

Linear Models and Model Evaluation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Pattern Recognition and Machine Learning



Lecture 08 Linear Models and Model Evaluation

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 regression	45
Regression evaluation	22.5
Logistic regression	22.5
Classification evaluation	22.5
Class imbalance	22.5

Three records, then a binary target

Fit a line; interpret its residuals; fit probabilities; choose decisions.

The numerical steps are shown on the slides.



The same inputs can support two different prediction tasks

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Three records define three residuals for one line.

Both models use $\mathbf{z} = \mathbf{b} + \mathbf{w}^\top \mathbf{x}$. Linear regression reports $\mathbf{z}$ directly; logistic regression reports $\sigma(\mathbf{z})$. The target type determines how the score is interpreted and evaluated.

$$y \in \mathbb{R} \quad y \in \{0, 1\}$$

$$p(y = 1 \mid \mathbf{x}) \quad \mathbf{z} = \mathbf{b} + \mathbf{w}^\top \mathbf{x}$$



A design matrix puts all sample equations together

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Rows are observations; columns are the constant and the measured feature.

$$x = (1, 2, 3) \quad y = (2, 3, 5)$$

$$\hat{y}_i = b + wx_i$$

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \beta = \begin{bmatrix} b \\ w \end{bmatrix}, \quad y = \begin{bmatrix} 2 \\ 3 \\ 5 \end{bmatrix}, \quad \hat{y} = X\beta.$$



Residuals define the least-squares objective

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

For the candidate line $\hat{y} = 1 + x$, predictions are (2, 3, 4); residuals are (0, 0, 1), so RSS is 1.

$$r_i = y_i - \hat{y}_i$$

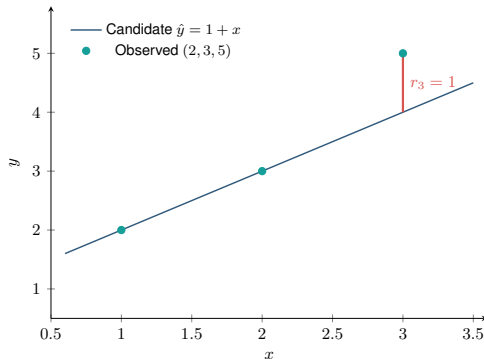
$$\text{RSS}(\beta) = \sum_i r_i^2 = \|y - X\beta\|_2^2.$$

$$\hat{y} = 1 + x$$



Measure a residual vertically at the same input

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



A residual compares observed and predicted responses at the same input. Its vertical difference need not equal the shortest distance to the line.



Differentiate the objective to obtain normal equations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The objective is a convex quadratic, so a stationary solution is globally minimizing.

$$J(\beta) = (\mathbf{X}\beta - \mathbf{y})^\top (\mathbf{X}\beta - \mathbf{y}) = \beta^\top \mathbf{X}^\top \mathbf{X} \beta - 2\beta^\top \mathbf{X}^\top \mathbf{y} + \mathbf{y}^\top \mathbf{y}.$$

$$\nabla J(\beta) = 2\mathbf{X}^\top (\mathbf{X}\beta - \mathbf{y}) = 0 \quad \Rightarrow \quad \mathbf{X}^\top \mathbf{X} \hat{\beta} = \mathbf{X}^\top \mathbf{y}.$$



Each matrix entry is a sum over the same three records

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

These entries give two scalar equations. No matrix inverse is needed to solve this example.

$$x = (1, 2, 3), \quad y = (2, 3, 5), \quad X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix}, \quad \beta = \begin{bmatrix} b \\ w \end{bmatrix}.$$

$$X^T X = \begin{bmatrix} 1+1+1 & 1+2+3 \\ 1+2+3 & 1^2+2^2+3^2 \end{bmatrix} = \begin{bmatrix} 3 & 6 \\ 6 & 14 \end{bmatrix},$$

$$X^T y = \begin{bmatrix} 2+3+5 \\ 1(2)+2(3)+3(5) \end{bmatrix} = \begin{bmatrix} 10 \\ 23 \end{bmatrix}.$$



Elimination solves the normal equations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

At $x = 1$, $\hat{y} = \frac{1}{3} + \frac{3}{2} = \frac{11}{6}$ and $y - \hat{y} = 2 - \frac{11}{6} = \frac{1}{6}$. The next table checks all three rows.

