

Pattern Recognition and Machine Learning

Lecture 09 Support Vector Machines



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 SVM	22.5
Soft margin	22.5
Dual formulation and kernels	45
Support vector regression	45

The separating corridor and the regression tube

Define geometric margin; price violations; use kernels; interpret target tolerance.

The numerical steps are shown on the slides.



Several separators fit the same training data

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A classifier score identifies a side of the boundary. Dividing by the normal-vector length converts it to geometric distance. This makes margin comparisons insensitive to multiplying all score parameters by a positive constant.

Each row describes one inspected item: standardized roughness x_1 , standardized vibration x_2 , and a verified label $y \in \{-1, +1\}$. Correct training signs do not uniquely determine a line.

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

$$\mathcal{D} = \{((-2, 0), -1), ((-1, 0), -1), ((1, 0), +1), ((2, 0), +1)\}$$

$$x_1 = 0 \quad x_1 = 0.8$$

Check. A larger minimum clearance tolerates a larger small measurement shift in the same feature geometry. It is a design principle, not a guaranteed probability of success.



A linear score has a direction and a position

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

$f(x) = w^T x + b$ is a signed score. The set $f(x) = 0$ is the decision boundary; w is perpendicular to that boundary, while b shifts its location.

$$f(x) = w^T x + b \quad f(x) = 0$$

$$w = (3, 4)^T, \quad b = -5, \quad x = (1, 1)^T \Rightarrow f(x) = 3 + 4 - 5 = 2$$

$$3x_1 + 4x_2 = 5$$

Check. A positive score predicts class +1. The numerical score 2 is neither a geometric distance nor a probability.



Derive the distance instead of memorizing it

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The boundary is $w^T x + b = 0$. Its unit normal is $n = w / \|w\|$. From point p , move a signed distance t to $q = p - tn$.

$$w^T x + b = 0 \quad n = w / \|w\|$$

$$q = p - tn$$

$$0 = w^T q + b = w^T (p - tn) + b = (w^T p + b) - t w^T n.$$

$$w^T n = \frac{w^T w}{\|w\|} = \|w\| \quad \Rightarrow \quad t = \frac{w^T p + b}{\|w\|}.$$

Let x_0 be the perpendicular projection of x onto $H = \{z : w^T z + b = 0\}$. Travel along the unit normal $w / \|w\|$.

$$H = \{z : w^T z + b = 0\} \quad w / \|w\|$$

$$x = x_0 + d \frac{w}{\|w\|} \quad \Rightarrow \quad w^T x + b = d \|w\|$$

$$d = \frac{w^T x + b}{\|w\|}, \quad \text{dist}(x, H) = |d|$$

$$w = (3, 4)$$

$$b = -5$$

$$x = (1, 1)$$

$$d = \frac{3+4-5}{\sqrt{3^2+4^2}} = 2/5 = 0.4$$

Normalize the closest functional margin to one

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For any $a > 0$, (w, b) and (aw, ab) give the same boundary and the same signs. A raw functional margin is therefore not an invariant measure of clearance.

$a > 0$

$$\frac{y_i((aw)^T x_i + ab)}{\|aw\|} = \frac{a y_i(w^T x_i + b)}{a \|w\|} = \frac{y_i(w^T x_i + b)}{\|w\|}$$

$20/50 = 0.4$

For separable data choose the parameter scale so that the smallest labeled score is 1. All training constraints can then use one formula.

$$y_i(w^T x_i + b) \geq 1 \quad (i = 1, \dots, N)$$

$$y_i = +1 \Rightarrow f(x_i) \geq 1, \quad y_i = -1 \Rightarrow f(x_i) \leq -1$$

$$w = (1, 0) \quad b = 0$$

$$1/\|w\|$$

Check. The constraint value 1 fixes score scale. The physical clearance is still $1/\|w\|$.



The two supporting planes form a corridor

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The decision plane is $f(x) = 0$. Its two parallel supporting planes are $f(x) = +1$ and $f(x) = -1$. Under hard-margin constraints, no training point lies inside the open corridor.

$$f(x) = 0 \quad f(x) = +1$$

$$f(x) = -1$$

$$\text{distance to each supporting plane} = \frac{1}{\|w\|} \quad \text{full width} = \frac{2}{\|w\|}$$

$$w = (1, 0) \quad b = 0$$

$$x_1 = 1 \quad x_1 = -1$$

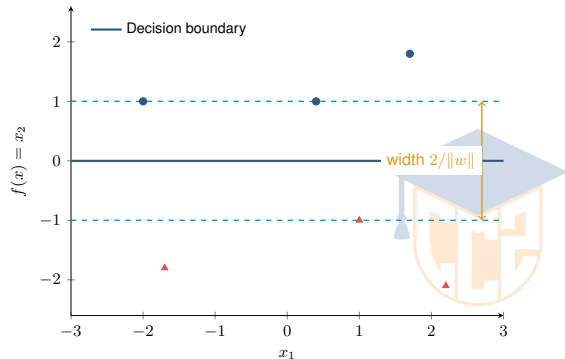
Check. Some texts call the half-width the margin. State which convention is used before comparing numerical answers.



