

Bayesian Decisions and Generative Classifiers

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Pattern Recognition and Machine Learning

Lecture 03 Bayesian Decisions and Generative Classifiers

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
Bayesian decisions	45
Probability estimation	45
Gaussian classifiers	45

An alarm in a factory

Count possible sources; update class probabilities; compare action costs.

The numerical steps are shown on the slides.



Count all four outcomes before taking a ratio

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

An alarm is known. Whether the part is defective is still unknown.

Factory model. In 10,000 parts, 2% are defective. The sensor alarms on 90% of defective parts and 5% of sound parts.

$$Y \in \{D, S\} \quad X \in \{A, \neg A\}$$

Interpretation. The same numerator appears in two different fractions. The second denominator includes every explanation of an alarm, including false alarms.

True state	Alarm A	No alarm $\neg A$	Sum
Defective D	$200 \times .90 = 180$	20	200
Normal S	$9,800 \times .05 = 490$	9,310	9,800
Sum	670	9,330	10,000

$$P(A | D) = \frac{180}{200} = .90, \quad P(D | A) = \frac{180}{670} \approx .269.$$



The alarm population supplies the denominator

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The denominator contains every route compatible with the observation.

$$10,000(0.02)(0.90) = 180 \quad \text{defective alarms ,}$$

$$10,000(0.98)(0.05) = 490 \quad \text{sound alarms .}$$

$$\# \text{ all alarms} = 180 + 490 = 670, \quad P(D | A) = 180/670 \approx .269.$$

$$P(A | D) = 180/200 = .90$$



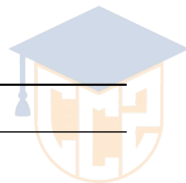
Bayes keeps every competing explanation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

Multiply within one explanation; add across alternative explanations; normalize to obtain class probabilities.

$$P(Y = k | x) = \frac{p(x | Y = k)\pi_k}{\sum_j p(x | Y = j)\pi_j}, \quad \pi_k = P(Y = k).$$

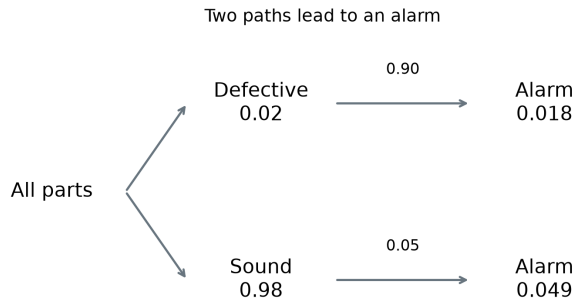


Term	Meaning in the factory
Priors $\pi_D = .02$	Fraction defective before reading the sensor
Likelihood $P(A D) = .90$	Alarm frequency among defective parts
Evidence $P(A) = .018 + .049$	Alarm frequency from all possible states
Posterior $P(D A) = .018/.067$	Defect fraction after restricting to alarms

Two routes lead to the same alarm

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Multiply probabilities along each class path. Add the path weights that lead to the same observed alarm, then normalize to obtain the posterior.



Evidence multiplies the prior odds

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Two observations. An alarm is 18 times more likely under a defect, so it is useful evidence. Sound parts were 49 times more common, so this evidence does not make a defect the most likely state.

$$\underbrace{\frac{P(D | A)}{P(S | A)}}_{\text{posterior odds}} = \underbrace{\frac{P(A | D)}{P(A | S)}}_{\text{likelihood ratio}} \underbrace{\frac{P(D)}{P(S)}}_{\text{prior odds}}.$$

$$18 \times \frac{1}{49} = \frac{18}{49}, \quad P(D | A) = \frac{18}{18 + 49}.$$



A green light and a new population give new posteriors

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

What changed? The first comparison changes the observation. The second changes the class prior while keeping the sensor likelihoods fixed.

