



Introduction to Data Science

Fundamentals of Probability

17 Sep 2026

Fundamentals of Probability

Today's Route

1. Bayes' Theorem

Move from observed symptoms or test results to an updated belief.

2. Entropy, Cross Entropy, and KL Divergence

Measure uncertainty and the cost of using a mismatched distribution.

3. Continuous Probability Distributions

Replace sums with integrals, using uniform and Gaussian examples.

Update beliefs, compare distributions, and model continuous values.

Bayes' Theorem

Why Reverse a Conditional Probability?

FORWARD QUESTION

If a disease is present, how likely are particular symptoms?

This follows the textbook's disease-to-symptom description.

$P(\text{symptoms} \mid \text{disease})$

INFERENCE QUESTION

Given the symptoms, how likely is the disease?

This is the question a clinician would want to answer.

$P(\text{disease} \mid \text{symptoms})$



The forward question and the inference question use different conditions.

Bayes' Theorem

Joint Probability: Both Values Occur

Let A represent symptoms and B represent disease.

$$P(A = a, B = b)$$

This is the probability that A takes value a and B takes value b together.

$$P(A, B) = P(A \mid B) P(B)$$

$$P(A, B) = P(B \mid A) P(A)$$

The joint distribution describes combinations of values of two random variables.

Bayes' Theorem

The Same Joint Probability, Rearranged

$$P(B \mid A) P(A) = P(A \mid B) P(B)$$

$$P(B \mid A) = \frac{P(A \mid B) P(B)}{P(A)}$$

The numerator is the probability that the disease and the symptoms occur together.

The denominator is the probability of the observed symptoms. Use $P(A) > 0$.

Bayes' theorem calculates the reverse conditional from quantities we already know.

Bayes' Theorem

Marginalization: Where Does $P(A)$ Come From?

$$P(A) = \sum_b P(A, B = b)$$

To obtain the probability of symptoms A , include every possible value of B .

The variable B disappears from the final expression because we sum over its values.

For a test example, the two cases are disease present and disease absent.

These cases must cover the possibilities without overlap.

Sum over the possibilities that could produce the observed evidence.

Bayes' Theorem

A Positive Test

Illustrative counts for 1,000 people—not clinical data.



Group	Positive	Negative	Total
Disease present	8	2	10
Disease absent	99	891	990
Total	107	893	1,000

Question: among the 107 positive results, how many people have the disease?

A positive result can occur with or without the disease.

Bayes' Theorem

A Positive Test: Follow the Denominator

$$\begin{aligned} P(\text{disease} \mid \text{positive}) &= \frac{8}{8 + 99} \\ &= 8/107 \approx 0.0748 = 7.48\% \end{aligned}$$

There are eight positive results from people with the disease.

There are also 99 positive results from people without it.

Both groups must be included in the denominator.

The posterior considers the positive cases, not the whole population.

Bayes' Theorem

Prior, Likelihood, and Posterior in the Example

Quantity	Question	Example value
Prior	Before the test, how common is the disease?	1%
Likelihood	With the disease, how often is the test positive?	80%
Posterior	After a positive test, how likely is the disease?	About 7.48%

The direction of the condition changes which group we are counting within.

Do not confuse $P(\text{positive given disease})$ with $P(\text{disease given positive})$.

Bayes' Theorem

Quick Check: Read the Condition

1. What do we observe?

The test result is positive. Disease status is still unknown.

2. Which group do we use?

Restrict attention to everyone with a positive result.

3. What do we count?

Within that group, count the people with the disease.

Always ask: what information is already given?

Entropy, Cross Entropy, and KL Divergence

Recording Experiment

1. Researcher

Roll a dice and report the outcome over a phone line.



2. Recorder

Write a binary string for each reported outcome.

3. Goal

Recover the results correctly while writing as few bits as possible.

Less uncertainty can mean fewer bits are needed to describe an outcome.