$$3b + 6w = 10 \quad (1), \quad 6b + 14w = 23 \quad (2).$$

$$(2) - 2(1): \quad (6b - 6b) + (14w - 12w) = 23 - 20 \Rightarrow 2w = 3 \Rightarrow w = \frac{3}{2}.$$

$$3b + 6\left(\frac{3}{2}\right) = 10 \Rightarrow 3b = 1 \Rightarrow b = \frac{1}{3}.$$

$$x = 1 \quad \hat{y} = \frac{1}{3} + \frac{3}{2} = \frac{11}{6}$$

$$y - \hat{y} = 2 - \frac{11}{6} = \frac{1}{6}$$



Solve the running example and check its residuals

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Use $\hat{y} = \frac{1}{3} + \frac{3}{2}x$ on the same three observations. Define residual $r = y - \hat{y}$. A negative residual means the prediction is too high. Square and add the three residuals to obtain RSS.

x	True y	Predicted $\hat{y}$	Residual r
1	2	11/6	1/6
2	3	10/3	-1/3
3	5	29/6	1/6



Worked calculation

$$\text{RSS} = \left(\frac{1}{6}\right)^2 + \left(-\frac{1}{3}\right)^2 + \left(\frac{1}{6}\right)^2 = \frac{1}{36} + \frac{1}{9} + \frac{1}{36} = \frac{1}{6}.$$

The second residual is negative, so its prediction is too high. The new line has RSS 1/6, below that of the candidate line $1 + x$ RSS is 1.

At the optimum, residuals are orthogonal to fitted columns

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The fitted response is the orthogonal projection of y onto the column space of X .

$$\begin{aligned} X^T r &= 0. \\ \sum_i r_i &= 0, \quad \sum_i x_i r_i = 1/6 - 2/3 + 1/2 = 0. \end{aligned}$$



Rank deficiency and small singular values need different care

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Use a numerical least-squares solver rather than explicit inversion. Remove redundant features or compare regularized fits when coefficient stability matters.

$$X = U\Sigma V^T \quad \hat{\beta} = X^+y = V\Sigma^+U^Ty$$

Design	What is identifiable?	What needs care?
Two identical columns: x, x	Only $(w_1 + w_2)x$ determines the prediction	$(w_1 + c, w_2 - c)$ gives the same fitted values.
A very small nonzero singular value	A solution may be unique	Dividing by a small value amplifies perturbations.
A numerical least-squares solver	Uses the available directions	Check coefficient stability and compare regularization.

Gaussian noise turns maximum likelihood into least squares

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For fixed σ^2 , only the squared-residual term changes with β . Maximizing likelihood therefore minimizes RSS.

$$y_i = \mathbf{x}_i^\top \beta + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2).$$
$$\log p(y \mid X, \beta, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_i (y_i - \mathbf{x}_i^\top \beta)^2.$$



An outlier changes a squared-loss fit more than a small miss

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

With $\delta = 1$, residuals 1 and 10 have Huber losses 0.5 and 9.5, compared with squared losses 0.5 and 50.

$$L_2(r) = \frac{1}{2}r^2, \quad L_1(r) = |r|, \quad L_\delta(r) = \begin{cases} \frac{1}{2}r^2, & |r| \leq \delta, \\ \delta|r| - \frac{1}{2}\delta^2, & |r| > \delta. \end{cases}$$

$$\delta = 1$$



Ridge adds a positive term to the weak directions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Keep the intercept separate; use centered features Z and response y_c .

$$J(w) = \|Zw - y_c\|^2 + \lambda \|w\|^2, \quad \hat{w} = (Z^T Z + \lambda I)^{-1} Z^T y_c.$$

$$Z = (-1, 0, 1)^T \quad Z^T Z = 2$$

$$Z^T y_c = 3$$

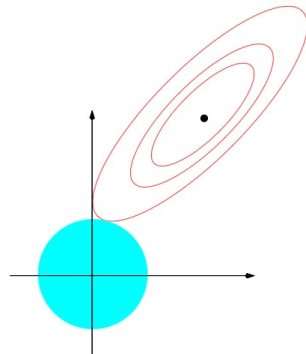
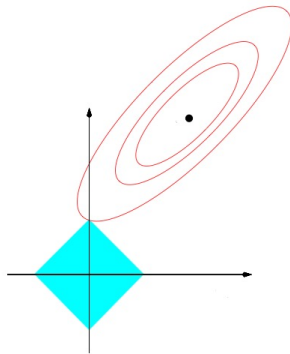
$$\lambda = 1: \quad \hat{w} = 3/(2 + 1) = 1, \quad \hat{b} = \bar{y} - \hat{w}\bar{x} = 4/3.$$



Compare the L1 corners with the L2 circle

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The contours show equal fitting loss.
The penalty restricts which coefficients are affordable. An L1 corner can select a zero coefficient; L2 usually shrinks coefficients continuously.