Situation	Computation	Defect posterior
No alarm; defect prior 2 %	$20/(20 + 9,310)$	.00214
Alarm; defect prior 2 %	$.90(.02)/[.90(.02) + .05(.98)]$	.269
Alarm; defect prior 20 %	$.90(.20)/[.90(.20) + .05(.80)]$	.818



Largest posterior minimizes zero-one error

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Zero-one classification

A correct class label costs 0. Every wrong class label costs 1.

$$R(k | x) = \sum_{j \neq k} P(Y = j | x) = 1 - P(Y = k | x).$$

$$\hat{y} = \arg \max_k P(Y = k | x).$$

Use the same alarm posterior

$$P(D | A) = 18/67, \quad P(S | A) = 49/67.$$

Predict	Conditional error probability
Defective D	$49/67 \approx .731$
Sound S	$18/67 \approx .269$

Under equal error costs, predict S. Next compare actions when missed defects and false alarms have different costs.

Choose an action by its conditional expected loss

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

Read the table. Rows are chosen actions; columns are unknown true states. Each risk averages one row using the same posterior probabilities.

Action a	True D	True S	Risk after alarm
Quarantine Q	0	2	$2(49/67) = 1.463$
Release R	30	0	$30(18/67) = 8.060$

$$R(a | x) = \sum_j L(a, j) P(Y = j | x), \quad a^*(x) = \arg \min_a R(a | x).$$



A threshold follows from the loss table

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Let $p = P(D | x)$. Keep the preceding costs: a missed defect costs $c_M = 30$, and quarantining a sound part costs $c_{FA} = 2$.

Derive the threshold

$$R(Q | x) = 2(1 - p), \quad R(R | x) = 30p.$$

$$2(1 - p) < 30p \Rightarrow 2 < 32p \Rightarrow p > 2/32 = .0625.$$

In general,

$$c_{FA}(1 - p) < c_M p \iff p > \frac{c_{FA}}{c_M + c_{FA}}.$$

Change one cost

c_M	c_{FA}	Threshold
30	2	$2/32 = .0625$
30	10	$10/40 = .25$
90	2	$2/92 \approx .0217$

At the same posterior .269, quarantine is cheaper than release. A higher miss cost lowers the intervention threshold.



Review creates an intermediate action region

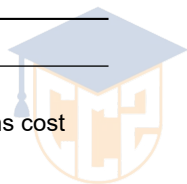
ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

At the alarm posterior $p = .269$, review has the smallest risk. This example treats .5 as the total expected review loss; delays and reviewer errors would need to be included in a richer loss model.

$$R(R | x) = 30p, \quad R(Q | x) = 2(1 - p), \quad R(V | x) = .5.$$

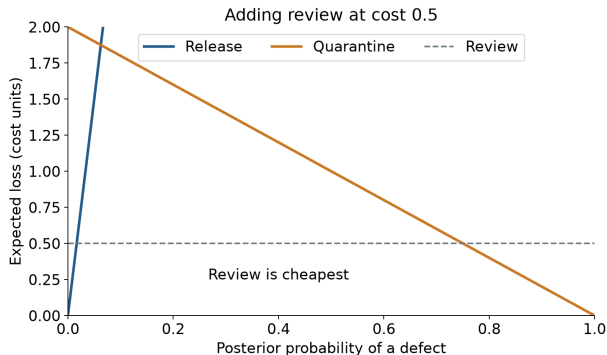
Posterior region	Lowest-risk action	Reason
$p < 1/60$	Release	$30p < .5$
$1/60 < p < .75$	Review	Both automatic actions cost more
$p > .75$	Quarantine	$2(1 - p) < .5$

$p = .269$



Choose the lowest risk curve at each probability

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER



At each posterior probability, compare all action risks vertically. The lowest curve determines the action. A constant review cost of .5 creates an intermediate review region.



Pointwise choices determine the overall risk

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Evaluate the earlier two-action policy: quarantine alarms and release non-alarms. Use the original 10,000 parts and the same costs.

Outcome	Count	Total loss
Defect, alarm	180	0
Sound, alarm	490	$490(2) = 980$
Defect, no alarm	20	$20(30) = 600$
Sound, no alarm	9310	0

$$R = \frac{980 + 600}{10000} = .158 \text{ per part.}$$

Conditional risks combine into overall risk

$$R(a) = E_X[R(a(X) | X)],$$

$$R^* = E_X[\min_a R(a | X)].$$

Each observation contributes its chosen action's loss.

With review loss .5, review all 670 alarms and release non-alarms:

$$R = \frac{670(.5) + 20(30)}{10000} = .0935.$$

Compare the policies with the same probabilities, population and cost units.