Entropy, Cross Entropy, and KL Divergence

Why the Code Is Ambiguous

Outcome	Code
1	0
2	1
3	10
4	11
5	100
6	101

Recorded string: 0110

$0 + 1 + 1 + 0$ gives (1, 2, 2, 1)

$0 + 1 + 10$ gives (1, 2, 3)

$0 + 11 + 0$ gives (1, 4, 1)



The same record has multiple valid readings.

One binary string can have several readings under the first code.

Entropy, Cross Entropy, and KL Divergence

Prefix Codes Remove the Ambiguity

Outcome	Prefix code	Code length
1	0	1 bit
2	10	2 bits
3	110	3 bits
4	1110	4 bits
5	11110	5 bits
6	111110	6 bits



Recorded string: 0110 0 + 110 gives (1, 3)

Entropy, Cross Entropy, and KL Divergence

Prefix Codes Remove the Ambiguity

Outcome	Prefix code	Code length
1	0	1 bit
2	10	2 bits
3	110	3 bits
4	1110	4 bits
5	11110	5 bits
6	111110	6 bits



Once a complete codeword is found, it cannot be the start of a longer valid word.

No complete codeword may be the beginning of another codeword.

Entropy, Cross Entropy, and KL Divergence

Average Code Length for a Fair Dice

PREFIX CODE

Code lengths: 1, 2, 3, 4, 5, 6.

Each face has probability $1/6$.

$$(1 + 2 + 3 + 4 + 5 + 6)/6$$

$$= 3.5 \text{ bits per roll}$$

FIXED-LENGTH CODE

000, 001, 010, 011, 100, 101.

Every outcome uses three bits.

3 bits per roll



Different valid codes can require different amounts of writing.

Entropy, Cross Entropy, and KL Divergence

A Biased Dice Changes the Best Allocation

Suppose a weighted dice produces a 6 almost all the time.

Giving 6 a short code, such as “0”, reduces the average number of written bits.

Rare outcomes can use longer codes because they are written less often.

If the result is completely known in advance, the outcome carries no new uncertainty.



Assign short descriptions to frequent outcomes.

Entropy, Cross Entropy, and KL Divergence

Information in One Outcome

$$\text{Information length} = \log_2(1/p(x)) = -\log_2 p(x)$$

Probability of the outcome	Information length
1/2	1 bit
1/4	2 bits
1/8	3 bits

This is an ideal information length; a single binary codeword has an integer length.

A less probable outcome carries more information when it occurs.

Entropy, Cross Entropy, and KL Divergence

Entropy Is the Average Information

$$H(p) = -\sum_x p(x) \log_2 p(x)$$

1. Compute the information

For each outcome, use $-\log_2 p(x)$.

2. Weight by how often it occurs

Multiply that information by $p(x)$, then add over all outcomes.

Entropy averages the information over the true outcome probabilities.

Entropy, Cross Entropy, and KL Divergence

Entropy: Certain, Fair Coin, Fair Dice

Distribution	Entropy	Meaning
One outcome has probability 1	0 bits	No uncertainty remains.
Fair coin: two outcomes	1 bit	Two equally likely possibilities.
Fair dice: six outcomes	$\log_2 6 \approx 2.585$ bits	Six equally likely possibilities.

For a fixed finite outcome set, a uniform distribution has the highest entropy.

Entropy, Cross Entropy, and KL Divergence

Cross Entropy: True p , Assumed q

p : WHAT ACTUALLY HAPPENS

The true distribution determines how often each outcome appears.

It supplies the weights in the expectation.

q : WHAT WE ASSUME OR PREDICT

The assumed distribution determines the information length assigned to each outcome.

It appears inside the logarithm.

$$CE(p, q) = -\sum_x p(x) \log_2 q(x)$$

Outcomes follow p , but their descriptions are designed using q .

