

Pattern Recognition and Machine Learning

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Pattern Recognition and Machine Learning

Lecture 01 Pattern Recognition and Machine Learning



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
Objects, measurements and labels	45
Prediction, uncertainty and evidence	45
The recognition system	45

Six fruits and one new measurement

Fit class centers; predict an unseen fruit; inspect the errors.

The numerical steps are shown on the slides.



Recognition must tolerate variation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A fruit is the object, its weight is a measurement, and its species is the label. Known labels let us estimate class centers before predicting a new fruit.

Writers produce different shapes for the same digit. Different digits can also look similar. Recognition must work on handwriting excluded from training.



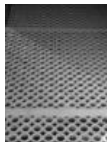
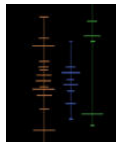
An object becomes measurements, then a prediction

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The object exists in the world. Measurements describe it. The predicted class is the model's answer, which can be wrong.

$$\mathbf{x} = (x_1, \dots, x_{784}) \in \mathbb{R}^{784}, \quad \hat{y} = f(\mathbf{x}) \in \{0, 1, \dots, 9\}.$$

Texture, shape and repeated structure can define different notions of a pattern.



Samples, features and labels play different roles

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

Learning requires a specified task, training experience and evaluation rule. Remembering training images gives weak evidence about recognizing a new writer.

Each row describes one fruit. Weight and red-pixel fraction are inputs. The species supplies the training label, which remains separate from the model prediction.

Case	Weight in grams	Red-pixel fraction	Label y
Fruit 1	40	0.80	Tomato
Fruit 2	50	0.90	Tomato
Fruit 3	110	0.30	Mandarin

Check. An 85 g tomato may receive a mandarin prediction. The measured input, the true label and the predicted label remain separate.



One feature or two: keep rows and labels aligned

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Two measured features

$$X = \begin{bmatrix} 40 & .8 \\ 50 & .9 \\ 60 & .7 \\ 100 & .2 \\ 110 & .3 \\ 120 & .1 \end{bmatrix}, \quad y = \begin{bmatrix} T \\ T \\ T \\ M \\ M \\ M \end{bmatrix}.$$

Six rows are six fruits. Columns are weight in grams and red-pixel fraction: $n = 6$, $d = 2$.

The same fruits, weight only

$$X_{\text{weight}} = \begin{bmatrix} 40 \\ 50 \\ 60 \\ 100 \\ 110 \\ 120 \end{bmatrix}, \quad x_{\text{new}} = 75 \text{ g.}$$

Now $d = 1$. The row order and labels stay unchanged. The new fruit has an input but an unknown label.

Two operations on a feature vector

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The vector $(3, 4)$ has length 5. Dividing by 5 preserves direction and gives unit length. This transformation removes magnitude, which may matter for the task.

$$\mathbf{a} = (3, 4), \quad \|\mathbf{a}\|_2 = \sqrt{3^2 + 4^2} = 5, \quad \mathbf{u} = \mathbf{a}/5 = (0.6, 0.8).$$

An inner product multiplies each feature by its coefficient and adds the contributions. Here the score is 2. A separate decision rule converts the score to a class.

$$\mathbf{w} = (1, -2), \quad \mathbf{x} = (4, 1), \quad \mathbf{w}^T \mathbf{x} = 1(4) + (-2)(1) = 2.$$

Input change	Change in the inner product	Reason
x_1 increases by 1	+1	Its coefficient is 1.
x_2 increases by 1	-2	Its coefficient is -2.
Add the orthogonal vector $(2, 1)$	0	The positive and negative contributions cancel.

Changing units can change a naive distance

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Training and prediction must use the same feature order, units and measurement procedure. A shape-compatible input can still carry the wrong meaning.