Measure the complete separating corridor

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The two margin planes have scores $+1$ and -1 . Their perpendicular separation is $2/\|w\|$. Maximizing this width is equivalent to minimizing the squared normal-vector length under the margin constraints.



Solve a one-dimensional hard margin step by step

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Keep only the closest opposite-class observations $x_- = -1$ and $x_+ = 1$. Their constraints limit the width that any feasible separator can achieve.

$$x_- = -1 \quad x_+ = 1$$

$$\begin{aligned} -(-w + b) &\geq 1, \quad w + b \geq 1 \quad \Rightarrow \quad w \geq 1 \\ \min \frac{1}{2} w^2 &\Rightarrow \quad w^* = 1, \quad b^* = 0, \quad \text{width} = 2 \end{aligned}$$



Hard-margin SVM is a constrained quadratic problem

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

After fixing score scale, maximizing corridor width is equivalent to minimizing the squared weight norm. The bias remains a position parameter and is not included in this standard norm penalty.

$$\min_{w,b} \frac{1}{2} \|w\|^2 \quad \text{subject to} \quad y_i(w^T x_i + b) \geq 1 \quad \forall i$$

Check. Convexity gives a global optimum for this problem when feasible; it does not make wrong labels or unrepresentative data reliable.



Support vectors are the active training constraints

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

In a nondegenerate hard-margin solution, critical rows lie on $y_i f(x_i) = 1$. Rows farther away leave room in their inequalities; a small movement of them may leave the optimum unchanged.

$$y_i f(x_i) = 1$$

$$x = -1, +1 : y f(x) = 1; \quad x = -2, +2 : y f(x) = 2$$

$$w = 1, b = 0$$



Condition to check

Support does not mean representative, frequent, or certainly correct. It means the row participates in determining the optimized boundary.

Feature units are part of the margin definition

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Euclidean clearance depends on the coordinates supplied to the model. Changing one feature from millimeters to micrometers changes distances unless preprocessing restores a deliberate scale.

$$z_j = \frac{x_j - \mu_j}{s_j}, \quad \mu_j, s_j \text{ estimated on the training fold only}$$

Check. Scaling is a modeling choice. Save training statistics and reuse them for validation and prediction; do not refit them on held-out rows.



Conflicting duplicate labels reveal the role of costs

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Slack measures how far a case falls short of its required signed score. A point may be correctly classified and still incur slack. The penalty adds these shortfalls to the model-size term.

At one identical input location, a positive and negative label receive the same score f . Their combined hinge loss has an irreducible floor.

$$\max(0, 1 - f) + \max(0, 1 + f) = 2 \quad \text{for } -1 \leq f \leq 1$$

$$2 \max(0, 1 - f) + \max(0, 1 + f) \text{ is minimized at } f = 1$$



Compare two compromises with the same objective

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Write $m_i = y_i f(x_i)$. The smallest feasible slack is $\xi_i = \max(0, 1 - m_i)$, so being inside the corridor and being misclassified are different events.

$$m_i = y_i f(x_i) \quad \xi_i = \max(0, 1 - m_i)$$

m_i	ξ_i	Sign correct?	Position
1.4	0	Yes	outside/on margin
0.4	0.6	Yes	inside corridor
0	1	boundary	decision plane
-0.3	1.3	No	wrong side

Soft-margin SVM balances a short weight vector against total violation. C is the price assigned to one unit of slack in the chosen objective convention.

$$\min_{w,b,\xi} \frac{1}{2} \|w\|^2 + C \sum_i \xi_i$$

$$\text{subject to } y_i f(x_i) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

$$J_A = 0.5 + 1.2C, \quad J_B = 2 + 0.2C$$

$$C = 0.5 \quad J_A = 1.1 < J_B = 2.1$$

$$C = 3 \quad J_B = 2.6 < J_A = 4.1$$

Check. Increasing C changes the preferred tradeoff. It does not promise a monotonic improvement in validation accuracy.

