

Semi-supervised Learning

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Pattern Recognition and Machine Learning



Lecture 12 Semi-supervised 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
Self-training and multiple views	45
Generative models and S3VM	45
Graphs and constrained clustering	45

A few verified labels and many unknown targets

Check the structural assumption before turning a prediction into supervision.

The numerical steps are shown on the slides.



More observations do not automatically mean more supervision

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Unlabeled inputs reveal where observations occur, while verified labels give class meanings.

Unlabeled data reveal where inputs are common and which observations are similar.

$$\mathcal{D}_L = \{(x_i, y_i)\}_{i=1}^L, \quad \mathcal{D}_U = \{x_j\}_{j=1}^U, \quad U \gg L.$$



Unlabeled data locate dense regions, not class names

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Deploying a fault classifier next month is inductive. Labeling one fixed archive using its entire similarity graph can be transductive.

Input density can guide a classifier only when density structure and label structure agree.

$(x = 0, y = 0)$ $(x = 10, y = 1)$



Three assumptions link input structure with labels

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

Similarity must be relevant to the target, not just to background or collection conditions.

Assumption	Structural assumption
Smoothness assumption	Relevant nearby observations tend to share labels
Low-density boundary	Classes align with dense regions; boundaries cross sparse regions
Manifold assumption	Labels vary smoothly along the low-dimensional structure

Check. A shared background does not establish the same animal class. A shared recording date does not establish the same fault.

A shared background can make the cluster assumption fail

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Images cluster by bright versus dark backgrounds. The required classes are cats versus dogs, and both occur in both backgrounds.

Condition to check

An accurate input-density model is not necessarily an accurate model of the target classes.

Input group	Animals inside it	Risk of propagating one seed label
Bright background	Cats and dogs	A cat seed can give dogs the cat pseudo-label.
Dark background	Cats and dogs	A dog seed can give cats the dog pseudo-label.
More unlabeled images	Estimate the background groups more precisely	The target mismatch can become more confident.

Start with a supervised baseline and an untouched labeled check

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Reserve labeled validation and test sets before pseudo-labeling or graph construction choices.

Check. It measures agreement with a model's own targets unless independently verified.

Comparison item	Fair evidence	Misleading substitute
Real training-label budget	Keep it fixed for baseline and SSL	Give one method more verified labels.
Validation and test	Reserve true labels before choosing the method	Reuse them as pseudo-label training targets.
Reported improvement	Independent labeled performance, including each class	Agreement with the model's own targets.
Few-label variability	Repeat labeled subsets and report variation	Rely on one fortunate split.

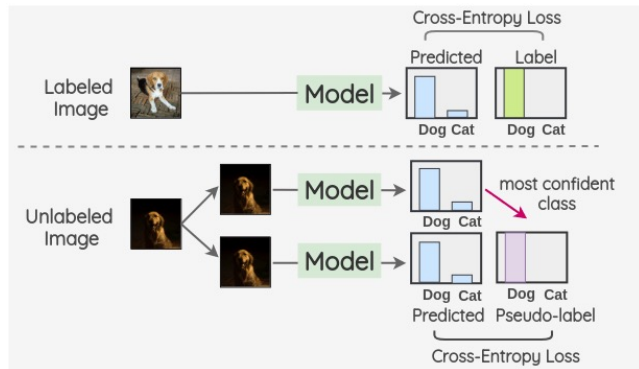
Trace a prediction becoming a training target

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Set $\hat{y} = \arg \max_y p_{\theta}(y | x)$ and confidence $c = \max_y p_{\theta}(y | x)$.

First train with verified labels. Then predict unlabeled observations and accept only those meeting the stated rule. Their new training targets remain model guesses; another training round does not verify them.

$$\hat{y} = \arg \max_y p_{\theta}(y | x)$$
$$c = \max_y p_{\theta}(y | x)$$



Worked self-training: accept two candidates, not all four

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Rejected does not mean useless or incorrect; it means evidence is insufficient under the current rule.

Candidate	$p(y = 0 \mid x)$	$p(y = 1 \mid x)$	Pseudo-label	Press $\tau = 0.90$	Acceptance
A	0.97	0.03	0		Yes
B	0.58	0.42	0		No
C	0.08	0.92	1		Yes
D	0.25	0.75	1		No



Give pseudo-labels controlled influence in the loss

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A: accepted pseudo-labeled samples; skip the second term if none are accepted.

$$\mathcal{L} = \frac{1}{L} \sum_{i \in L} \text{CE}(y_i, p_{\theta}(\mathbf{x}_i)) + \lambda \frac{1}{|A|} \sum_{j \in A} a_j \text{CE}(\hat{y}_j, p_{\theta}(\mathbf{x}_j)).$$



Self-training can fail through confidence, imbalance or feedback

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Overconfidence

Of 100 predictions near confidence .95, only 70 are correct.

$$70/100 = .70.$$

Confidence needs calibration.