Only the weight unit changed. A scale-aware representation or a task-specific distance must state how feature differences should trade off.

$$d^2 = (\Delta \text{ weight})^2 + (\Delta \text{ red fraction})^2.$$

Comparison	Weight delta	Color delta	Squared distance
A	10 g	0.8	100.64
B	20 g	0.1	400.01
A, kilograms	0.01 kg	0.8	0.6401
B, kilograms	0.02 kg	0.1	0.0104

Learning settings differ in the information available

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

First define the target

Tomato or mandarin is a class. Weight in grams is a quantity.

Class codes 0 and 1 name categories. Their difference has no physical unit. A 10 g weight difference does.

Then check the training information

Setting	Information received
Supervised	Inputs with known targets
Unsupervised	Inputs without target labels
Semi-supervised	Labeled and unlabeled inputs
Reinforcement	States, actions and rewards

The same images can support different tasks and learning settings.

Labels and groups answer different questions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Fruit labels let us check predictions. The customer table has no target labels: grouping depends on the selected measurements and similarity rule.

Customer	Visits	Spend
A	2	20
B	3	22
C	15	80
D	16	75



Check. These groups reflect selected measurements and similarity. They do not automatically reveal a true social or business category.

Additional data and interactive feedback supply different learning signals

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

More unlabeled data help only when their structure relates to the task. Delayed rewards require reasoning about action consequences.

Example	How learning gets information
Semi-supervised digits	1,000 labeled and 100,000 unlabeled images. Similarity can help if it agrees with digit identity.
Self-supervised text	Hide a token and predict it from context. The data themselves provide training targets.
Reinforcement game	Choose an action, observe a changed state and later reward. A winning game does not label every move as optimal.

Fit class centers, then predict a new fruit

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The three tomato weights average 50 g, and the three mandarin weights average 110 g. These fitted centers summarize the labeled training records.

Known class	Training weights / g		
Tomato	40	50	60
Mandarin	100	110	120

$$\mu_T = \frac{40 + 50 + 60}{3} = 50 \text{ g},$$

$$\mu_M = \frac{100 + 110 + 120}{3} = 110 \text{ g}.$$

Evaluate both distances. $\arg \min$ returns the class that gives the smaller distance, not the distance value.

$$\hat{y}(x) = \arg \min_{c \in \{T, M\}} |x - \mu_c|.$$

Weight / g	$ x - 50 $	$ x - 110 $	Label
midrule 75	25	35	T
95	45	15	M
80	30	30	M (tie)

Check. Prediction uses the existing centers. A new input does not automatically change them. At equal distance, our implementation chooses mandarin.

A model family, its parameters and its learning algorithm

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A prediction now follows a fixed rule. We can ask why it may fail, what its output means, and which unseen observations would test it.

Nearest-center classification is the model form. The values 50 and 110 are fitted parameters. Computing a mean within each class is the training procedure.

```
center_t = sum([40, 50, 60]) / 3
center_m = sum([100, 110, 120]) / 3
def predict(x):
    return "Tomato" if abs(x-center_t) < abs(x-center_m) else "Mandarin"
print(predict(75)) # Tomato
```

The decision boundary follows from equal distances

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Equal distances define the boundary. Squaring preserves their order because distances are nonnegative. Expanding both sides gives the threshold 80 g.

$$(x - 50)^2 = (x - 110)^2$$

$$x^2 - 100x + 2500 = x^2 - 220x + 12100$$

$$120x = 9600$$

$$x = 80.$$

$$\hat{y}(x) = \begin{cases} T, & x < 80, \\ M, & x \geq 80. \end{cases}$$



Two-feature centroid prediction is a complete vector calculation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The weight-only calculation used one coordinate. This separate geometric example uses two already scaled coordinates. Its class centers are supplied.

$$\mu_T = (0, 0), \quad \mu_M = (4, 2), \quad \mathbf{x} = (1, 1).$$

Distance to tomato

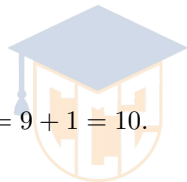
$$d_T^2 = (1 - 0)^2 + (1 - 0)^2 = 1 + 1 = 2.$$

$2 < 10$, so predict tomato. Comparing square roots gives the same ordering.