Eliminate slack and recover hinge loss

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For fixed w, b , positive violation costs make the optimal slack as small as possible. Substituting it gives a loss-plus-regularization objective.

$$\ell_{\text{hinge}}(y, f) = \max(0, 1 - yf)$$
$$J = \frac{1}{2} \|w\|^2 + C \sum_i \max(0, 1 - y_i f(x_i))$$

$$m = (1.4, 0.4, -0.3)$$

Check. Hinge loss rewards correct signs with sufficient margin. Logistic loss keeps decreasing smoothly even beyond that margin.



Take one hinge-loss subgradient step and inspect the score

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For one positive row $x = (2, 1)$, start with $w = 0, b = 0, C = 1$. Its margin is below 1, so the hinge contribution has gradients $-yx$ and $-y$.

$$x = (2, 1) \quad w = 0, b = 0$$

$$C = 1$$

$$\eta = .1 : \quad w^{\text{new}} = .1(2, 1) = (.2, .1), \quad b^{\text{new}} = .1$$

$$f^{\text{new}}(x) = .2(2) + .1(1) + .1 = .6, \quad \ell^{\text{new}} = .4$$

Check. Later updates must include regularization and all relevant rows. Beyond the margin, this row has zero hinge gradient rather than an endless demand for larger score.



Weights and parameter conventions change the tradeoff

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

If missing a defect is more costly, use row or class weights in the violation term. Distinguish this cost choice from the later choice of a prediction threshold.

$$J = \frac{1}{2} \|w\|^2 + C \sum_i a_i \ell_i \quad (a_i > 0)$$

$$\frac{\lambda}{2} \|w\|^2 + \frac{1}{N} \sum_i \ell_i \iff C = \frac{1}{N\lambda}$$

$$N = 100 \quad \lambda = 0.02$$

$$C = 0.5$$

Check. A bigger C means a smaller equivalent regularization strength here. Compare conventions before copying a number from another model.



Why rewrite the model using training rows?

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Feature coefficients

The primal problem fits one weight per input feature and a bias. In d dimensions, w has d entries.

Training-row coefficients

The dual will attach one coefficient α_i to each training row. The labels determine the sign of its contribution.

$$w = \sum_i \alpha_i y_i x_i \quad \Rightarrow \quad f(x) = \sum_i \alpha_i y_i x_i^T x + b.$$

We will derive this representation from the constraints. First, a one-variable example explains what a constraint multiplier does.

The dual is useful for kernels. It is not automatically cheaper when the training set is large.



A constraint multiplier in one dimension

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Primal: the best feasible value

$$\min_u \frac{1}{2}u^2 \quad \text{subject to } u \geq 1.$$

The best feasible point is $u = 1$, with value $1/2$.

$$L(u, \alpha) = \frac{1}{2}u^2 + \alpha(1 - u), \quad \alpha \geq 0.$$

For feasible u , the added term is nonpositive. Minimizing L over all u gives a lower bound on the feasible minimum.

Dual: make that bound tight

$$\partial_u L = u - \alpha = 0 \Rightarrow u = \alpha.$$

$$q(\alpha) = L(\alpha, \alpha) = \alpha - \frac{1}{2}\alpha^2.$$

$$q'(\alpha) = 1 - \alpha = 0 \Rightarrow \alpha = 1.$$

The largest bound is $q(1) = 1/2$, matching the primal value. SVM applies the same construction to many row constraints.

Put both soft-margin constraints into a Lagrangian

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

Write constraints as $1 - \xi_i - y_i f(x_i) \leq 0$ and $-\xi_i \leq 0$. Attach nonnegative multipliers α_i and μ_i respectively.

$$L = \frac{1}{2} \|w\|^2 + C \sum_i \xi_i + \sum_i \alpha_i (1 - \xi_i - y_i (w^T x_i + b)) - \sum_i \mu_i \xi_i$$

Check. Keep every sign consistent with the chosen inequality direction. We need only the SVM-specific stationarity and complementarity conditions.



Three derivatives reveal three pieces of structure

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Minimize the Lagrangian over primal variables. Stationarity eliminates w, b, ξ and leaves a constrained optimization in the row multipliers.

$$\nabla_w L = 0 \Rightarrow w = \sum_i \alpha_i y_i x_i$$

$$\partial_b L = 0 \Rightarrow \sum_i \alpha_i y_i = 0$$

$$\partial_{\xi_i} L = 0 \Rightarrow C - \alpha_i - \mu_i = 0 \Rightarrow 0 \leq \alpha_i \leq C$$



Check. The upper bound comes from soft-margin penalties. Hard-margin multipliers have no finite C cap.