More classes: compare weights, then compare actions

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Use posteriors (.48, .36, .16). Each action risk is the probability-weighted sum of its costs.

Action	True A	True B	True C
Choose A	0	4	8
Choose B	2	0	3
Choose C	5	1	0

$$R(\text{Choose A}) = 0(.48) + 4(.36) + 8(.16) = 2.72,$$

$$R(\text{Choose B}) = 2(.48) + 0(.36) + 3(.16) = 1.44,$$

$$R(\text{Choose C}) = 5(.48) + 1(.36) + 0(.16) = 2.76.$$

A is most probable; action B minimizes expected loss.



Worked review: learn rates and choose an action

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

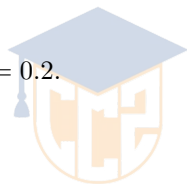
New labelled records. 10 defective parts include 8 alarms. 90 sound parts include 18 alarms. Use raw frequency estimates in this exercise.

$$P(D) = 10/100 = 0.1, \quad P(A | D) = 8/10 = 0.8, \quad P(A | \neg D) = 18/90 = 0.2.$$

$$P(D | A) = \frac{0.1(0.8)}{0.1(0.8) + 0.9(0.2)} = \frac{0.08}{0.26} \approx 0.308.$$

Eight of the 26 alarmed parts are defective, giving 8/26.

Check. .8 is the fraction of defective parts that trigger an alarm. .308 is the fraction of alarmed parts that are defective.



Solution: probability and action are separate steps

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Without review: quarantine.

With review: review. **What**

changed? The available actions changed. The posterior 4/13 did not. In the next period, we study how frequency estimates can become unreliable when the training sample is small.

$$\hat{\pi}_D = .10, \quad \hat{P}(A | D) = .80, \quad \hat{P}(A | S) = .20.$$

$$w_D = .08, \quad w_S = .18, \quad P(D | A) = \frac{.08}{.26} = \frac{4}{13}.$$

Action	Computation	Conditional risk
Release	30(4/13)	120/13 $\approx$ 9.231
Quarantine	2(9/13)	18/13 $\approx$ 1.385
Review	.5	.5

A distribution family still needs parameter values

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Estimate a simple distribution first. Then combine several feature probabilities in Naive Bayes.

Likelihood. Keep these data fixed and compare candidate parameters. $L(\theta)$ is not automatically a normalized density over θ .

$$L(\theta) = p(\mathcal{D} \mid \theta) = \theta^7 (1 - \theta)^3.$$



Derive the Bernoulli maximum-likelihood estimate

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

For independent trials with common success probability θ , the observed sequence has likelihood $L(\theta) = \theta^7(1 - \theta)^3$.

$$L(\theta) = \theta^7(1 - \theta)^3$$

$$\ell = 7 \log \theta + 3 \log(1 - \theta), \quad \ell' = \frac{7}{\theta} - \frac{3}{1 - \theta}.$$

$$\frac{7}{\theta} = \frac{3}{1 - \theta} \Rightarrow 7(1 - \theta) = 3\theta \Rightarrow 7 = 10\theta \Rightarrow \hat{\theta} = .7.$$

For s successes in n independent trials, differentiate the log likelihood. An interior optimum gives $\hat{\theta} = s/n$. All-success or all-failure samples require a boundary optimum.

$$\ell(\theta) = s \log \theta + (n - s) \log(1 - \theta),$$

$$\ell'(\theta) = \frac{s}{\theta} - \frac{n - s}{1 - \theta} = 0 \quad \Rightarrow \quad \hat{\theta}_{ML} = \frac{s}{n}.$$

$$\ell''(\theta) = -\frac{s}{\theta^2} - \frac{n - s}{(1 - \theta)^2} < 0 \quad (0 < s < n).$$

Beta updating records evidence and uncertainty

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Multiply prior and likelihood

$$p(\theta) \propto \theta^{2-1}(1-\theta)^{2-1},$$

$$L(\theta) = \theta^7(1-\theta)^3.$$

$$p(\theta | D) \propto \theta^8(1-\theta)^4.$$

The exponents are parameter minus one, so the posterior is Beta(9, 5).