This is a distance-based class decision. It does not yet give a class probability.

Distance to mandarin

$$d_M^2 = (1 - 4)^2 + (1 - 2)^2 = 9 + 1 = 10.$$



A hard label and a probability express different information

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A label names a class, a distance measures separation, and a probability describes uncertainty under a model. Nearest-center distances alone do not supply class probabilities.

$$P(Y = M \mid \mathbf{x}) = 0.65$$

Condition to check

The centroid calculation alone does not produce calibrated probabilities. A smaller distance cannot be read directly as a confidence percentage.

Output	Example	What it means
Class label	Mandarin	A chosen category.
Distance	15 g	Separation from a fitted center.
Probability	$P(Y = M \mid \mathbf{x}) = .65$	An uncertainty assignment from an additional probability model.

Conditional probabilities use a selected denominator

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

The numerator describes the same 48 fruits. Conditioning on tomato selects 60 fruits, while conditioning on red selects 56. Reversing the condition changes the question and the denominator.

Among 100 fruits	red	not red	Sum
Tomato	48	12	60
Mandarin	8	32	40
Sum	56	44	100

$$P(\text{red} \mid T) = 48/60 = 0.8, \quad P(T \mid \text{red}) = 48/56 \approx 0.857.$$

Unseen fruits reveal an error that training examples hide

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The model must work beyond the exact fruits that determined its parameters. The heavy tomato shows that weight alone can be insufficient.

Weight in grams	True class	Predicted	0-1 loss
75	T	T	0
95	M	M	0
85	T	M	1
115	M	M	0

$$\text{accuracy} = 3/4 = 75\%, \quad \bar{\ell} = (0 + 0 + 1 + 0)/4 = 0.25.$$



Expectation is a probability-weighted average

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A simple fruit-weight model.

Suppose weight W is 50 g with probability .6 and 110 g with probability .4. These are model outcomes, not all individual training weights.

$$P(W = 50) = .6, \quad P(W = 110) = .4.$$

$$E[W] = 50(.6) + 110(.4) = 30 + 44 = 74 \text{ g.}$$

Quantity	Value	Meaning
One outcome	50 or 110 g	A single draw from this model
Expected weight	74 g	Probability-weighted average
Observed mean	Depends on the sample	Finite draws need not average 74 g



Variance describes spread around an average

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Subtract the mean 50 g, square the deviations, then average them using divisor 3. This gives 66.67 g². Its square root is 8.16 g.

Worked calculation

$$x = (40, 50, 60), \quad \bar{x} = 50, \quad \frac{1}{3} \sum_i (x_i - 50)^2 = \frac{200}{3} \text{ g}^2.$$

Summary	Value	Interpretation
Mean	50 g	Center of the three observations.
Mean squared deviation	66.67 g squared	Spread under the divisor-3 convention.
Standard deviation	8.16 g	Square root restores the weight unit.

A loss turns a target-dependent mistake into a number

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A loss assigns a number to the consequence of a prediction. Define the target first.

$$\ell_{01}(y, \hat{y}) = \mathbf{1}\{y \neq \hat{y}\}, \quad \ell_{\text{sq}}(y, \hat{y}) = (y - \hat{y})^2.$$

Task	True target	Prediction	Loss
Classify a fruit	T	M	$\mathbf{1}\{T \neq M\} = 1$
Predict weight	50 g	60 g	$(50 - 60)^2 = 100 \text{ g}^2$
Predict weight	50 g	80 g	$(50 - 80)^2 = 900 \text{ g}^2$



For classification, zero-one loss is 0 for a correct label and 1 for an incorrect label. Squared loss prices larger numeric errors more strongly.

$$\widehat{R}(f) = \frac{1}{n} \sum_{i=1}^n \ell(y_i, f(x_i)).$$

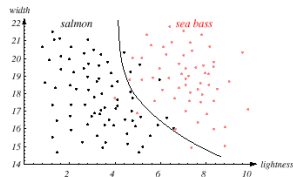
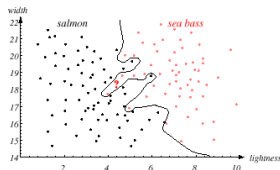
For the preceding four unseen fruits, losses $(0, 0, 1, 0)$ give average loss $(0 + 0 + 1 + 0)/4 = .25$.