Entropy, Cross Entropy, and KL Divergence

A Small Cross-Entropy Calculation

Outcome	True p	Assumed q
Heads	0.5	0.75
Tails	0.5	0.25

$$CE(p, q) = -0.5 \log_2(0.75) - 0.5 \log_2(0.25)$$

$$\approx 1.2075 \text{ bits}, \quad H(p) = 1 \text{ bit}$$



The extra average information length is about 0.2075 bits.

A mismatched distribution can require more average bits than the true distribution.

Entropy, Cross Entropy, and KL Divergence

KL Divergence: The Extra Cost of a Mismatch

$$\begin{aligned} \text{KL}(p, q) &= \text{CE}(p, q) - H(p) \\ &= \sum_x p(x) \log_2(p(x) / q(x)) \end{aligned}$$

In the coin example: $\text{KL} = 1.2075 - 1 = 0.2075$ bits.

The minimum is zero when q matches p .

Cross entropy includes the uncertainty in p ; KL isolates the mismatch.

KL divergence subtracts the uncertainty already present in p .

Entropy, Cross Entropy, and KL Divergence

Why Do Models Minimize Cross Entropy?

$$CE(p, q) = H(p) + KL(p, q)$$

FIXED TARGET p

$H(p)$ does not depend on the model parameters.

Training cannot change this term when the target distribution is fixed.

LEARNED DISTRIBUTION q

The model changes q through its parameters.

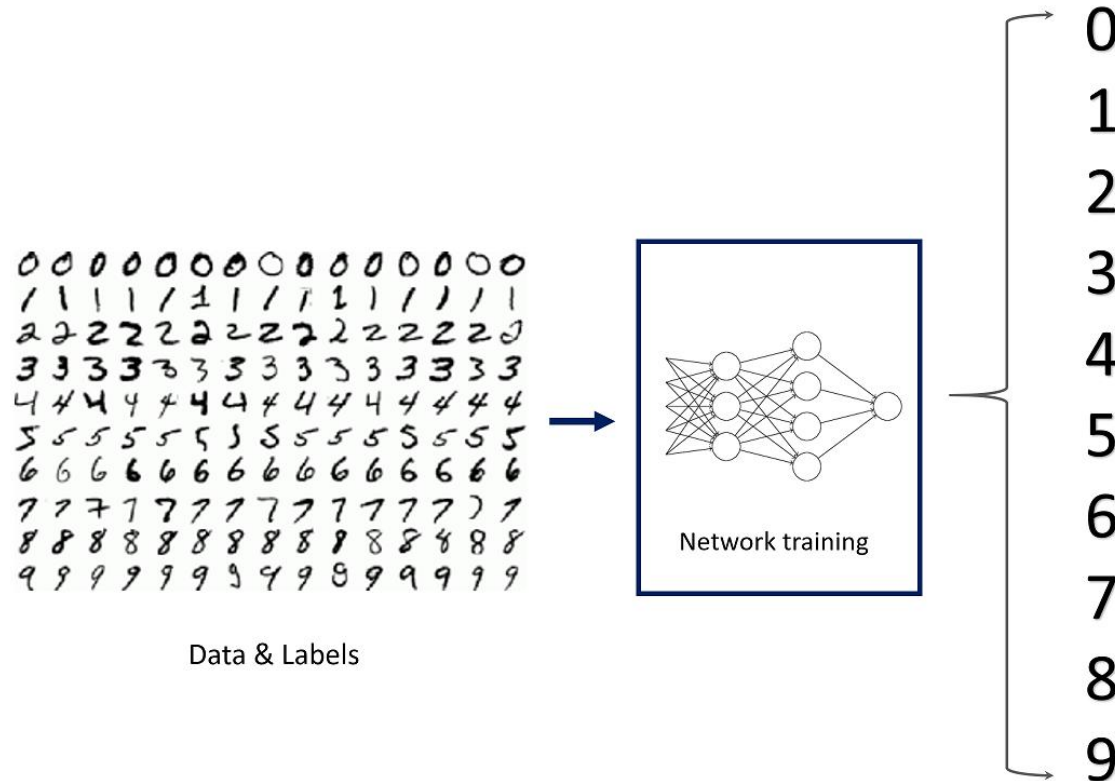
Reducing cross entropy therefore reduces the KL term.