Class imbalance

The same threshold accepts 900 majority cases and only 20 minority cases.

Track acceptance and validation recall separately by class.

Error reinforcement

A wrong confident target is reused as training evidence. Later confidence can rise without correcting it.

Use verified-label validation to decide when to stop.

Worked review: decide which pseudo-labels deserve influence

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A digit image shifted slightly should usually keep its label. A feature record with small measurement noise may still represent the same state.

$$\mathcal{L}_{\text{cons}} = \|p_{\theta}(y \mid \mathbf{a}_1(x)) - p_{\theta}(y \mid \mathbf{a}_2(x))\|^2.$$

A gets class 0; C gets class 1; B remains unlabeled. Calibration, class balance and independent validation still matter.



Complementary views need enough information and different errors

Self-training teaches itself; co-training exchanges selected targets between views.

Sufficiency: each view contains enough information to predict the class reasonably well.

$$p(x^{(1)}, x^{(2)} | y) = p(x^{(1)} | y)p(x^{(2)} | y)$$

Check. Practical co-training variants may relax the classical assumptions; theoretical guarantees do not transfer automatically.



Worked co-training: the stronger view teaches the uncertain view

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Exchange the target for the same sample, while training each model on its own view.

Case	Image View $p(\text{cat})$	Text View $p(\text{cat})$	Action at threshold 0.9
A	0.97	0.60	Image proposes cat to text model
B	0.55	0.04	Text proposes dog to image model
C	0.93	0.06	Confident disagreement: inspect before exchanging targets

Co-regularization encourages agreement without hard exchange

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

True labels anchor both views; unlabeled observations penalize inconsistent predictions.

$$\mathcal{L} = \mathcal{L}_L^{(1)} + \mathcal{L}_L^{(2)} + \lambda \frac{1}{U} \sum_{x \in U} \|p_1(y | x^{(1)}) - p_2(y | x^{(2)})\|^2.$$



A generative model links class names to input density

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For a mixture, sum over the missing class. For S3VM, penalize points close to the boundary.

Unlabeled data have a missing class variable; they still inform means, covariances and mixing proportions.

$$p(x, y = k \mid \theta) = \pi_k \mathcal{N}(x; \mu_k, \Sigma_k), \quad p(x \mid \theta) = \sum_k p(x, y = k \mid \theta).$$



Worked EM: calculate a soft responsibility

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

An unlabeled sample sums over possible classes before taking the logarithm.

$$\ell(\theta) = \sum_{i \in L} \log p(x_i, y_i | \theta) + \lambda \sum_{j \in U} \log \sum_k p(x_j, y = k | \theta).$$

Means are $\mu_0 = 0$, $\mu_1 = 4$; both variances are 1; priors are 0.5. The unlabeled point is $x = 1$.

$$\mu_0 = 0 \quad \mu_1 = 4$$

$$x = 1$$

$$r_0 = \frac{e^{-(1-0)^2/2}}{e^{-(1-0)^2/2} + e^{-(1-4)^2/2}} = \frac{e^{-0.5}}{e^{-0.5} + e^{-4.5}} \approx 0.982.$$
$$r_1 = 1 - r_0 \approx 0.018.$$



A better density fit can produce a worse classifier

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

One verified class-0 point is 0, one verified class-1 point is 4, and the unlabeled point 1 has responsibilities (0.982, 0.018).

$$\mu_0^{\text{new}} = \frac{0 + 0.982 \times 1}{1 + 0.982} \approx 0.495, \quad \mu_1^{\text{new}} = \frac{4 + 0.018 \times 1}{1 + 0.018} \approx 3.947.$$

Likelihood improvement checks the fitted density; labeled validation checks the classification task.



The unlabeled hinge chooses a side without a class name

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For a labeled point, the known sign gives hinge loss $\max(0, 1 - yf)$. An unlabeled point has no verified sign.

Try both possible labels at $f = .3$

Proposed y	yf	Hinge loss
--------------	------	------------

+1	.3	.7
----	----	----

-1	-.3	1.3
----	-----	-----

favorable sign gives $+.3$ margin and loss $.7$.

This term encourages distance from the boundary. It does not verify which class is correct. Labeled observations must anchor the class meanings.

Eliminate the unknown sign

$$\min_{y \in \{-1, +1\}} \max(0, 1 - yf) = \max(0, 1 - |f|).$$

At $f = -.3$, choosing $y = -1$ also gives loss $.7$. At $f = 0$, either sign gives loss 1 .