Lasso uses a thresholded coordinate update

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Hold the other coefficients fixed. The partial residual removes their predictions from the centered response. Soft-thresholding then updates one coefficient; the denominator accounts for its feature scale.

$$r^{(-j)} = y_c - \sum_{k \neq j} \mathbf{z}_k w_k$$

$$J = \frac{1}{2} \|\mathbf{Z}w - y_c\|^2 + \lambda \sum_j |w_j|,$$

$$w_j \leftarrow \frac{S(\mathbf{z}_j^\top r^{(-j)}, \lambda)}{\mathbf{z}_j^\top \mathbf{z}_j}, \quad S(a, \lambda) = \text{sign}(a) \max(|a| - \lambda, 0).$$

$$\lambda = 1 \quad w = 1$$

$$\lambda = 4 \quad w = 0$$

$\mathbf{Z}^\top \mathbf{Z} = 2$ and $\mathbf{Z}^\top y_c = 3$, so $\hat{w} = S(3, \lambda)/2$. With $\lambda = 1$, the result is $(3 - 1)/2 = 1$. With $\lambda = 4$, it is zero.



One gradient step shows what the residual signal does

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For half mean squared error $J = \|X\beta - y\|^2 / (2n)$, the gradient averages feature-weighted residuals. Here $n = 3$, the initial coefficients are zero and the learning rate is .1.

$$\mathbf{g} = \frac{1}{3} X^T (X\beta - y) = \frac{1}{3} \begin{bmatrix} -10 \\ -23 \end{bmatrix} = \begin{bmatrix} -10/3 \\ -23/3 \end{bmatrix}.$$

$$\beta_{\text{new}} = \beta - \eta \mathbf{g} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.1 \begin{bmatrix} -10/3 \\ -23/3 \end{bmatrix} = \begin{bmatrix} 1/3 \\ 23/30 \end{bmatrix}.$$



Check. Mean squared error without the factor $1/2$ doubles this gradient. Keep the loss convention with its learning rate. One step generally does not reach the optimum.

Scaling and step size determine optimization behavior

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For a quadratic least-squares objective, the Hessian is $H = X^T X/n$. Full-batch gradient descent is stable with a step satisfying $0 < \eta < 2/\lambda_{\max}(H)$.

$$H = X^T X/n \quad 0 < \eta < 2/\lambda_{\max}(H)$$

One-dimensional check

 $\eta = .5$ $\eta = 3$

$$L(\theta) = \theta^2/2, H = 1$$

$$\text{Start at } \theta_0 = 1$$

$$\text{Update } \theta_{t+1} = (1 - \eta)\theta_t$$

Within $0 < \eta < 2$

$$1 \rightarrow .5 \rightarrow .25$$

Distance to zero shrinks

Outside the stable range

$$1 \rightarrow -2 \rightarrow 4$$

Magnitude grows while
signs alternate

Residual structure reveals what the model has left unexplained

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A useful residual check asks what structure remains: a curve suggests a missed nonlinear effect, a widening band suggests changing noise, and time dependence suggests related errors.

Observed pattern	Possible explanation to check
Mean residual curves with the input	The linear form may miss a curved relationship
Residual spread grows with the input	Noise scale may depend on the input
Residuals retain one sign over consecutive periods	Possible time dependence, drift or omitted variables



Check. Check splits, measurements and assumptions, then compare repairs within training and validation. A small aggregate error can conceal structured residuals.

Start the metrics from three visible residuals

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Compute several metrics from the same residual vector before comparing models.