$$\theta | D \sim \text{Beta}(\alpha + s, \beta + n - s).$$

Three different summaries

Method	Calculation	Value
MLE	7/10	.700
Posterior mean	9/(9 + 5)	.643
Posterior mode	(9 - 1)/(14 - 2)	.667

The mean averages the posterior. The mode locates its peak. The displayed mode formula needs both shape parameters above 1.

$$\text{Beta}(2, 2) \rightarrow \text{Beta}(5, 3) \rightarrow \text{Beta}(9, 5).$$

Prediction and Laplace smoothing average uncertainty

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Add-one smoothing is the posterior-mean prediction for a uniform Beta prior. Do not call this the MAP estimate: under the same uniform prior, MAP equals MLE.

$$P(X_{new} = 1 \mid \mathcal{D}) = \int \theta p(\theta \mid \mathcal{D}) d\theta = \frac{\alpha + s}{\alpha + \beta + n}.$$

$$E[\theta \mid \mathcal{D}] = \frac{n}{n + \alpha + \beta} \frac{s}{n} + \frac{\alpha + \beta}{n + \alpha + \beta} \frac{\alpha}{\alpha + \beta}.$$

Priors	Data	Predictive success probability
Beta (2, 2)	7 of 10	9/14 = .643
Beta (2, 2)	70 of 100	72/104 = .692
Beta (1, 1)	0 of 4	1/6 Not zero.



More outcomes require a normalized probability vector

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The posterior is Dirichlet(4, 2, 1); the estimates sum to 1. A categorical feature with M values adds M counts to its denominator. A word-count model instead uses the total number of tokens in that class.

$$\hat{\theta}_{j,ML} = \frac{n_j}{n}, \quad E[\theta_j | \mathcal{D}] = \frac{n_j + \alpha_j}{n + \sum_v \alpha_v}.$$

Three observed categories	First	Second	Third
Counts	3	1	0
MLE probabilities	3/4	1/4	0
Dirichlet (1, 1, 1) Posterior mean	4/7	2/7	1/7



Bernoulli features: present and absent both count

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Key idea

A joint model of d binary features has $2^d - 1$ free probabilities per class. Conditional independence needs only d . For $d = 20$, that is 1,048,575 versus 20. The simplification replaces interactions with an independence assumption **within each class**.

$$p(x_1, \dots, x_d \mid Y = k) \approx \prod_{j=1}^d p(x_j \mid Y = k).$$

$d = 20$

Representation.

$x = (1, 0)$ means free is present and meeting is absent. Repeating free twice still gives the same binary value 1.

Training messages	Spam	Ordinary
Emails in this class n_k	4	6
Messages containing free	3	1
Messages containing meeting	0	4

$$\phi_{kj} = \frac{n_{kj} + 1}{n_k + 2}, \quad p(x \mid k) = \prod_j \phi_{kj}^{x_j} (1 - \phi_{kj})^{1-x_j}.$$

$$x = (1, 0) \quad \phi = (4/6, 1/6)$$

A complete Bernoulli message prediction

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Read the factors. Prior $\times$ probability of free $\times$ probability of no meeting. Both classes explain the same observation.

$$x = (1, 0)$$

$$w_S = .4 \left(\frac{4}{6} \right) \left(1 - \frac{1}{6} \right) = \frac{2}{9} \approx .2222,$$

$$w_O = .6 \left(\frac{2}{8} \right) \left(1 - \frac{5}{8} \right) = .05625,$$

$$P(S | x) = \frac{w_S}{w_S + w_O} = \frac{320}{401} \approx .798.$$



Categorical features: borrow the account example

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Reference example. Predict whether an account is genuine from posting density, friend density, and a real profile photo. Density takes s, m, l ; photo takes yes/no.

New input (m, m, yes)	Genuine	Not genuine
Priors	7/10	3/10
Posting likelihood	2/5	1/3
Friend likelihood	1/2	1/6
Photo likelihood	5/9	3/5

$$w_G = 7/90, \quad w_N = 1/100, \quad P(G | \mathbf{x}) = \frac{70}{79} \approx .886.$$



Multinomial features count tokens, not messages

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

New message. free free project has count vector $x = (2, 0, 1)$.