Read the dual objective by its two terms

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Stationarity gives $C - \alpha_i - \mu_i = 0$ and $\sum_i \alpha_i y_i = 0$, removing the slack and bias terms. Substitute the weight vector to leave only pairwise inner products in the objective.

$$\begin{aligned} L &= \frac{1}{2} \|\mathbf{w}\|^2 - \mathbf{w}^T \sum_i \alpha_i y_i \mathbf{x}_i + \sum_i \alpha_i \\ &= -\frac{1}{2} \|\mathbf{w}\|^2 + \sum_i \alpha_i \\ &= \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j. \end{aligned}$$



Maximize this objective subject to $0 \leq \alpha_i \leq C$ and $\sum_i \alpha_i y_i = 0$. Inputs enter this objective only through pairwise inner products.

KKT separates inactive rows, free supports, and violations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Let $m_i = y_i f(x_i)$. At an optimum, feasibility, stationarity and complementary slackness hold together:

$$\alpha_i(1 - \xi_i - m_i) = 0, \quad (C - \alpha_i)\xi_i = 0.$$

Derive the free-support case

$$0 < \alpha_i < C \Rightarrow C - \alpha_i > 0 \Rightarrow \xi_i = 0.$$

$$\alpha_i > 0 \Rightarrow m_i = 1 - \xi_i.$$

Combining both results gives $m_i = 1$.

Coefficient	Required margin
$\alpha_i = 0$	$m_i \geq 1$
$0 < \alpha_i < C$	$m_i = 1, \xi_i = 0$
$\alpha_i = C$	$m_i = 1 - \xi_i \leq 1$

For $m_i = .4$, slack is .6. Its sign is correct, although its margin is too small.

Equalities can be degenerate. Nonzero coefficients identify the rows used in prediction.



Recover the intercept from a free support vector

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For $0 < \alpha_i < C$, KKT gives $y_i f(x_i) = 1$, hence $f(x_i) = y_i$. Use this equality to recover the bias.

$$b_i = y_i - w^T x_i = y_i - \sum_j \alpha_j y_j x_j^T x_i.$$

$$x = 1, y = 1 : \quad b = 1 - (1)(1) = 0,$$

$$x = -1, y = -1 : \quad b = -1 - (1)(-1) = 0.$$

Both free support vectors give the same bias. Numerically, average the bias estimates from several free support vectors.

Check. Do not apply this equality to an arbitrary at-bound support with positive slack. If there are no free supports, use feasible KKT bounds for b .



Solve the same two-point problem in the dual

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Take $x_- = -1, y_- = -1$ and $x_+ = 1, y_+ = 1$, with $C > 1/2$. Label balance requires $\alpha_- = \alpha_+ = a$.

$$x_- = -1, y_- = -1 \quad x_+ = 1, y_+ = 1$$

$$C > 1/2 \quad \alpha_- = \alpha_+ = a$$

$$w = a(-1)(-1) + a(+1)(+1) = 2a$$

$$W(a) = 2a - \frac{1}{2}(2a)^2 = 2a - 2a^2$$



Worked calculation

$$W'(a) = 2 - 4a = 0 \Rightarrow a = \frac{1}{2}, \quad w = 1, \quad b = 0$$

$$W = 1/2 \quad \|w\|^2/2 = 1/2$$

Check. This calculation verifies the representation. If $C < 1/2$, the coefficient cap changes the solution and slack is needed.

Prediction is a signed similarity sum

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Only rows with nonzero α_i contribute. Their labels set the sign of each contribution, while the coefficient sets its strength.

$$f(x) = \sum_{i \in SV} \alpha_i y_i x_i^T x + b$$

$$f(0.6) = \frac{1}{2}(-1)(-1 \cdot 0.6) + \frac{1}{2}(+1)(1 \cdot 0.6) = 0.6$$

$$x = -0.2 \quad w = 1, b = 0$$

Check. Signed contributions are not individual probabilities or votes. Kernelization changes the similarity calculation, not the need to sum them.



SMO updates a pair to preserve label balance

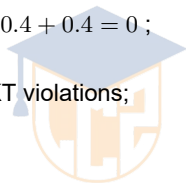
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Changing one nonzero coefficient alone usually violates $\sum_i \alpha_i y_i = 0$. Sequential minimal optimization fixes all other coefficients and optimizes a feasible pair.

$$y_i \Delta \alpha_i + y_j \Delta \alpha_j = 0, \quad y_i = -1, y_j = 1 \Rightarrow \Delta \alpha_i = \Delta \alpha_j.$$

Set $C = 1$, the coefficient pair changes from $(0.2, 0.2)$ together to $(0.4, 0.4)$, preserving $-0.4 + 0.4 = 0$; changing only the second instead gives $-0.2 + 0.4 = 0.2 \neq 0$.

Check. The pair optimum is clipped to its feasible interval. Working-set selection uses KKT violations; numerical tolerance controls when updates stop.