When p is fixed, minimizing cross entropy also minimizes KL divergence.

Entropy, Cross Entropy, and KL Divergence

Classifier Training: The Observed Label

For an image labelled 5, the training target assigns $p(5) = 1$ and $p(\text{other digits}) = 0$.



Entropy, Cross Entropy, and KL Divergence

Classifier Training: The Observed Label

For an image labelled 5, the training target assigns $p(5) = 1$ and $p(\text{other digits}) = 0$.

$$\text{CE}(p, q) = -\log_2 q(5)$$

$$\text{If } q(5) = 0.80: \text{CE} = -\log_2(0.80) \approx 0.322$$

Only the term for the observed label remains because all other target weights are zero.

For a one-hot target, cross entropy depends on the probability of the correct label.

Entropy, Cross Entropy, and KL Divergence

Which Prediction Has the Smaller Loss?

Probability assigned to label 5	Cross-entropy loss
$q(5) = 0.80$	$-\log_2(0.80) \approx 0.322$ bits
$q(5) = 0.20$	$-\log_2(0.20) \approx 2.322$ bits

The label is the same in both examples.

The second prediction gives the label much less probability, so it receives a larger penalty.

Giving the observed label more probability reduces its cross-entropy loss.

Entropy, Cross Entropy, and KL Divergence

Cross Entropy and Negative Log Likelihood

$$L(\theta) = \prod_i q_{\theta}(y_i \mid x_i)$$

Multiply the model probabilities assigned to the observed labels.

$$-\log_2 L(\theta) = -\sum_i \log_2 q_{\theta}(y_i \mid x_i)$$

Taking the negative logarithm turns the product into the sum of per-example losses.

This writes the usual likelihood under independent-example modelling.

One-hot cross entropy is negative log likelihood of the observed labels.