Training token counts	free	meeting	project	Sum
Spam	6	0	2	8
Ordinary	1	5	4	10

$$\phi_{kj} = \frac{n_{kj} + 1}{\sum_v n_{kv} + V}, \quad V = 3.$$

$$x = (2, 0, 1)$$

$$w_S = .4(7/11)^2(3/11) = .04418, \quad w_O = .6(2/13)^2(5/13) = .005462.$$



Log scores connect the formula to the program

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Rows are classes; columns are words. Subtracting a common maximum preserves score differences and the normalized posterior. Predict by argmax when probabilities are not needed.

$$s_k = \log \pi_k + \sum_j x_j \log \phi_{kj}, \quad P(k | x) = \frac{e^{s_k - m}}{\sum_j e^{s_j - m}}, \quad m = \max_j s_j.$$

```
import numpy as np
counts = np.array([[6.,0.,2.], [1.,5.,4.]])
phi = (counts+1)/(counts.sum(axis=1, keepdims=True)+3)
x = np.array([2.,0.,1.])
s = np.log([.4,.6]) + np.log(phi) @ x
q = np.exp(s-s.max())
print(q/q.sum()) # approximately [.890, .110]
```

What independence and representation can lose

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Conditional, not marginal.

With equal class priors and independent feature probabilities .9 in A and .1 in B,
 $P(X_1 = 1) = P(X_2 = 1) = .5$, but $P(X_1 = X_2 = 1) = .41 \neq .25$.

Example	Model effect	What was lost?
Copy one clue into two columns	The odds change from 2 : 1 Change to 4 : 1	Dependence is counted twice
“Dog bites man” versus “Man bites dog”	Identical count vectors	Word order and roles
20 binary clues with little training data	Few parameters help estimation	Interactions are omitted

$$P(X_1 = 1) = P(X_2 = 1) = .5 \quad P(X_1 = X_2 = 1) = .41 \neq .25$$

Model assumption

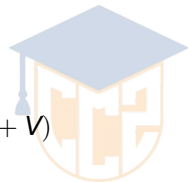
Repeated or dependent clues can make a naive product overconfident.

Choose a distribution from the observation definition

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

There are four records and three possible feature categories with counts $(3, 1, 0)$. Adding one to every count gives denominator $4 + 3 = 7$.

Observation	Model	Smoothed or fitted quantity
Present/absent word	Bernoulli	$(n_{kj} + 1)/(n_k + 2)$
M mutually exclusive categories	Categorical	$(n_{kjm} + 1)/(n_k + M)$
Word counts over V terms	Multinomial	$(n_{kj} + 1)/(\text{class tokens} + V)$
Continuous feature	Gaussian	Class mean and variance



One training table produces priors and feature models

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Split training observations by class; define each feature's sample space.

$$\pi_k = P(Y = k) \quad s_k = \log \pi_k + \log p(x | k)$$

Step	What comes from training?	What happens for a new message?
1. Prior	Class counts: 4 spam, 6 ordinary	Compare both candidate classes.
2. Feature model	Use the sample space: presence, category or token count	Evaluate the same observation under each class.
3. Score	Keep the fitted parameters fixed	$s_k = \log \pi_k + \log p(x k)$.
4. Decide	Normalize scores if a posterior is needed	Choose an action using its loss.

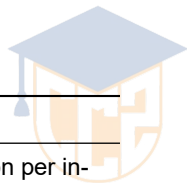
Continuous measurements need densities, not point masses

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Counts work for discrete outcomes, but a measured length can take continuously many values.

Class posteriors compare class-weighted densities at the same input. A density height is not itself a class probability and can exceed 1.

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right], \quad P(a < X < b) = \int_a^b p(x) dx.$$



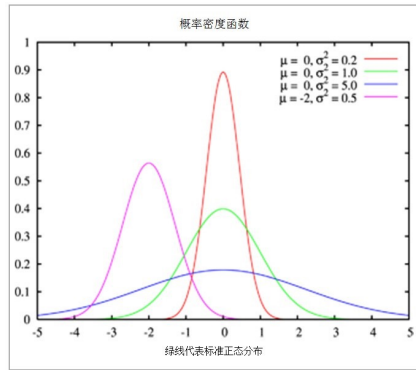
$X \sim \mathcal{N}(0, .01)$	Value	Meaning
Density $p(0)$	3.989	Local concentration per input unit
$P(-.1 < X < .1)$	.683	Area under the density
$P(X = 0)$	0	A single point has zero width

Locate the center and spread of each Gaussian

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The mean locates the Gaussian center.