A feature map can turn rings into a threshold

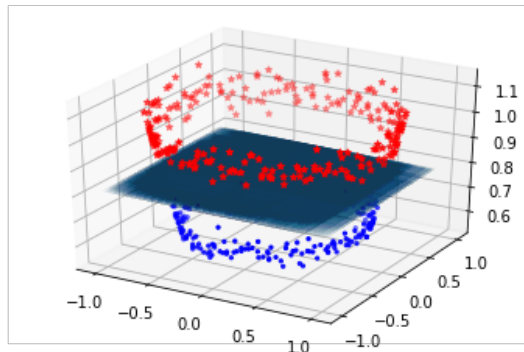
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The ring pattern motivates a new coordinate: squared radius. A threshold in this coordinate becomes a circle in the input plane.

$$z = x_1^2 + x_2^2 \quad f = z - 1$$

$$f(.5, 0) = .25 - 1 = -.75$$

$$f(1.5, 0) = 2.25 - 1 = 1.25$$



The kernel trick computes a mapped inner product

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A kernel $K(x, z) = \phi(x)^T \phi(z)$ supplies exactly the quantity required by the dual and prediction. We may compute it without explicitly forming every mapped coordinate.

$$K(x, z) = \phi(x)^T \phi(z)$$

$$f(x) = \sum_{i \in \text{SV}} \alpha_i y_i K(x_i, x) + b$$

$$W(\alpha) = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

$$K(x, z) = x^T z$$

Check. The trick avoids explicit features, not training data dependence. Prediction still needs kernels to retained support vectors.



Expand a quadratic kernel and verify one calculation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For two-dimensional inputs, $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T \mathbf{z})^2$ includes squares and an interaction. The square-root factors preserve the exact inner product.

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

$$(1 + \mathbf{x}^T \mathbf{z})^2 = (1 + 4)^2 = 25,$$

$$\phi(\mathbf{x})^T \phi(\mathbf{z}) = 1 + 4 + 4 + 4 + 8 + 4 = 25.$$

A kernel computes the same result without explicitly constructing every mapped coordinate.



A valid kernel must have positive-semidefinite Gram matrices

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

An inner-product kernel is symmetric. For every finite sample set its Gram matrix must satisfy $v^T K v \geq 0$ for every coefficient vector v .

$$K = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \quad v = (1, -1)^T \Rightarrow v^T K v = -2 < 0$$

Check. Symmetry alone is insufficient. A visually plausible similarity or an arbitrary sigmoid parameter choice may fail validity.



RBF similarity has a distance scale you can calculate

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The radial basis kernel compares squared distance after the chosen preprocessing. γ sets how quickly similarity falls as two rows separate.

$$K(\mathbf{x}, \mathbf{z}) = \exp(-\gamma \|\mathbf{x} - \mathbf{z}\|^2)$$

$$\gamma = 0.5 : \quad d = 0 \Rightarrow 1, \quad d = 1 \Rightarrow 0.607, \quad d = 2 \Rightarrow 0.135$$

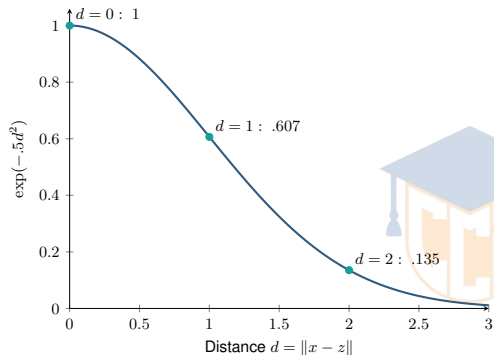


Check. A kernel value is a similarity, not the posterior probability that two rows have the same label.

Read how RBF similarity falls with distance

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For an RBF kernel with $\gamma = .5$, distances 0, 1 and 2 give similarities 1, about .607 and about .135. A larger distance lowers similarity; feature scaling therefore changes the kernel.



Gamma controls locality, while bandwidth controls width

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Write the same RBF kernel as $\exp(-\|x - z\|^2/(2\sigma^2))$. Then $\gamma = 1/(2\sigma^2)$: larger gamma means a narrower neighborhood.

$$\exp(-\|x - z\|^2/(2\sigma^2)) \quad \gamma = 1/(2\sigma^2)$$

$$K = e^{-1} \text{ at } d = \frac{1}{\sqrt{\gamma}} \quad \gamma = 0.25 \Rightarrow d = 2; \quad \gamma = 4 \Rightarrow d = 0.5$$



Check. Locality changes boundary flexibility. A narrower kernel can fit small islands, but noise may create those islands too.