Model complexity can follow the signal or memorize the sample

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Training estimates parameters, validation selects methods, and testing evaluates the fixed choice. To test recognition of new writers, keep every image from one writer in the same split.

Both panels contain the same training fish. Only the boundary changes. A complex boundary may fit individual observations, so compare performance on data excluded from fitting.



Five families ask different modeling questions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

We will learn one mechanism at a time. For every family, identify the input, target and evidence before using its formulas.

Family	Main question	Lectures
Generative classifier	How does each class generate measurements?	3
Discriminative classifier	Which score or decision rule predicts the target?	4–5, 8–9
Neural network	Which intermediate features help prediction?	15–18
Clustering / dimension reduction	How can unlabeled inputs be organized?	10–12
Ensemble	How can several predictors work together?	13–14



XOR shows why a different feature can simplify a boundary

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The class-1 corners lie on opposite diagonals, so one affine boundary cannot separate them. For binary inputs $x_1, x_2 \in \{0, 1\}$, the feature $(x_1 - x_2)^2$ makes their labels separable by a threshold.

x_1	x_2	$z = (x_1 - x_2)^2$	Class y
0	0	0	0
0	1	1	1
1	0	1	1
1	1	0	0

$$z = (x_1 - x_2)^2 = x_1 + x_2 - 2x_1x_2.$$

$$\hat{y} = \mathbb{1}\{z > 0.5\}$$



Follow one recognition pipeline

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The classifier is one part of a recognition system. Acquisition, object separation and measurement determine the evidence that reaches it. A lighting error may require better acquisition. A missing property may require better representation. Each failed prediction is evidence about the whole chain.

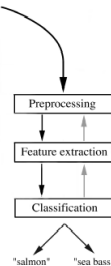
Each arrow connects an output to the next input. Follow acquisition, object separation, measurement, classification and action in order.



Fish sorting makes the full recognition chain concrete

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

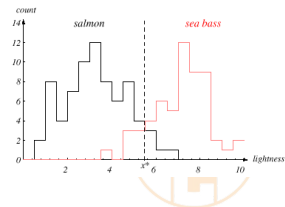
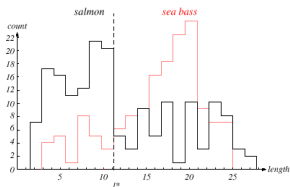
Overlapping objects or unstable illumination can spoil useful measurements before the decision rule runs.



A useful fish feature separates distributions more clearly

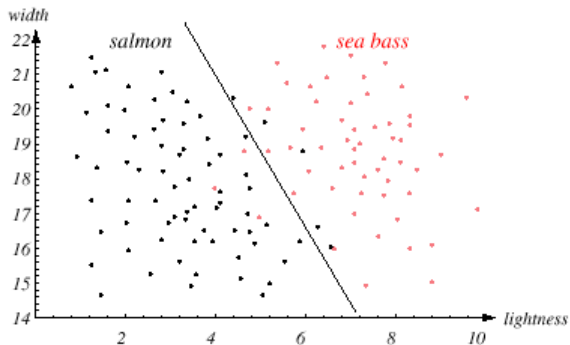
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The horizontal axes show different fish measurements. Lightness separates these distributions more clearly, but the remaining overlap still permits errors.



Read one fish as two measurements

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



Horizontal: lightness. Vertical: width.
Each mark is one fish with a known class.