The middle point is overpredicted because $r < 0$. Signed residuals sum to zero here, but the absolute and squared errors do not.

$$\hat{y} = 1/3 + 1.5x \quad r = y - \hat{y}$$

x	y	$\hat{y}$	r	$ r $	r^2
1	2	11/6	1/6	1/6	1/36
2	3	10/3	-1/3	1/3	1/9
3	5	29/6	1/6	1/6	1/36



A negative residual means the prediction is too high

Absolute and squared error use different prices

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

MAE and RMSE use the response unit. MSE uses its square. Taking the square root restores units but does not remove the larger price assigned to large errors.

$$\text{MAE} = \frac{1/6 + 1/3 + 1/6}{3} = \frac{2}{9} \approx .222,$$

$$\text{RSS} = \frac{1}{36} + \frac{1}{9} + \frac{1}{36} = \frac{1}{6},$$

$$\text{MSE} = \frac{1/6}{3} = \frac{1}{18} \approx .0556,$$

$$\text{RMSE} = \sqrt{1/18} \approx .236.$$



R-squared compares the fit with a constant baseline

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

This is training-set arithmetic on three rows. A high training R^2 does not establish future accuracy.

$$y = (2, 3, 5) \quad \bar{y} = 10/3$$

$$\begin{aligned} \text{TSS} &= (2 - 10/3)^2 + (3 - 10/3)^2 + (5 - 10/3)^2 \\ &= 16/9 + 1/9 + 25/9 = 14/3. \end{aligned}$$

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}} = 1 - \frac{1/6}{14/3} = 1 - \frac{1}{28} = \frac{27}{28}.$$



Metric choice can reverse a comparison

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

MAE prefers A; RMSE prefers B because A has one large miss. Choose the metric from consequences of errors, then inspect subgroups and signed bias.

$$\text{MAE}_A = (0 + 0 + 6)/3 = 2,$$

$$\text{MAE}_B = 2.5,$$

$$\text{RMSE}_A = \sqrt{(0 + 0 + 36)/3} \approx 3.46,$$

$$\text{RMSE}_B = \sqrt{3(2.5^2)/3} = 2.5.$$



A negative test R-squared has a direct numerical meaning

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The model's squared error is ten times this constant baseline's error. This negative value is meaningful, not a software failure.

$$y = (0, 1)$$

$$\text{RSS} = (0 - 2)^2 + (1 - 2)^2 = 5,$$

$$\text{TSS} = (0 - .5)^2 + (1 - .5)^2 = .5,$$

$$R^2 = 1 - 5/.5 = -9.$$



Condition to check

The evaluation mean defines this descriptive benchmark. It is not permission to use future outcomes when training a deployable baseline.

Worked interpretation: a line, its errors and its limits

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Extrapolation and future accuracy require additional support. Evaluate on suitable held-out observations.

$$\hat{y} = 1/3 + 1.5x \quad x = 1, 2, 3$$

Check	Completed result
at $x = 4$ Predicted	$1/3 + 6 = 19/3$, outside the observed response range
Largest overprediction	at $x = 2$: $y = 3$, $\hat{y} = 10/3$, $r = -1/3$
Copy the input column	Any two slopes summing to 1.5 give the same predictions; coefficients are nonunique
Price one large miss	RMSE emphasizes it more strongly than MAE



A sigmoid converts a linear score into an event probability

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The target now indicates whether an event occurs.

The score is unrestricted; the probability is bounded. A threshold then converts probability into a class.

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

$$z = b + wx, \quad p(y = 1 \mid x) = \sigma(z) = \frac{1}{1 + e^{-z}}.$$

z	-2	-1	0	1	2
p	0.119	0.269	0.500	0.731	0.881

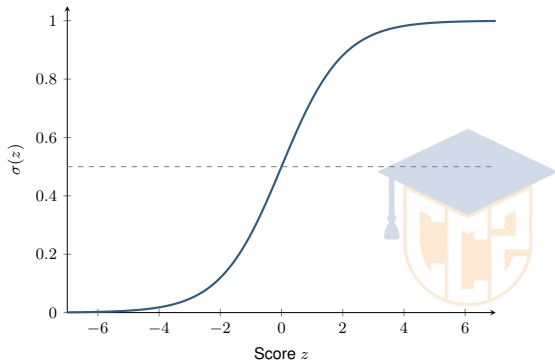
$$\sigma'(z) = p(1 - p) \quad p = 0.5$$



Read probability from a score on the sigmoid curve

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The sigmoid maps score zero to probability .5. Its slope is greatest near zero and approaches zero at both extremes. A probability threshold corresponds to a score threshold.



