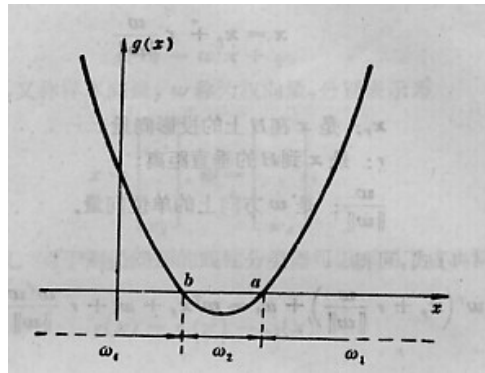
$$(z_1, z_2) = (x^2, x) \quad z_1 - 4z_2 + 3 = 0$$

$$z_1 = z_2^2$$

Two crossings isolate a middle interval

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A quadratic score can cross zero at a and b , with $a < b$. The sign between the roots differs from the sign outside them.



One interaction resolves the diagonal-label pattern

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

What changed? The new feature records sign agreement, which no weighted sum of the original two coordinates can represent on these four corners.

x_1	x_2	$x_1 x_2$	Target
-1	-1	+1	+1
-1	+1	-1	-1
+1	-1	-1	-1
+1	+1	+1	+1

$$g(x) = x_1 x_2, \quad \hat{y} = \text{sign } g(x).$$



A full quadratic map expresses several boundary shapes

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For input $(1, 2)$, the feature vector is $(1, 2, 4, 1, 2, 1)$.

Different weights select different conic or degenerate shapes.

The model remains linear in its six coefficients; adding features changes capacity before any fitting procedure is chosen.

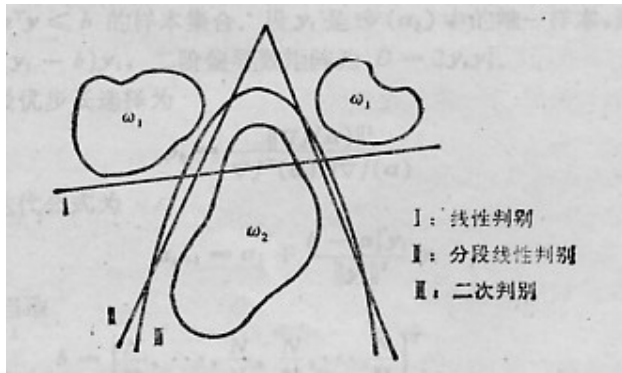
$$\phi(\mathbf{x}) = (x_1^2, x_1x_2, x_2^2, x_1, x_2, 1)^T.$$

$$g(\mathbf{x}) = ax_1^2 + bx_1x_2 + cx_2^2 + dx_1 + ex_2 + f.$$

Equation $g(\mathbf{x}) = 0$	Example geometry
$x_1^2 + x_2^2 - 4 = 0$	Circle of radius 2
$x_1^2 - x_2 = 0$	Parabola
$x_1x_2 = 0$	Two intersecting lines

Straight, piecewise and curved boundaries

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



The reference compares straight, piecewise linear and curved boundaries. These show different representational choices. Their shapes alone do not establish predictive performance.



Least squares fits targets but can overemphasize extreme scores

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

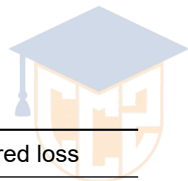
Encode labels by $t_i \in \{-1, +1\}$ and include a constant column in Φ :

$t_i \in \{-1, +1\}$

$$\min_w \|\Phi w - t\|^2.$$

$$\nabla_w \|\Phi w - t\|^2 = 2\Phi^T(\Phi w - t) = 0 \Rightarrow \Phi^T \Phi w = \Phi^T t.$$

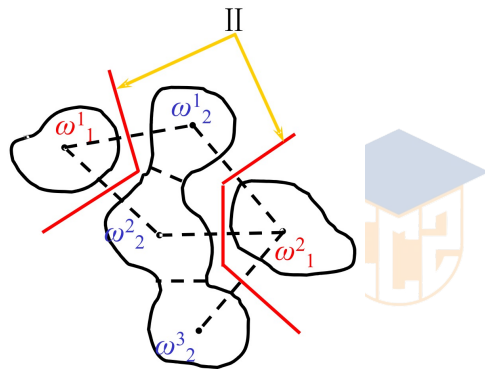
Positive sample	Score	Residual	Squared loss
Correct but extreme	8	7	49
Wrong but near zero	-.1	-1.1	1.21



Several prototypes preserve separate class regions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Read the class regions. Red labels mark two subregions of class 1; blue labels mark three of class 2. A point inside each region represents a prototype.