Entropy, Cross Entropy, and KL Divergence

Three Measures, Three Questions

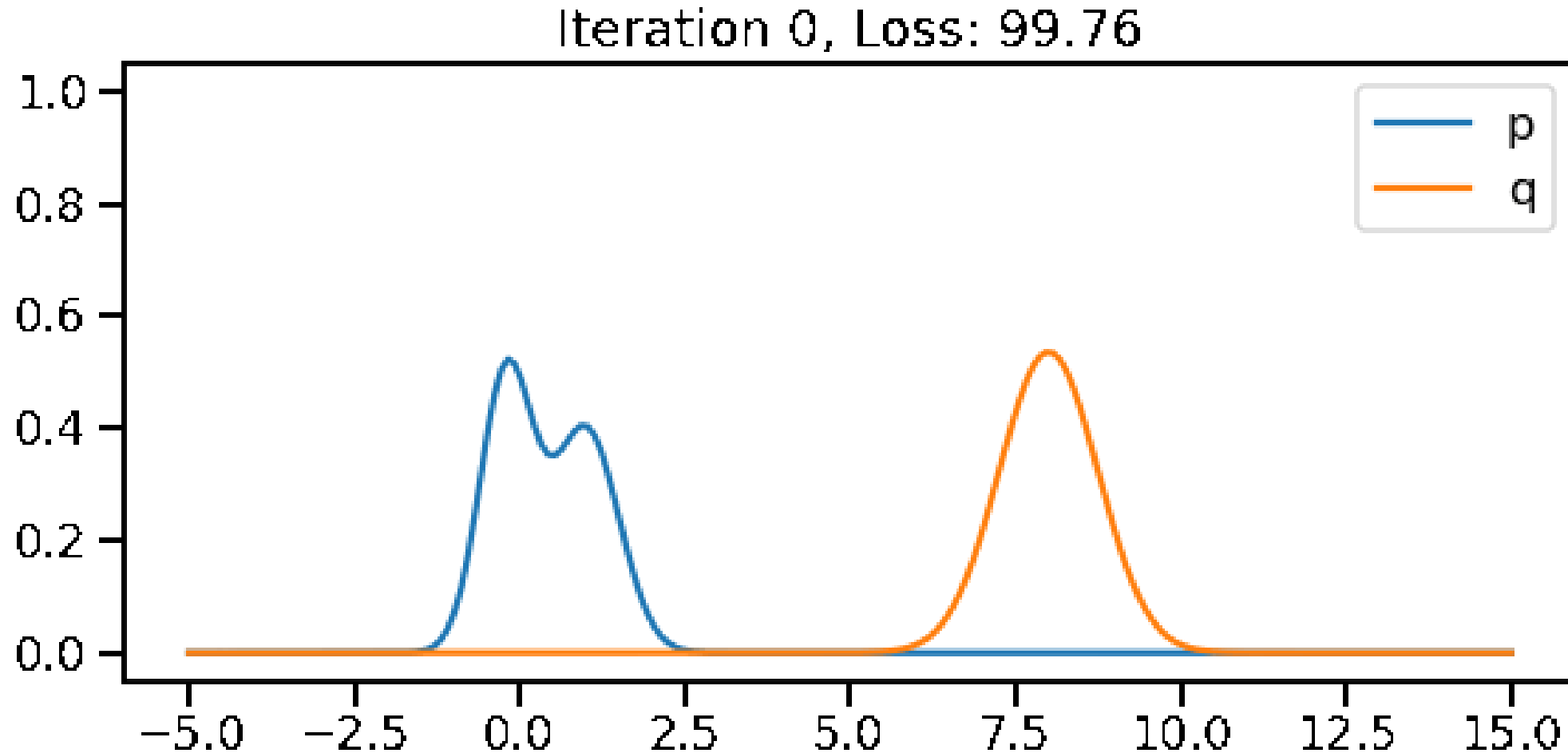
Measure	Question	Coin example
Entropy $H(p)$	How uncertain is the true distribution?	1 bit
Cross entropy $CE(p, q)$	What is the cost when we use q for data from p ?	1.2075 bits
KL divergence $KL(p, q)$	How much extra cost comes from the mismatch?	0.2075 bits

The same true distribution $p = (0.5, 0.5)$ is used throughout this comparison.

Keep the roles of p and q fixed when reading the formulas.

Entropy, Cross Entropy, and KL Divergence

Three Measures, Three Questions



Continuous Probability Distributions

From a List of Outcomes to an Interval

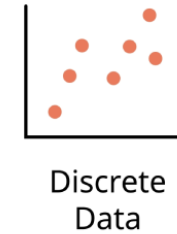
DISCRETE

Examples: coin outcomes or MNIST classes.

Add the probabilities assigned to individual outcomes.

Sum of probabilities = 1

CONTINUOUS



VS



Examples: values on a real interval.