Log odds are linear; the probability is not

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Increasing x_j by one while holding other features fixed multiplies the odds by e^{w_j} . If $w_j = \log 2$, odds double.

Probability	Original odds	New odds	New probability
1/2	1	2	$2/(1+2) = 2/3$
1/5	1/4	1/2	$(1/2)/(1+1/2) = 1/3$

Check. Doubling the odds does not double the probability. A probability threshold of t Equivalent score threshold $\log[t/(1-t)]$.



Bernoulli likelihood explains cross entropy and confident errors

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

Natural logarithms are used. This smooth objective rewards probability assigned to the observed label. Zero-one error counts wrong classes equally; cross entropy penalizes confident mistakes more strongly.

$$y_i \in \{0, 1\} \quad p_i = \sigma(\mathbf{b} + \mathbf{w}^\top \mathbf{x}_i)$$

$$p(y_i | x_i) = p_i^{y_i} (1 - p_i)^{1 - y_i}.$$

$$J = -\frac{1}{n} \sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)].$$

True label	Predicted positive probability	True-label probability	Loss
1	0.9	0.9	0.105
1	0.01	0.01	4.605
0	0.9	0.1	2.303

The chain rule reduces the gradient to probability minus label

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

One example: differentiate in order

$$s = w^T x + b, \quad p = \sigma(s), \quad \ell = -y \log p - (1-y) \log(1-p).$$

$$\frac{\partial \ell}{\partial p} = -\frac{y}{p} + \frac{1-y}{1-p} = \frac{p-y}{p(1-p)}.$$

$$\frac{\partial p}{\partial s} = p(1-p) \Rightarrow \frac{\partial \ell}{\partial s} = p - y.$$

$$\nabla_w \ell = (p - y)x, \quad \partial_b \ell = p - y.$$

Check the direction

If $y = 1$, $p = .2$, and scalar $x = 2$, then

$$p - y = -.8, \quad \partial_w \ell = -1.6.$$

Gradient descent increases w and raises this example's score.

Average over training examples

$$\nabla_w J = \frac{1}{n} \sum_i (p_i - y_i) x_i,$$

$$\partial_b J = \frac{1}{n} \sum_i (p_i - y_i).$$

Compute one logistic gradient update on three observations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The two negatives move below 0.5, the positive moves above it; average cross entropy falls from 0.693 to about 0.567. This is one step, not the final optimum.

$$\mathbf{x} = (-1, 0, 1) \quad \mathbf{y} = (0, 0, 1)$$

$$\mathbf{b} = \mathbf{w} = 0 \quad p_i = 0.5$$

$$\mathbf{g}_b = (0.5 + 0.5 - 0.5)/3 = 1/6,$$

$$\mathbf{g}_w = [(-1)(0.5) + 0(0.5) + 1(-0.5)]/3 = -1/3.$$

$$\eta = 1 \quad \mathbf{b} = -1/6, \mathbf{w} = 1/3$$



Decision costs determine a probability threshold

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Let false-positive cost be C_{FP} and false-negative cost be C_{FN} , with zero cost for correct decisions.

$$R(\text{predict } 1 \mid x) = C_{FP}(1 - p), \quad R(\text{predict } 0 \mid x) = C_{FN}p.$$

$$p > \frac{C_{FP}}{C_{FP} + C_{FN}}.$$

$$1/(1 + 4) = 0.2$$



Read precision, recall, specificity and F1

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The same probabilities can produce different alarms at different thresholds.

On 1,000 independent held-out inspection records, 50 truly need follow-up. Positive means "needs follow-up"; rows are true labels.

	Predict positive	Predict negative
True positive	TP = 40	FN = 10
True negative	FP = 95	TN = 855

accuracy = $(40 + 855)/1000 = 0.895$ All-negative accuracy = $950/1000 = 0.95$

Recall conditions on true positives; precision conditions on predicted alerts. Balanced accuracy averages recall and specificity: $(0.8 + 0.9)/2 = 0.85$. None of these values alone specifies the cost of mistakes.