Nearest-prototype scores become affine comparisons

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The term $-x^T x$ is common to every prototype. Therefore compare affine scores

$$\hat{j} = \arg \min_j \|x - \mu_j\|^2.$$

$$-\|x - \mu_j\|^2 = -x^T x + 2\mu_j^T x - \|\mu_j\|^2.$$

$$h_j(x) = 2\mu_j^T x - \|\mu_j\|^2.$$

$$g_c(x) = \max_{j: \text{label}(j)=c} h_j(x), \quad \hat{y} = \arg \max_c g_c(x).$$



Worked prototype decisions retain a multimodal class

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

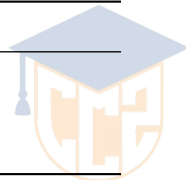
Class A has prototypes $\mu_{A1} = (0, 0)$ and $\mu_{A2} = (4, 0)$; class B has $\mu_B = (2, 2)$.

$$\mu_{A1} = (0, 0) \quad \mu_{A2} = (4, 0)$$

$$\mu_B = (2, 2)$$

Input	d_{A1}^2	d_{A2}^2	d_B^2	Predicted
(.5, .5)	.5	12.5	4.5	A (prototype 1)
(3.5, .5)	12.5	.5	4.5	A (prototype 2)
(2, 1.5)	6.25	6.25	.25	B

$$x_1 + x_2 = 2 \quad -x_1 + x_2 = -2$$



k-nearest neighbors use training samples rather than fixed prototypes

For a new x , find its k nearest labelled training samples and predict the majority class.

training points; queries $x = (2, 1)$	Label	Squared distance
$(2, 1.1)$	A	.01
$(1.5, 1.8)$	B	.89
$(2.5, 1.8)$	B	.89
$(0, 0)$	A	5
$(4, 0)$	A	5



$k = 1$ $k = 3$

$k = 5$ $k = 1$

Neighbor identity depends on scaling; output depends on the task

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Fit appropriate scaling from training data; distance units determine which observations are local.

$$q = (0, 0) \quad a = (1, 10)$$

$$b = (3, 1)$$

Coordinates used	$\ q - a\ ^2$	$\ q - b\ ^2$	Nearest
Raw features	101	10	b
Second feature divided by 10	2	9.01	a

Classification votes: $(A, B, B) \rightarrow B$. Regression averages: $(8 + 9 + 13)/3 = 10$.

A maximum of affine pieces supplies a definite score

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Class A occupies two tails, while B occupies $-1 < x < 3$. The active A piece changes at $x = 1$, but this is not a class boundary. The class boundaries occur at $x = -1$ and $x = 3$.

$$g_{c\ell}(x) = w_{c\ell}^T x + b_{c\ell}, \quad g_c(x) = \max_{\ell} g_{c\ell}(x).$$

$$\hat{y} = \arg \max_c g_c(x).$$

$$g_A(x) = \max(x - 3, -x - 1) \quad g_B(x) = 0$$

x	g_A	g_B	Class
-2	1	0	A
0	-1	0	B
4	1	0	A

$$-1 < x < 3 \quad x = 1$$

$$x = -1 \quad x = 3$$

Choose a representation, then verify its calculation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Interval. Class B for $2 < x < 5$, class A outside. Write one quadratic score and its feature weights.

$$2 < x < 5$$

For the interval, $\phi = (x^2, x, 1)$ and $w = (1, -7, 10)$. For the corner pattern, adding an interaction changes the representation, so the old affine impossibility proof no longer applies.

Question	Computation	Rule
Interval	$(x - 2)(x - 5) = x^2 - 7x + 10$	Negative: B; positive: A
Diagonal corners	$g = x_1 x_2$	Sign gives the target
Representative query point (3, 1)	Squared distance (10, 2, 2)	A2/B tie; use declared tie policy

$$\phi = (x^2, x, 1) \quad w = (1, -7, 10)$$

What does a discriminative score establish?

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Question

A linear boundary, a Fisher projection and a feature map all use coefficients. Are their meanings interchangeable?

Evidence and reasoning

The linear normal determines a boundary orientation. Fisher maximizes between-class separation relative to within-class scatter. A feature map changes which boundaries can be represented. Their objectives and assumptions explain their different roles.

Meaning of a score

A large score becomes a probability only through a suitable probability model and evaluation.