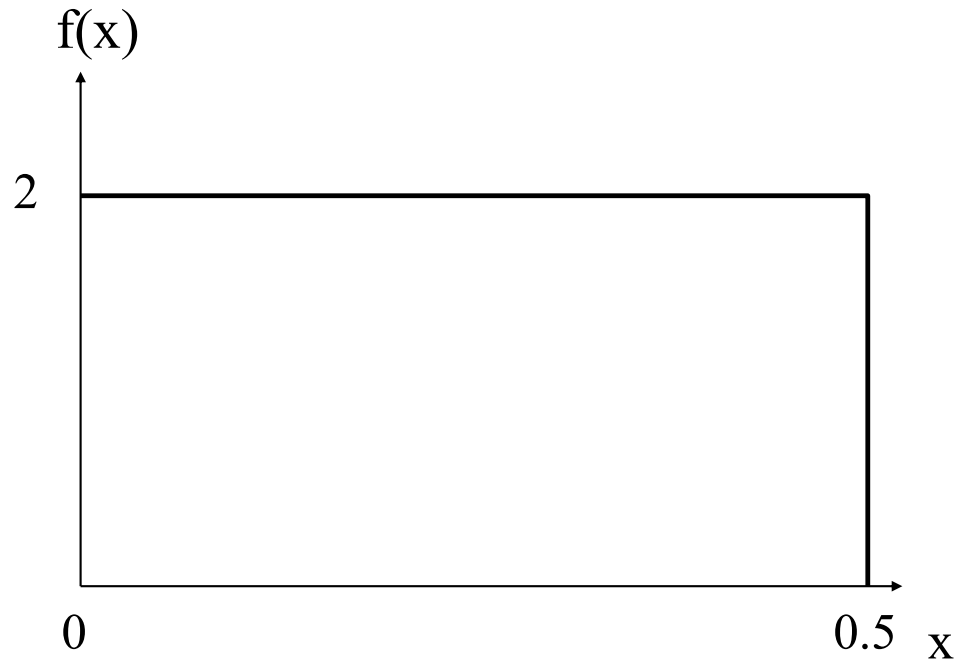
Integrate a probability density over an interval.

Total area under $f(x) = 1$

For continuous variables with a density, probability comes from area.

Continuous Probability Distributions

Uniform on $[0, 0.5]$



All values from 0 to 0.5 have the same density height.

The height is 2, not 1.

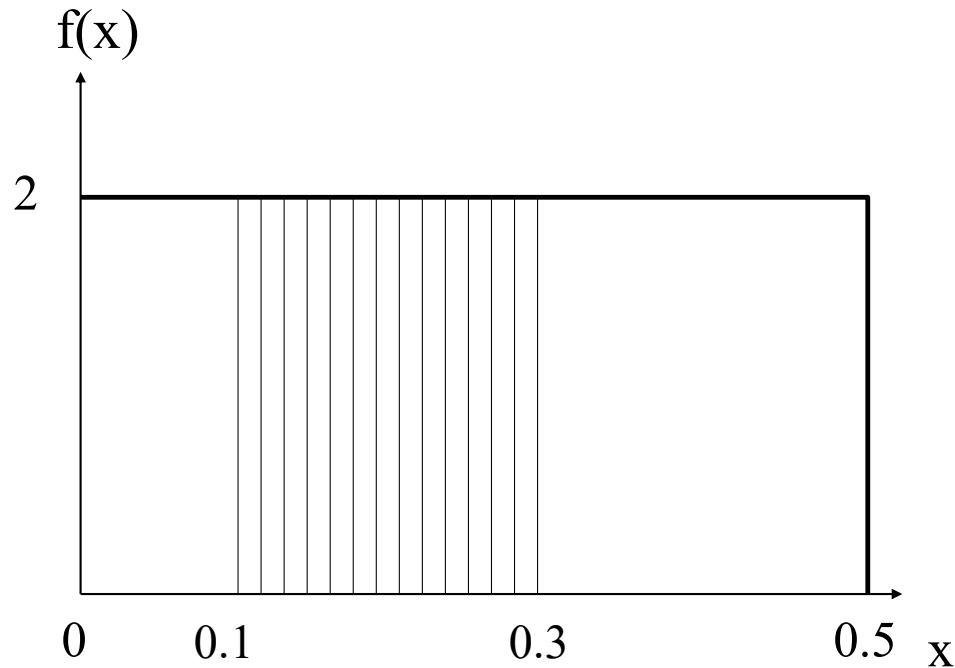
Area = width $\times$ height = $0.5 \times 2 = 1$.

Outside this interval, the density is zero.

A density of 2 over an interval of width 0.5 gives total probability 1.

Continuous Probability Distributions

Calculate an Interval Probability



$$P(0.1 \leq X \leq 0.3)$$

$$= (0.3 - 0.1) \times 2$$

$$= 0.4$$

We integrate only over the selected interval, not the entire domain.