A prediction may use only information available at that moment

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Estimate the mean and standard deviation from training data. With saved values 50 and 10, a new value 70 becomes 2. Apply the same transformation during validation and testing.

$$\mu_{\text{train}} = \frac{1}{n_{\text{train}}} \sum_{i \in \text{train}} x_i, \quad z = \frac{x - \mu_{\text{train}}}{s_{\text{train}}}.$$

$$z = 2$$

A field in a historical table may have been recorded after the predicted event. Training on that field teaches a rule that cannot run when needed.



A majority baseline exposes what accuracy can hide

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Overall accuracy counts correct products. It does not say whether the rare, important class is found.

True state	Count among 100 products	Always predict normal
Normal	95	95 correct
Defective	5	0 detected

$$\text{accuracy} = 95/100 = 95\%, \quad \text{defect recall} = 0/5 = 0.$$



A confusion matrix identifies which kind of error occurred

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Was the target truly positive? Did the classifier predict positive? These two yes/no questions produce four cells.

Of 10 defective parts, 8 trigger alarms and 2 are missed. Of 90 sound parts, 4 trigger false alarms and 86 pass. The four cells total 100.

	Predicted defect	Predicted normal	Sum
True defect	TP=8	FN=2	10
True normal	FP=4	TN=86	90
Sum	12	88	100

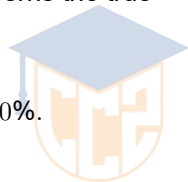


Precision, recall and accuracy use different denominators

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A reinspection workload concerns the alarms. A missed-defect requirement concerns the true defects. Both can matter in one system.

$$\text{precision} = \frac{TP}{TP + FP} = \frac{8}{12} = 66.7\%, \quad \text{recall} = \frac{TP}{TP + FN} = \frac{8}{10} = 80\%.$$
$$\text{accuracy} = \frac{TP + TN}{N} = \frac{94}{100} = 94\%.$$



Unequal error costs can favor a lower-accuracy policy

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A miss costs 100 units, a false alarm costs 1, and a correct decision costs 0.

Option	Misses	False alerts	Total errors
A	2	4	6
B	1	8	9

$$C_A = 2(100) + 4(1) = 204, \quad C_B = 1(100) + 8(1) = 108.$$



An uncertain case can be sent for a second examination

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A wrong automatic class costs 10. A second examination costs 2. Correct automatic classification costs 0.

Largest class probability	Expected automatic loss	Second-examination loss	Preferred action
0.9	$10(1 - 0.9) = 1$	2	Classify
0.6	$10(1 - 0.6) = 4$	2	Review

A threshold moves the trade-off between missed cases and alarms

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Alarm when score $\geq t$. Moving t from .7 to .4 detects one more defect and creates one more false alarm. Choose t on validation data using costs or requirements; then test the chosen rule.

Score	True label	Threshold .7 Alarm?	Threshold .4 Alarm?
.9	Defect	Yes	Yes
.6	Defect	No	Yes
.5	Normal	No	Yes
.2	Normal	No	No



A recognition task needs inputs, outputs and an independent test

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Changing inputs

A camera, lighting condition, product batch or writer can change the measurements.

Inspect missed cases as well as alarms under the new condition.

Required outputs

Classification returns a label.

Detection returns a label and location. Counting returns a number.

The target determines the evaluation.

A test on new writers

$50 \times 10 = 500$ images.

Split by writer. Fit preprocessing on training rows. Compare a baseline and inspect errors on excluded writers.

What would make our fruit rule fail?

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Question

Training centers of 50 g and 110 g give a boundary at 80 g. Does that establish a reliable species rule?

Evidence and reasoning

An 85 g tomato is a counterexample: the rule predicts mandarin. Weight separates the observed centers, while overlapping fruit weights limit the rule. Add informative measurements and evaluate fruits excluded from fitting.

Check. A few correct examples establish a calculation, not a population error rate.

