機率與統計：從入門到精通

第 12 章　資訊理論

Information Theory

難度：專家

1948 年，Claude Shannon 在貝爾實驗室發表《通訊的數學理論》，以一篇論文獨力創立了一門學科。它的出發點樸素得驚人：把「一則訊息帶來多少資訊」定義為「它讓你多驚訝」，再用機率把驚訝變成可以計算的數字。由此誕生的幾個量——**熵**、**交叉熵**、**KL 散度**、**互資訊**——並不是機器學習課本裡的異國點綴：它們**就是**機器學習的損失函數與訓練目標本身。本章要把你在訓練分類器與語言模型時反覆遇到的 cross-entropy 一次講精確，並揭示一件值得回味的事：訓練一個分類器，就是在字面意義上**最小化模型與世界之間的資訊差距**。

學習目標

讀完本章後，你將能夠：

- 以**驚訝度**（surprisal）定義**熵** $H(p)$，用位元（bit）與 nat 計算離散分布的熵，並說明「熵 = 最短平均編碼長度」的源編碼意義。
- 定義**交叉熵**與 **KL 散度**，證明 $H(p,q) = H(p) + \mathrm{KL}(p\|q)$ 與 Gibbs 不等式 $\mathrm{KL}(p\|q) \geq 0$，並解釋前向 / 反向 KL 的不對稱行為。
- 說明「最小化交叉熵 = 最大化概似 = 最小化 KL」為何是同一件事；計算 one-hot 與 label smoothing 標籤下的分類損失，並以**困惑度**（perplexity）評估語言模型。
- 定義**聯合熵**、**條件熵**與**互資訊**，手算聯合分布表的 $I(X;Y)$，並說明互資訊為何能捕捉相關係數看不見的非線性相依。
- 用 Python 實作熵、交叉熵、KL 散度與互資訊，數值驗證本章的每一條恆等式。

12.1　熵：不確定性的度量

資訊理論的第一步，是替「資訊」找一個不依賴語意的度量。Shannon 的洞見是：一則訊息的資訊量，不在於它說了什麼，而在於你收到它時有多驚訝。「明天太陽會升起」幾乎沒有資訊；「明天會有日全食」資訊量巨大。差別不在字數，在機率。

定義 12.1　驚訝度（自資訊）

設結果 x 的機率為 $p(x)>0$。**其驚訝度**（surprisal，又稱自資訊 self-information）定義為

$$I(x) = -\log p(x).$$

對數底數決定單位：底數 2 時單位是**位元**（bit），自然對數時單位是 **nat**，換算為 $1\text{ nat} = 1/\ln 2 \approx 1.4427\text{ bit}$、$1\text{ bit} = \ln 2 \approx 0.6931\text{ nat}$。

這個定義把三條直覺同時變成定理：**確定的事零驚訝**（$-\log 1 = 0$）；**越罕見越驚人**（$p\to 0$ 時 $-\log p\to\infty$，機率 $1/1024$ 的事件驚訝度恰為 10 bit）；**獨立事件的驚訝可相加**（$-\log(p_A\,p_B) = -\log p_A - \log p_B$）。為什麼偏偏是對數？因為機率是相乘的、驚訝理應相加，把乘法變加法的連續函數只有對數；Shannon 更證明了唯一性——任何滿足連續性、均勻時隨選項數遞增、分組可加性的不確定性度量，必為熵的常數倍。熵不是方便的約定，而是被公理逼出的答案。

定義 12.2　熵

離散分布 p 的**熵**（entropy）是驚訝度的期望值：

$$H(p) = -\sum_{x} p(x)\log p(x) = \mathbb{E}_{p}\big[-\log p(X)\big],$$

約定 $0\log 0 = 0$（因 $\lim_{t\to 0^+} t\log t = 0$）。熵度量的是：**在看到結果之前，你平均有多不確定**。

例 12.1　熵的四張基準照片

計算下列分布的熵：(a) 公平硬幣；(b) 公平六面骰；(c) 正面機率 0.9 的偏硬幣；(d) 確定事件。

解　(a) $H = -\tfrac12\log_2\tfrac12 - \tfrac12\log_2\tfrac12 = 1\text{ bit}$（$=\ln 2\approx 0.693\text{ nat}$）：公平硬幣是「1 bit 不確定性」的原器。

(b) 六個結果各 $1/6$：$H = 6\times\tfrac16\log_2 6 = \log_2 6 \approx 2.585\text{ bit}$。等可能的選項越多，熵越大。

(c) 偏硬幣：

$$H = -0.9\log_2 0.9 - 0.1\log_2 0.1 = 0.9\times 0.152 + 0.1\times 3.322 \approx 0.469\text{ bit} \approx 0.325\text{ nat}.$$