S3VM encourages a low-density classification boundary

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A labeled point must be far enough on its known correct side.

$$\min_{y \in \{-1, +1\}} \max(0, 1 - yf) = \max(0, 1 - |f|).$$

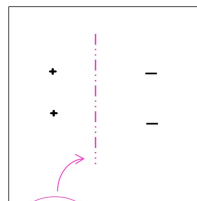
$$\mathcal{J} = \frac{1}{2} \|w\|^2 + C_L \sum_{i \in L} \max(0, 1 - y_i f_i) + C_U \sum_{j \in U} \max(0, 1 - |f_j|).$$



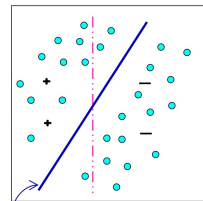
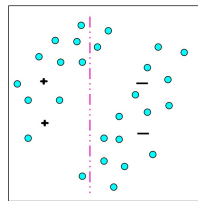
Unlabeled density can change a boundary

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The labeled points anchor class meanings. Unlabeled points reveal dense regions and encourage the boundary to pass between them. This helps only when low-density separation agrees with the required classes.



SVM
Labeled data *only*



Transductive SVM

Worked S3VM: loss is largest at the boundary

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

At $f = 0$, the point is on the boundary and receives the largest unlabeled penalty.

Unlabeled score $f(x)$	-1.4	-0.3	0	0.8
$\max(0, 1 - f)$	0	0.7	1	0.2

$f = 0 \Rightarrow$ Unlabeled loss = 1



Worked review: three ways to use one unlabeled point

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The supervised hinge objective is convex in (w, b) ; adding $\max(0, 1 - |f|)$ makes the objective nonconvex.

Hard target 0 is inferred, not verified; mixture updates use fractional weights; S3VM loss is 0.8; inspect view validity, pairing and calibration.

Method	Use of this point
Self-training	Probability (0.98, 0.02) ; crossing the threshold produces hard target 0
Generative method	Retain fractional responsibilities (0.98, 0.02) ; fractional weights contribute to both class updates
S3VM	For current score $f = 0.2$, the unlabeled loss is $1 - 0.2 = 0.8$

These methods use hard targets, fractional weights and margin preferences respectively. Unlabeled data enter their objectives in different ways.

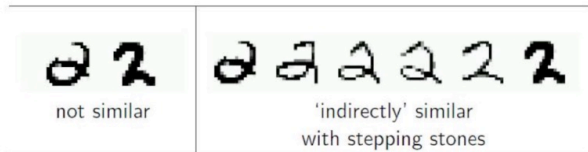
A graph makes the smoothness assumption explicit

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Nodes are observations. Strong edges indicate that similar scores are preferred.

The same neighborhood structure that supported embedding can support label propagation.

Handwritten digits recognition with pixel-wise Euclidean distance



Build weights using a relevant representation and metric

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Choose k -nearest-neighbor or radius edges and make the graph symmetric where required.

$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad \text{for selected neighbor edges; otherwise } W_{ij} = 0.$$

 $\sigma = 1$

$$\sigma = 1 : \quad d = 1 \Rightarrow W = e^{-1/2} \approx 0.607; \quad d = 3 \Rightarrow W = e^{-9/2} \approx 0.011.$$



Graph energy gives the weighted-average equation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

f_i is a class score at node i ; it is the quantity we want to learn.

$$D_{ii} = \sum_j W_{ij}, \quad L_G = D - W,$$

$$E(f) = \frac{1}{2} \sum_{i,j} W_{ij} (f_i - f_j)^2 = f^\top L_G f.$$

$$f_A = 0 \quad f_C = 1$$

It connects to fixed scores 0 and 1 with weights 1 and 3. Count each undirected edge once:

$$E(u) = 1(u - 0)^2 + 3(u - 1)^2.$$

$$E'(u) = 2u + 6(u - 1) = 8u - 6 = 0 \Rightarrow u = 6/8 = .75.$$

$$u = \frac{1(0) + 3(1)}{1 + 3} = .75.$$

Worked calculation

With a threshold of .5, the point receives class 1. Its score .75 is inferred from the graph and fixed anchors, rather than a newly verified label.

Worked iteration: two unknown nodes settle between fixed anchors

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Known anchors are $f_A = 0$, $f_D = 1$ on the chain A–B–C–D. Every edge has weight 1.

Minimize the edge energy

$$E = f_B^2 + (f_B - f_C)^2 + (f_C - 1)^2.$$

$$\partial_{f_B} E = 2f_B + 2(f_B - f_C) = 0,$$

$$\partial_{f_C} E = 2(f_C - f_B) + 2(f_C - 1) = 0.$$

$$2f_B - f_C = 0, \quad -f_B + 2f_C = 1.$$

Thus $f_B = 1/3$, $f_C = 2/3$.