The variance controls spread. Density height describes concentration, while area over an interval gives probability.



A covariance matrix describes variation and correlation

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Variances describe spread; covariances describe how features vary together.

$$\mu = E[x], \quad \Sigma = E[(x - \mu)(x - \mu)^T], \quad x = (x_1, x_2)^T.$$

Entry	Interpretation	Example
Σ_{11}, Σ_{22}	Variances of the two measurements	4, 1
$\Sigma_{12} = \Sigma_{21}$	Joint deviations around the means	Positive: often increase together
$\rho = \Sigma_{12} / \sqrt{\Sigma_{11} \Sigma_{22}}$	Unit-free correlation	$\Sigma_{12} = 1 \Rightarrow \rho = .5$

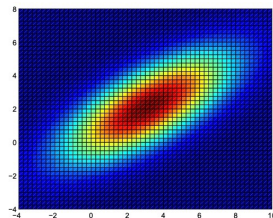
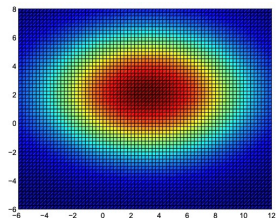
$$\Sigma = \begin{bmatrix} 4 & 1 \\ 1 & 1 \end{bmatrix}$$



Same center, different covariance: read the Gaussian shapes

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

$$\mu = (3, 2)^T, \quad \Sigma_L = \begin{bmatrix} 25 & 0 \\ 0 & 9 \end{bmatrix}, \quad \Sigma_R = \begin{bmatrix} 15 & 5 \\ 5 & 5 \end{bmatrix}.$$



Compare the two models. Both models center at (3, 2). Left: zero covariance and axis-aligned spread. Right: positive covariance and upward tilt.



Mahalanobis distance measures deviation in class units

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Mahalanobis distance measures deviation relative to the modeled spread. The same raw displacement can have different significance along different directions.

$$d_M^2(x, \mu) = (x - \mu)^T \Sigma^{-1} (x - \mu).$$

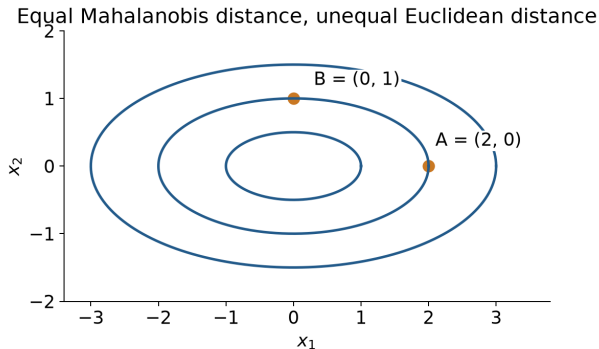
$$\mu = 0 \quad \Sigma = \text{diag}(4, 1)$$

Point	$\ x\ ^2$	d_M^2	Interpretation
(2, 0)	4	$2^2/4 = 1$	Along x_1 1 s.d.
(0, 1)	1	$1^2/1 = 1$	Along x_2 1 s.d.
(0, 2)	4	$2^2/1 = 4$	Along x_2 2 s.d.

Equal relative deviations can have different raw lengths

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

A and B lie on the same density contour despite different raw displacements. Mahalanobis distance measures these displacements relative to the class spread.