Question	Formula	Running value
Are alerts credible?	precision = $TP/(TP + FP)$	$40/135 = 0.296$
How many positives found?	recall = $TP/(TP + FN)$	$40/50 = 0.8$
How many negatives cleared?	specificity = $TN/(TN + FP)$	$855/950 = 0.9$
Balance precision and recall	$F_1 = 2TP/(2TP + FP + FN)$	$80/185 = 0.432$

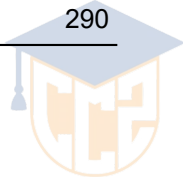
Balanced accuracy = $(0.8 + 0.9)/2 = 0.85$

Moving the threshold changes the counts, not the fitted scores

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Lower thresholds find more positives while creating more false alarms. Here the lowest listed cost is at 0.2, but only three candidate thresholds were evaluated. Select an operating threshold on validation data, consider follow-up capacity, and then freeze it before independent test reporting.

Threshold	TP	FP	FN	TN	Cost $20FN + FP$
0.8	25	19	25	931	519
0.5	40	95	10	855	295
0.2	45	190	5	760	290

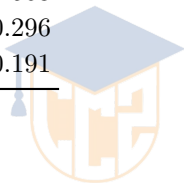


ROC and PR plot different views of the same threshold sweep

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

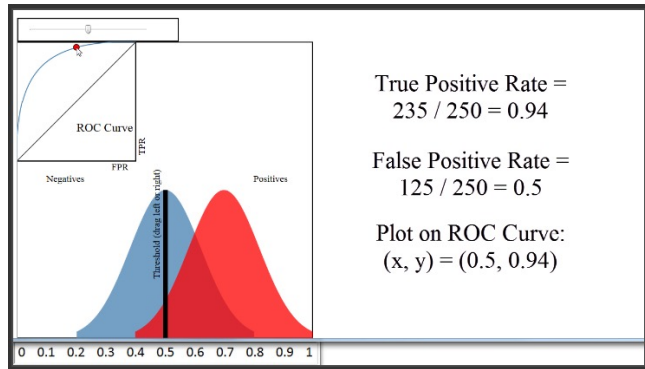
ROC uses FPR horizontally and recall vertically: discrimination relative to negatives.

Threshold	False-positive rate = $FP/950$	recall = $TP/50$	precision
0.8	0.02	0.50	$25/44 = 0.568$
0.5	0.10	0.80	$40/135 = 0.296$
0.2	0.20	0.90	$45/235 = 0.191$

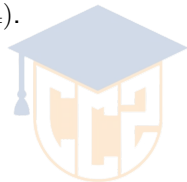


One threshold produces one point on the ROC curve

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



In this reference example, the threshold detects 235 of 250 positives and flags 125 of 250 negatives. The ROC point is therefore $(\text{FPR}, \text{TPR}) = (.50, .94)$.



AUC and average precision summarize ranking in different ways

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Sort ten distinct scores from highest to lowest. The three positives occur at ranks 1, 3 and 6. Count the negatives below each positive for AUC, and precision at those ranks for average precision.

Positive rank	Cumulative TP	Precision	Negatives below
1	1	1/1	7
3	2	2/3	6
6	3	3/6	4



$$\text{AUC} = \frac{7 + 6 + 4}{3 \times 7} = \frac{17}{21} \approx 0.810, \quad \text{AP} = \frac{1 + 2/3 + 3/6}{3} = \frac{13}{18} \approx 0.722.$$

AUC counts correctly ordered positive-negative pairs. AP averages precision at each positive rank. The no-skill precision baseline is $3/10 = 0.3$.

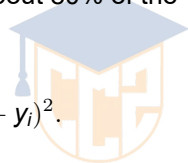
AUC counts correctly ordered positive-negative pairs. AP averages precision at the ranks occupied by positives. Both use the same displayed ranking.

Evaluate probabilities as well as thresholded labels

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Calibration asks whether objects assigned about 0.8 risk experience the event about 80% of the time.

$$\text{LogLoss} = -\frac{1}{n} \sum_i [y_i \log p_i + (1 - y_i) \log(1 - p_i)], \quad \text{Brier} = \frac{1}{n} \sum_i (p_i - y_i)^2.$$



Macro and micro averages emphasize different classes

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Class recalls are 0.90, 0.60, 0.50. Macro recall = $(0.9 + 0.6 + 0.5)/3 = 0.667$; micro recall = $107/130 = 0.823$ for this single-label task.