Or repeatedly average old neighbors

Iteration	f_B	f_C
0	.5	.5
1	.25	.75
2	.375	.625
3	.3125	.6875
Limit	1/3	2/3



$$f_B^{t+1} = f_C^t/2, \quad f_C^{t+1} = (f_B^t + 1)/2.$$

Keep anchors fixed. Use the previous full iteration for both updates.

General block system: U denotes unknown nodes and L labeled nodes. Partition the graph Laplacian L_G and solve $(L_G)_{UU}f_U = -(L_G)_{UL}f_L$. Each unknown component needs a labeled anchor.

Multiclass propagation uses one score column per class

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Apply neighbor averaging to each score column, keeping verified rows fixed.

$$\mathbf{F} \in \mathbb{R}^{(L+U) \times C}, \quad \mathbf{Y}_i = (0, \dots, 1, \dots, 0) \text{ for a verified class,} \quad \hat{y}_i = \arg \max_c F_{ic}.$$



A graph can spread errors as efficiently as labels

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Each step propagates through the normalized graph and restores some influence from initial labels.

$$S = D^{-1/2}WD^{-1/2}, \quad F^{(t+1)} = \alpha SF^{(t)} + (1 - \alpha)Y, \quad 0 < \alpha < 1.$$

$$F^* = (1 - \alpha)(I - \alpha S)^{-1}Y.$$

$$F^* = \alpha SF^* + (1 - \alpha)Y \Rightarrow (I - \alpha S)F^* = (1 - \alpha)Y.$$

Fix or qualify the graph assumption before adding more propagation iterations.



Partial supervision can specify relationships without class names

Seed label: a few observations have known cluster or class identities.

Check. Different identities need not be different clusters for every other clustering objective.

Knowledge	Required relation	What it does not specify
Must-link A, B	A and B share a cluster	Its arbitrary cluster number or name.
Cannot-link A, B	A and B occupy different clusters	All other pair relations or the total cluster count.
Seed class label	One observation has a known identity	Every unlabeled observation excludes target.

Worked constrained assignment: the nearest cluster may be forbidden

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

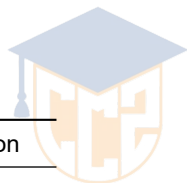
If A must-link B and B must-link C, then A, B and C must belong to one connected must-link component.

Centers are $\mu_1 = 0$ and $\mu_2 = 5$. Point A at $x_A = 0$ is assigned to cluster 1. Point B is at $x_B = 1$ and cannot-link A.

$$\mu_1 = 0 \quad \mu_2 = 5$$

$$x_A = 0 \quad x_B = 1$$

Candidate cluster for B	Squared distance	Constraint satisfied?	Decision
1	$(1 - 0)^2 = 1$	No	Forbidden
2	$(1 - 5)^2 = 16$	Yes	Assign B here



Constraints need a consistency check before an evaluation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Evaluate gains at the same verified-label budget. Check the constraints before interpreting a failed assignment.

Given constraints	Supported conclusion
A must-link B; B must-link C; A cannot-link C	Contradiction: must-link is transitive.
A cannot-link B; B cannot-link C	A and C may still share a group.
Three pairwise-excluding components	At least three groups are required.
One greedy assignment fails	Global infeasibility has not been proved.

Validation and test true labels remain excluded from pseudo-label training targets.

Match the method to the structural evidence

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Match the method to its structural assumption, then check the corresponding failure mode.

Method	Required evidence	Failure to examine
Self-training	Reliable high-confidence predictions	Confident errors
Co-training and consistency	Complementary views or valid transformations	Shared errors
Generative method	Suitable class-conditional distributions	Density-label mismatch
S3VM	Class boundaries in low-density regions	Single-class collapse
Graph propagation	Label-relevant neighborhoods	Error spread through wrong edges
Constrained clustering	Reliable same/different-group constraints	Contradiction or infeasibility

Keep a supervised baseline and verified-label validation. Without credible structure, improve the representation or label informative cases first.

Worked review: design a safe experiment for rare faults

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

We have 40 verified training labels and 5000 unlabeled rows. Keep separate verified-label validation and test sets for rare-fault evaluation.

Step	Experiment
Baseline	Fit 40 verified labels; measure rare-fault recall
Assumption	Relevant similar waveforms tend to share labels
First step	$\tau = .95$: $(.98, .02) \rightarrow 0$; $(.6, .4)$ stays unlabeled
Validation	Check acceptance by class, calibration and recall
Stop criterion	Stop for worse rare-fault recall or invalid source assumptions



Check. Higher training confidence does not verify a target. Freeze choices before the final labeled test.