The Gaussian score has three interpretable terms

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

$-\log(2\pi)/2$ is common and can be dropped. A class-specific determinant cannot be dropped. Select the largest score under zero-one loss, or normalize and use the loss table.

$$p(\mathbf{x} | k) = \frac{\exp[-\frac{1}{2} \mathbf{d}_{M,k}^2]}{(2\pi)^{d/2} |\Sigma_k|^{1/2}},$$

$$g_k(\mathbf{x}) = -\frac{1}{2} \log |\Sigma_k| - \frac{1}{2} \mathbf{d}_{M,k}^2 + \log \pi_k.$$



Shared covariance cancels the quadratic terms

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

With shared covariance, expand the quadratic form in each class score. The term involving $x^T \Sigma^{-1} x$ is common to every class and cancels in comparisons.

$$A = \Sigma^{-1}$$

$$\begin{aligned}(x - \mu_k)^T A (x - \mu_k) &= x^T A x - x^T A \mu_k - \mu_k^T A x + \mu_k^T A \mu_k \\ &= x^T A x - 2x^T A \mu_k + \mu_k^T A \mu_k.\end{aligned}$$

If every class shares Σ , expand the distance term:

$$g_k(x) = x^T \Sigma^{-1} \mu_k - \frac{1}{2} \mu_k^T \Sigma^{-1} \mu_k + \log \pi_k + \text{common terms}.$$

$$g_A(x) - g_B(x) = w^T x + b, \quad w = \Sigma^{-1}(\mu_A - \mu_B),$$

$$b = -\frac{1}{2} \mu_A^T \Sigma^{-1} \mu_A + \frac{1}{2} \mu_B^T \Sigma^{-1} \mu_B + \log(\pi_A / \pi_B).$$



A prior shifts a linear Gaussian threshold

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Model. $X | A \sim \mathcal{N}(0, 1)$ and $X | B \sim \mathcal{N}(4, 1)$.

$$g_A(x) - g_B(x) = -\frac{1}{2}x^2 + \frac{1}{2}(x-4)^2 + \log \frac{\pi_A}{\pi_B} = -4x + 8 + \log \frac{\pi_A}{\pi_B}.$$

$$x^* = 2 + \frac{1}{4} \log \frac{\pi_A}{\pi_B}; \quad \text{choose A if } x < x^*.$$

Priors	Threshold	Interpretation
(.5, .5)	2	Midpoint between centers
(.8, .2)	2.347	More common A receives a larger region



Unequal variances can produce two thresholds

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Model. $X | A \sim \mathcal{N}(0, 1)$, $X | B \sim \mathcal{N}(0, 4)$; equal priors.

$$g_A(x) - g_B(x) = \log 2 - \frac{3}{8}x^2.$$

$$g_A = g_B \quad \Rightarrow \quad |x| = \sqrt{\frac{8}{3} \log 2} \approx 1.360.$$

$|x| < 1.360$ gives a larger density for A; $|x| > 1.360$ gives a larger density for B; use a stated rule at equality.



Eight points let us train the whole model step by step

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Training quantities. Both classes contain 4 points, so both priors are .5.

Class	Labelled feature vectors	Class mean
A	(0, 0), (2, 0), (2, 2), (0, 2)	$(1, 1)^T$
B	(4, 4), (6, 4), (6, 6), (4, 6)	$(5, 5)^T$

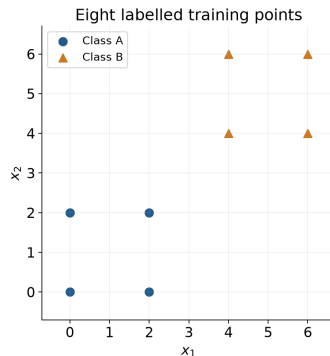
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} \mathbf{x}_i, \quad \hat{\pi}_k = \frac{n_k}{n}.$$



Locate the eight training observations

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Each class mean locates its cloud.
Deviations from that class mean
provide within-class spread. The same
eight points determine all later
parameter estimates.



Covariance is an average of centered outer products

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

MLE divides by $n_k = 4$. The unbiased sample covariance divides by $n_k - 1 = 3$ and gives $\frac{4}{3}I$; they are different estimators. An explicit denominator makes the derivation and code agree.

Centered vector	δ_1^2	$\delta_1\delta_2$	δ_2^2
$(-1, -1)$	1	1	1
$(1, -1)$	1	-1	1
$(1, 1)$	1	1	1
$(-1, 1)$	1	-1	1
Sum	4	0	4

$$\hat{\Sigma}_A = \frac{1}{4} \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} = I, \quad \hat{\Sigma}_B = I.$$



Within-class pooling and grand-mean centering

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Center each class by its own mean before pooling covariance. Centering all points by the grand mean also includes between-class separation.

$$\hat{\Sigma}_{\text{shared}} = \frac{1}{n} \sum_k \sum_{i: y_i=k} (\mathbf{x}_i - \hat{\mu}_k)(\mathbf{x}_i - \hat{\mu}_k)^T = I.$$

Check. Centering all eight points by the grand mean gives $(3, 3)^T$ gives

$$\text{Cov}_{\text{all}}(\mathbf{X}) = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}.$$

This includes between-class spread. The next page uses the within-class estimate to derive the boundary.