C and gamma control different parts of flexibility

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Gamma determines the locality of similarities; C determines the price of training margin violations. Tune them jointly because one may partly compensate for the other.

	Smaller C	Large C
Smaller γ	Broad influence; more margin violations tolerated	Broad influence; stronger margin-violation penalty
Large γ	Local influence; some margin violations tolerated	Local influence and strong fitting pressure; small regions may form

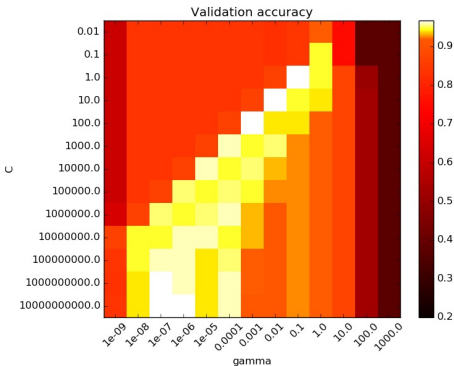
Check. These are structural tendencies. The selected combination must be supported by validation under the intended sampling process.

Read a validation heatmap as a region of evidence

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Each cell corresponds to one (C, γ) pair evaluated with the same folds and preprocessing. Read axis direction and color scale before locating a high-scoring region.

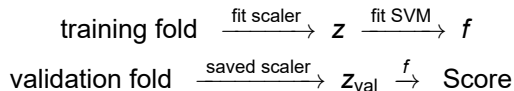
Check. The reference heatmap illustrates interaction on its dataset. Its best cell is not a universal default for inspection data.



Put scaling inside each validation fold

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Search over a pipeline, not over a classifier trained on globally transformed data. Every training fold fits its own scaler and SVM; the paired validation fold receives only their transforms and predictions.



Multiclass SVM 结合数字工程研究中心

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For M categories, one-vs-rest trains M binary models; one-vs-one trains $M(M - 1)/2$ pairwise models. Their combination rules are part of the estimator.

$$M = 4 : \quad \text{OvR} = 4 \text{ models} , \quad \text{OvO} = 6 \text{ models}$$

Check. Raw decision scores need a stated combination rule. They are not a vector of probabilities merely because one score is returned per class.



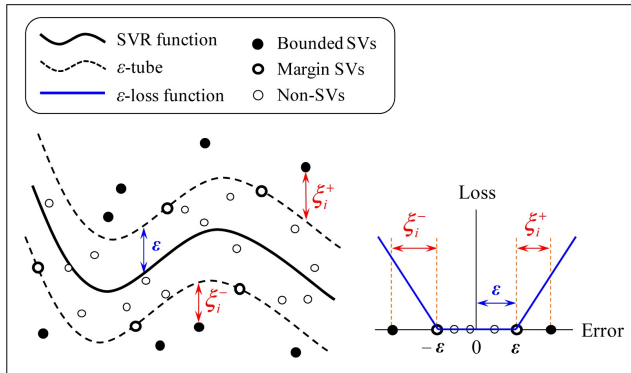
Regression replaces the corridor with an error tube

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

SVR places an ε -wide tolerance on each side of a numeric prediction. Errors within that tube cost zero; larger errors pay only for the excess distance.

Support vector regression predicts a continuous target. Around $f(x)$ place a tube of half-width ε : residuals within the tube contribute no data-fit loss.

Check. The tube is measured vertically in target units, not by perpendicular distance in the input space. The full vertical width is 2ε .



Calculate epsilon-insensitive loss on real residuals

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Epsilon is measured in target units. For wear measured in millimeters, a tolerance 0.1 mm and 1 mm represent very different requirements, even if both are familiar decimal parameter values.

$$y = 2.0 \text{ mm}, \quad f = 2.6 \text{ mm}, \quad |r| = .6 \text{ mm}$$

$$\ell_{.1} = .5 \text{ mm}, \quad \ell_1 = 0$$

Check. Instrument noise and operational tolerance are related considerations, not automatically the same quantity. Report the selected tube in original output units.

Let $r = y - f(x)$. The loss is zero inside the tube and grows linearly with only the excess beyond its edge.

$$r = y - f(x)$$

$$\ell_{\epsilon}(r) = \max(0, |r| - \epsilon)$$

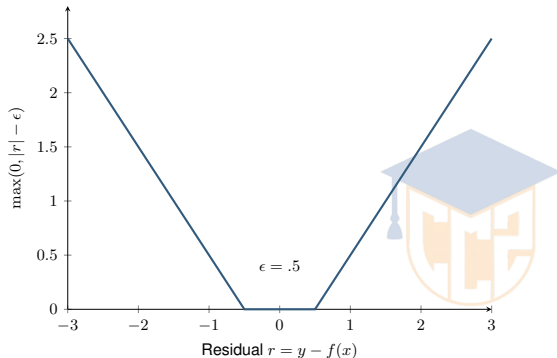
$$\epsilon = 0.2 : \quad r = (0.1, -0.2, 0.5, -0.7) \Rightarrow \ell = (0, 0, 0.3, 0.5)$$

Check. Epsilon ignores small residuals; it does not remove the regularization penalty or declare measurement errors nonexistent.