	A	B	C
A	90	8	2
B	3	12	5
C	1	4	5

$$\text{macroRecall} = (0.9 + 0.6 + 0.5)/3 = 0.667$$

$$\text{microRecall} = 107/130 = 0.823$$



Class weighting changes the importance of observations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A high accuracy can coexist with poor detection of rare positives.

Weighting encourages attention to minority errors without creating new observations.

$$\alpha_1 = \frac{1000}{2 \cdot 50} = 10, \quad \alpha_0 = \frac{1000}{2 \cdot 950} = \frac{10}{19}.$$
$$J = -\frac{1}{n} \sum_i \alpha_{y_i} [y_i \log p_i + (1 - y_i) \log(1 - p_i)] + \text{penalty term}.$$



Resampling changes training data, with specific tradeoffs

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For $x_i = (2, 4)$, $x_j = (6, 8)$, $t = 0.25$, the synthetic input is $(3, 5)$. It is not an observed person. Apply resampling only inside training folds; validation and test retain the intended natural distribution.

$$x_{\text{new}} = (1 - t)x_i + tx_j, \quad 0 < t < 1.$$

$$x_i = (2, 4), x_j = (6, 8), t = 0.25$$

$$x_{\text{new}} = 0.75(2, 4) + 0.25(6, 8) = (1.5 + 1.5, 3 + 2) = (3, 5).$$

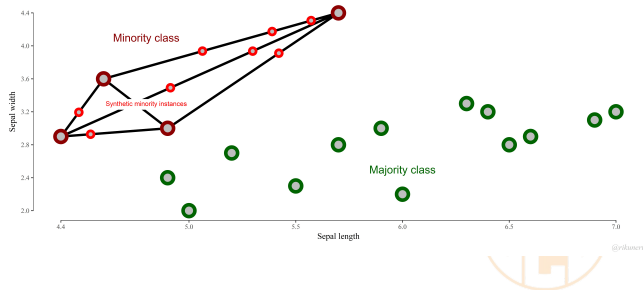


SMOTE creates a point between minority neighbors

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

SMOTE interpolates between minority-class training observations. It creates a point on their connecting segment, so its usefulness depends on whether that segment represents a plausible minority case.

Addressing class imbalance problems of ML via SMOTE: synthesising new dots between existing dots



Prevalence changes precision even when error rates stay fixed

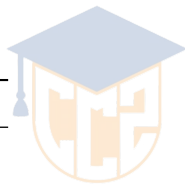
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

With fixed true-positive and false-positive rates, precision still changes with prevalence. Count true and false alarms under the new class proportions before interpreting an alarm.

$$\pi = P(Y = 1)$$

$$\text{Precision} = \frac{\text{TPR} \pi}{\text{TPR} \pi + \text{FPR}(1 - \pi)}.$$

Population	Positive prevalence	precision
Artificially balanced	0.50	$0.4 / (0.4 + 0.05) = 0.889$
Natural rare-positive	0.05	$0.04 / (0.04 + 0.095) = 0.296$



Worked decision: cost subject to a 150-alert limit

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Preserve class proportions when independent random splitting is appropriate. It cannot create more minority information or resolve patient/time dependence. Few positive test events imply substantial uncertainty.

Define the metric before comparison. A model can be good at one use and poor at another; state what the selected evidence supports.

A missed positive costs 20, a false alert costs 1, and at most 150 alerts can be reviewed.

t	Misses	False alerts	Cost	Alerts / feasible?
.8	25	19	$25(20) + 19 = 519$	44: yes
.5	10	95	$10(20) + 95 = 295$	135: yes
.2	5	190	$5(20) + 190 = 290$	235: no

Select among the feasible displayed settings $t = 0.5$: 135 inspections and total cost 295.



Which threshold would we actually use?

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Question

The model has an AUC, false-alarm cost 1, missed-positive cost 20 and capacity for 150 inspections. Does AUC choose the operating threshold?

Evidence and reasoning

No. In the displayed candidate table, threshold .5 uses 135 inspections and has total cost 295. It is the least-cost feasible displayed choice. AUC evaluates ranking across thresholds; it does not encode this capacity constraint.

Check. The choice depends on estimated counts, costs and the available candidate settings. Recheck it when prevalence or capacity changes.