Probability is the area above the interval of interest.

Continuous Probability Distributions

A Point Has No Width

DENSITY HEIGHT

For the uniform example, $f(0.2) = 2$.

The density can be larger than 1.

The textbook calls this evaluated height a likelihood.

The probability of an interval may be positive even though each exact point has probability zero.

A nonzero density at a point is not a nonzero point probability.

POINT PROBABILITY

$P(X = 0.2) = 0$.

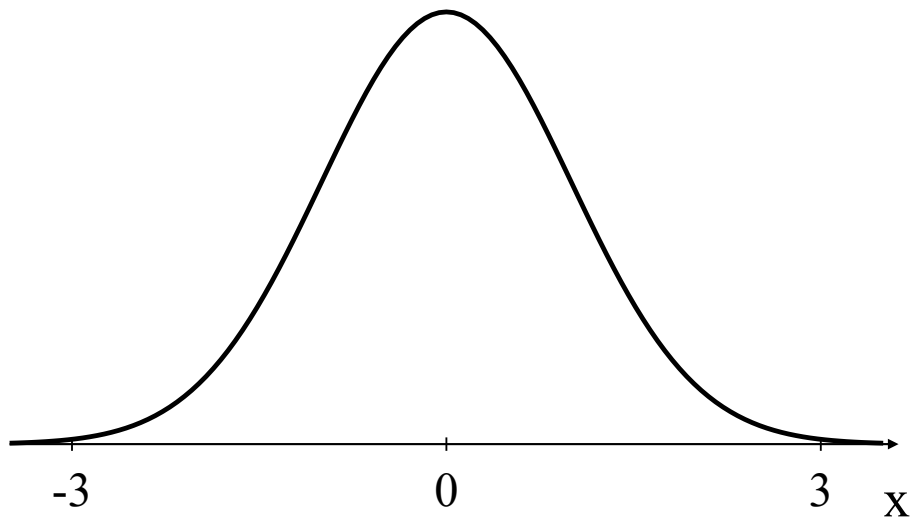
A single point has zero width.

Its area under the density curve is therefore zero.

Continuous Probability Distributions

The Gaussian Distribution

$f(x)$



The mean μ gives the centre.

The standard deviation σ describes the spread.

For a Gaussian, the mean is also the unique mode: the density is highest there.

The standard Gaussian has $\mu = 0$ and variance 1.

A Gaussian has a bell shape centred at its mean.

Continuous Probability Distributions

Read the Gaussian Formula by Its Parts

$$f(x; \mu, \sigma) = (1 / (\sigma\sqrt{2\pi})) \exp[-(x - \mu)^2 / (2\sigma^2)]$$

Distance from the mean

The term $x - \mu$ measures how far x is from the centre.

Spread and normalization

The parameter σ controls the spread; the complete density integrates to 1.

Use $\sigma > 0$. For the standard Gaussian, $\sigma = 1$.

The mean sets the centre; the standard deviation sets the scale.

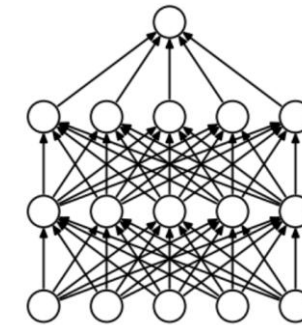
Continuous Probability Distributions

Central Limit Theorem: The Dropout Example

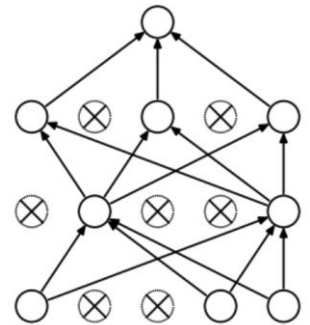
Let S be the number of masked neurons in a large layer.

$$E[S] = np, \quad \text{Var}(S) = np(1 - p)$$

$$Z = \frac{S - np}{\sqrt{np(1 - p)}}$$



(a) Standard Neural Net



(b) After applying dropout.