Only the residual outside the tube incurs loss

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

With $\epsilon = .5$, residuals between $-.5$ and $.5$ have zero loss. Outside this interval, loss increases linearly. The tolerance uses the units of the response.



Calculate both tube slacks for a batch of predictions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Fix prediction $\mathbf{f} = 2$ and tolerance $\epsilon = .2$. The smallest feasible slacks are $\xi = \max(y - \mathbf{f} - \epsilon, 0)$ and $\xi^* = \max(\mathbf{f} - y - \epsilon, 0)$. Only excess beyond the tube is penalized.

True response y	$y - \mathbf{f}$	ξ	ξ^*
2.5	0.5	0.3	0
1.5	-0.5	0	0.3

$$y_i - \mathbf{f}(x_i) \leq \epsilon + \xi_i, \quad \mathbf{f}(x_i) - y_i \leq \epsilon + \xi_i^*, \quad \xi_i, \xi_i^* \geq 0.$$

$$\min \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_i (\xi_i + \xi_i^*).$$

Use ξ for positive residual excess and ξ^* for negative residual excess. With epsilon 0.2, every row needs at most one strictly positive minimal slack.

$$\mathbf{y} = (1, 2, 4), \quad \mathbf{f} = (1.1, 2.6, 3.6), \quad \mathbf{r} = (-.1, -.6, .4)$$

$$\xi = (0, 0, .2), \quad \xi^* = (0, .4, 0), \quad \sum_i (\xi_i + \xi_i^*) = .6$$

$C = 3$ gives loss term $= 3 \times 0.6 = 1.8$. With an additional $\frac{1}{2} \|\mathbf{w}\|^2 = 2$ Full objective $= 2 + 1.8 = 3.8$.

Price outside-tube residuals when comparing two regression fits

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A has norm penalty .5 and total excess error 3; B has norm penalty 2 and excess error 1.

$$J_A = .5 + 3C, \quad J_B = 2 + C.$$

$$C = .25 : J_A = 1.25 < J_B = 2.25,$$

$$C = 2 : J_B = 4 < J_A = 6.5,$$

$$J_A = J_B : 2C = 1.5 \Rightarrow C = .75.$$



Dual SVR predicts with signed coefficient differences

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Two tube constraints need two nonnegative multipliers. Here y_i is a numerical target, not a binary class label, and $f_i = w^T x_i + b$.

$$y_i - f_i \leq \epsilon + \xi_i, \quad f_i - y_i \leq \epsilon + \xi_i^*.$$

$$L = \frac{1}{2} \|w\|^2 + C \sum_i (\xi_i + \xi_i^*)$$

$$+ \sum_i \alpha_i (y_i - f_i - \epsilon - \xi_i)$$

$$+ \sum_i \alpha_i^* (f_i - y_i - \epsilon - \xi_i^*) - \sum_i (\mu_i \xi_i + \mu_i^* \xi_i^*).$$

$$\partial_w L = 0 \Rightarrow w = \sum_i (\alpha_i - \alpha_i^*) x_i, \quad \partial_b L = 0 \Rightarrow \sum_i (\alpha_i - \alpha_i^*) = 0.$$

$$\partial_{\xi_i} L = 0 \Rightarrow 0 \leq \alpha_i \leq C, \quad 0 \leq \alpha_i^* \leq C.$$

$$f(x) = \sum_i (\alpha_i - \alpha_i^*) K(x_i, x) + b.$$



Read the SVR dual objective with two coefficient families

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Let $\beta_i = \alpha_i - \alpha_i^*$. The quadratic term penalizes function norm, the epsilon term charges both families, and observed target values reward signed coefficients.

$$\beta_i = \alpha_i - \alpha_i^*$$

$$\begin{aligned} \max W = & -\frac{1}{2} \sum_{i,j} \beta_i \beta_j K_{ij} - \epsilon \sum_i (\alpha_i + \alpha_i^*) + \sum_i y_i \beta_i \\ & 0 \leq \alpha_i, \alpha_i^* \leq C, \quad \sum_i \beta_i = 0 \end{aligned}$$



Check. The coefficient difference predicts the function, but the epsilon penalty depends on the sum of nonnegative coefficients. Keeping only beta too early loses that structure.