The fitted Gaussian scores give a linear boundary

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Insert the fitted class means and shared covariance into the score difference. Equal priors give the boundary $x_1 + x_2 = 6$.

$$\begin{aligned}(x_1 - 1)^2 + (x_2 - 1)^2 &= (x_1 - 5)^2 + (x_2 - 5)^2 \\ x_1^2 + x_2^2 - 2x_1 - 2x_2 + 2 &= x_1^2 + x_2^2 - 10x_1 - 10x_2 + 50 \\ 8x_1 + 8x_2 &= 48 \\ x_1 + x_2 &= 6.\end{aligned}$$



Worked calculation

$x_1 + x_2 < 6$ selects A; a sum above 6 selects B. State a rule at equality. The class-mean sums 2 and 10 check the direction of the inequality.

Scores, posterior, and boundary should agree

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Geometry check. The coordinate sums are 5 and 7, on opposite sides of $x_1 + x_2 = 6$. **Action check.** A cost-sensitive decision still requires a loss table; a confident class score alone does not specify the cheapest action.

$$\mu_A = (1, 1)^T \quad \mu_B = (5, 5)^T$$

$$\Sigma = I$$

New point	d_A^2	d_B^2	$g_A - g_B$
(2, 3)	5	13	4
(4, 3)	13	5	-4

$$P(A | x) = \frac{1}{1 + \exp[-(g_A - g_B)]}.$$

$$x_1 + x_2 = 6$$



Implement the same computation without matrix inversion

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

```
def score(x, mu, Sigma, prior):
    delta = x-mu
    np.linalg.cholesky(Sigma) # requires positive definiteness
    _, logdet = np.linalg.slogdet(Sigma)
    d2 = delta @ np.linalg.solve(Sigma, delta)
    return -.5*(logdet+d2) + np.log(prior)
means = [np.array([1.,1.]), np.array([5.,5.])]
s = np.array([score(np.array([4.,3.]), m, np.eye(2), .5)
              for m in means])
q = np.exp(s-s.max())
print(q/q.sum()) # approximately [.018, .982]
```

Solve $\Sigma v = \delta$, then compute $\delta^T v$. Check class order, covariance definiteness, and normalization. The numerical routine returns exactly the log score we derived, with the shared normalizer omitted.

$$\Sigma v = \delta$$

Small samples can leave covariance directions unlearned

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

The samples vary in only one direction. In general,

$$\hat{\Sigma} = \begin{bmatrix} 2/3 & 4/3 \\ 4/3 & 8/3 \end{bmatrix}, \quad |\hat{\Sigma}| = 0.$$

$$\text{rank}(\hat{\Sigma}_k) \leq \min(d, n_k - 1).$$

$$\tilde{\Sigma} = (1 - \lambda)\mathbf{S} + \lambda\tau I, \quad \tau = \text{tr}(\mathbf{S})/d.$$

$$\lambda = .2$$



Solution and the bridge to discriminative classifiers

ADVANCED COMPUTING AND DIGITAL TECHNOLOGY RESEARCH CENTER

Keep the fitted eight-point means and $\Sigma = I$. In a new population, priors are $(\pi_A, \pi_B) = (.8, .2)$. Observe $x = (3, 3)^T$.

$$\Sigma = I \quad (\pi_A, \pi_B) = (.8, .2)$$

$$x = (3, 3)^T \quad \Sigma = I$$

A has posterior .8, yet unequal error costs favor action B. The following risk comparison explains why the most likely class need not be the least costly action.



$$d_A^2 = d_B^2 = 8, \quad P(A | x) = .8, \quad P(B | x) = .2.$$

Action	Risk	Decision
Choose A	$10(.2) = 2$	Higher loss
Choose B	$1(.8) = .8$	Minimum risk