For many independent masks with fixed $0 < p < 1$, Z is approximately standard Gaussian.

The correctly standardized sum of many independent masks approaches a Gaussian.

Continuous Probability Distributions

Another Sum of Binary Variables

DROPOUT COUNT

Each neuron contributes 1 if masked and 0 otherwise.

It assumes independent mask decisions.

DISEASE COUNT IN A POPULATION

Each person contributes 1 if the disease is present and 0 otherwise.

It notes that these indicators may be dependent.

A count being a sum is not enough by itself to guarantee a Gaussian approximation.

Sums of binary variables may be approximately Gaussian under suitable conditions.

Continuous Probability Distributions

Continuous Expectation and Variance

$$E[X] = \int x f(x) dx$$

Average the values, weighted by their density.

$$\mu = E[X]$$

The mean is one constant describing the distribution.

$$\text{Var}(X) = \int (x - \mu)^2 f(x) dx$$

Average the squared distances from that mean.

Use the same ideas as before, replacing sums by integrals.

Continuous Probability Distributions

The Uniform Mean

For the uniform distribution on $[0, 0.5]$, the density is $f(x) = 2$.

$$E[X] = \int_0^{0.5} x \times 2 \, dx$$

$$= 0.5^2 - 0^2 = 0.25$$

Integrating $2x$ gives x squared. Evaluate at 0.5, then subtract the value at 0.

The result is halfway between the two interval endpoints.

The integral confirms the midpoint intuition: $E[X] = 0.25$.

Continuous Probability Distributions

Bayes' Theorem in the Continuous Setting

$$f(b \mid a) = \frac{f(a \mid b) f(b)}{\int f(a \mid u) f(u) du}$$

The observed value is a . The unknown value is b .

The numerator describes one candidate b together with the observation.

The denominator integrates over all candidates u .

The denominator now integrates over the possible values of the unknown variable.

Continuous Probability Distributions

Entropy, Cross Entropy, and KL as Integrals

$$H(f) = -\int f(x) \log_2 f(x) \, dx$$

$$CE(f, g) = -\int f(x) \log_2 g(x) \, dx$$

$$KL(f, g) = \int f(x) \log_2(f(x) / g(x)) \, dx$$

Continuous entropy is called differential entropy; its numerical interpretation differs from discrete entropy.

Continuous Probability Distributions

Why This Matters for Later Deep Learning

1. Choose a target distribution

The chapter points ahead to models that use Gaussian distributions.

2. Compare the learned distribution with the target

KL divergence measures the mismatch between the two.

3. Penalize the mismatch

The training objective encourages the learned distribution to move toward the target.

A divergence penalty encourages a learned distribution to match its target.

Summary

Quick Check

1. Density

Can $f(x) = 2$ be valid even though probabilities cannot exceed 1?

2. Bayes

Are “disease given positive” and “positive given disease” the same probability?

3. Cross entropy

What happens to loss when the model gives the observed label more probability?

Explain what each probability, density, or loss value means.

Summary

Quick Check: Answers

1. Yes: density is a height

The uniform example has height 2 and width 0.5, giving total area 1.

2. No: the conditioning group changes

The illustrative test example gives 80% in one direction and about 7.48% in the other.

3. The loss decreases

For a one-hot target, the loss is the negative log probability of the observed label.

Area, conditioning, and the observed label determine the interpretation.

Summary

What Did We Learn?

Topic	Core idea
Bayes' theorem	Update a belief using observed evidence.
Entropy	Measure the uncertainty within a distribution.
Cross entropy and KL	Compare a target distribution with a prediction.
Continuous distributions	Use density and area rather than point probabilities.
Deep learning connections	Classifiers, dropout, and distribution regularization.

Probability connects uncertainty, inference, and neural-network training.