Solve a two-point SVR in the dual, including epsilon

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Use $(x, y) = (0, 0), (1, 1)$, a linear kernel, epsilon 0.1 and $C = 1$. Let the first row have $\alpha_0^* = a$ and the second $\alpha_1 = a$, giving $\beta = (-a, a)$.

$$(x, y) = (0, 0), (1, 1) \quad C = 1$$

$$\alpha_0^* = a \quad \alpha_1 = a$$

$$\beta = (-a, a)$$

$$w = a, \quad W(a) = -\frac{1}{2}a^2 - .2a + a = .8a - \frac{1}{2}a^2$$

$$W'(a) = .8 - a = 0 \Rightarrow a = .8, \quad w = .8, \quad b = .1$$



Recover the SVR intercept from the correct tube edge

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A free positive-residual support satisfies $y_i - f_i = \epsilon$; a free negative-residual support satisfies $f_i - y_i = \epsilon$. Each edge gives a different sign of epsilon in the intercept formula.

$$y_i - f_i = \epsilon \quad f_i - y_i = \epsilon$$

$$0 < \alpha_i < C: \quad b = y_i - \epsilon - \sum_j \beta_j K(\mathbf{x}_j, \mathbf{x}_i)$$

$$0 < \alpha_i^* < C: \quad b = y_i + \epsilon - \sum_j \beta_j K(\mathbf{x}_j, \mathbf{x}_i)$$

$$w = 0.8 \quad b = 0 + 0.1 - 0 = 0.1$$

$$b = 1 - 0.1 - 0.8 = 0.1$$



C prices excess error; epsilon defines what counts as excess

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For fixed predictions, enlarging epsilon reduces the number of nonzero loss terms. Increasing C makes the remaining excesses more expensive. Neither parameter is a general accuracy knob.

Residual $ r $	$\epsilon = 0.2$ gives loss	$\epsilon = 0.5$ gives loss
0.1	0	0
0.4	0.2	0
0.8	0.6	0.3

Check. Larger tubes often yield fewer active rows, but the final fitted support count is data-dependent. Scaling the target changes parameter interpretation.



Wider tubes can reduce slope before changing residual signs

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For the two targets 0 and 1 at inputs 0 and 1, a sufficiently large C permits a zero-excess optimum. The shortest feasible slope depends on the tube overlap.

$$w = \max(0, 1 - 2\epsilon), \quad b = \epsilon \quad (0 \leq \epsilon \leq .5)$$
$$\epsilon = 0, .1, .4 : \quad (w, b) = (1, 0), (.8, .1), (.2, .4)$$



Forecast demand with a chronological validation design

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For a bike-demand example, predict a future count using features available at the prediction time. Earlier periods train the model, later periods validate it, and the final period tests it.

earlier training → later validation → future test

Check. RBF predictions outside the training feature range are not guaranteed to extrapolate a trend. Future information in features invalidates the exercise.



Construct chronological folds with a future test block

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Reserve June for the final test. Select the kernel, C and ϵ using only the validation folds.

Phase	Training months	Evaluation month
Fold 1	January–March	April validation
Fold 2	January–April	May validation
After freezing settings	Refit on January–May	One June test

Check. Fit scaling on training months within each fold. Use weather forecasts available at prediction time, rather than summaries known only after the day ends.



Choose the model from the question, not from its name

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

One instrument may classify a defect or predict wear. The measurements can match while the outputs, losses, and error units differ.

	Defect classification	Wear regression
Target	$y \in \{-1, +1\}$	Continuous wear units
Data penalty	$\max(0, 1 - yf)$	$\max(0, y - f - \epsilon)$
Design	Class costs and C kernel	Tolerance and C kernel
Evaluation	Recall, balanced accuracy	MAE, residual patterns



Explain the complete chain and carry similarity forward

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Each calculation below answers a different question. Check score normalization, violation cost, kernel scale and response units before comparing models. A decision score alone does not supply a calibrated probability.

Given quantity	What is obtained
$w = (3, 4), b = -5, x = (1, 1)$ Signed labeled score $yf = 0.4$	Distance $= 3 + 4 - 5 /5 = 0.4$ Minimum slack and hinge loss both equal .6
$0 < \alpha_i < C$ RBF Distance $d = 2, \gamma = 0.5$	By KKT. $yf = 1, \xi_i = 0$ $e^{-0.5(2^2)} = e^{-2} \approx 0.1353$
Residual $ r = 0.5, \epsilon = 0.2$	Loss outside the tube $= 0.5 - 0.2 = 0.3$

The next lecture has no target class labels. We must specify a useful similarity rule and purpose for grouping.