罕見面的單項貢獻反而較大（驚訝度 3.32 bit，但很少發生），加權後總熵不到公平硬幣的一半：**分布越偏斜、越可預測，熵越低**。

(d) 某結果 $p=1$：$-1\log 1 = 0$，其餘項按約定為零，$H=0$：確定的世界沒有驚訝。

圖 12.1 把 (a)(c)(d) 串成一條曲線。

熵曲線的端點行為可以推廣成一般定理：均勻分布是「最大無知」，退化分布是「完全秩序」。

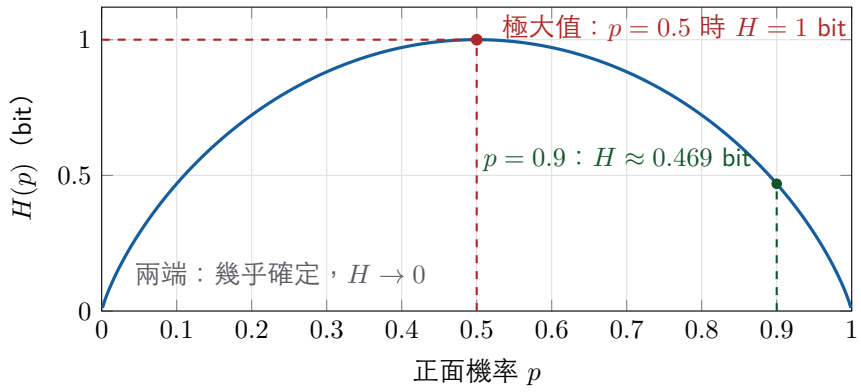


圖 12.1: 白努利分布的熵曲線 $H(p)=-p\log_2 p-(1-p)\log_2(1-p)$：最大無知（$p=0.5$）給出最大熵 1 bit，越接近確定熵越趨近零。

定理 12.3　熵的界限

設 p 是 K 個結果上的分布，則

$$0\le H(p)\le \log K.$$

下界等號成立若且唯若 p 為退化分布（某結果機率為 1）；上界等號成立若且唯若 p 為均勻分布。

證明　下界：每項 $-p(x)\log p(x)\ge 0$，故 $H\ge 0$；總和為零要求每個 $p(x)\in\{0,1\}$，配合 $\sum_x p(x)=1$ 得退化分布。上界：log 是凹函數，由 Jensen 不等式（第 6 章）

$$H(p)=\mathbb{E}_p\Big[\log\frac{1}{p(X)}\Big]\le \log \mathbb{E}_p\Big[\frac{1}{p(X)}\Big]=\log\big|\{x:p(x)>0\}\big|\le \log K.$$

log 嚴格凹，等號要求 $1/p(X)$ 為常數且支撐涵蓋全部 K 個結果，即 p 均勻。（練習 12.8 用 Lagrange 乘數法給出另一個證明。）∎

熵還有一個乍看毫不相干的身分：**壓縮的極限**。

例 12.2　最優編碼：常見的字給短碼

資料源輸出四個符號 A, B, C, D，機率為 $\frac12,\frac14,\frac18,\frac18$。求熵，並設計一組二元編碼使平均碼長達到熵。

解　熵為

$$H=\tfrac12\log_2 2+\tfrac14\log_2 4+\tfrac18\log_2 8+\tfrac18\log_2 8=\tfrac12+\tfrac12+\tfrac38+\tfrac38=1.75\text{ bit}.$$

給每個符號長度恰為驚訝度 $-\log_2 p(x)$ 的碼字：A $=0$、B $=10$、C $=110$、D $=111$。這是一組**前綴碼**（prefix code）：沒有碼字是另一碼字的開頭，位元流可即時無歧義解碼。平均碼長

$$\mathbb{E}[L]=\tfrac12(1)+\tfrac14(2)+\tfrac18(3)+\tfrac18(3)=1.75\text{ bit}=H(p),$$

優於固定長度編碼的 2 bit / 符號。「常見的字給短碼」正是霍夫曼編碼（Huffman coding）的核心直觀——本例機率都是 2 的負整數次方，直接取 $-\log_2 p(x)$ 為碼長即達極限，不需

完整演算法。

定理 12.4　Shannon 源編碼定理（敘述）

設資料源以分布 p 獨立產生符號。任何唯一可解碼的二元編碼，其平均碼長必滿足 $\mathbb{E}[L] \geq H(p)$（bit）；且存在前綴碼使 $\mathbb{E}[L] < H(p) + 1$，將符號打包成區塊更可任意逼近 $H(p)$。證明超出本書範圍。

核心觀念

熵的三位一體：**平均驚訝 = 最少平均是非題數 = 最短平均編碼長度**。不確定性與可壓縮性是同一個量的兩面：誰能精準預測，誰就能極致壓縮——這正是「大型語言模型本質上是壓縮器」一說的理論根據。

常見陷阱

log 底數混用是本章最常見的計算事故。同一個分布，$H = 1$ bit 也 $= 0.693$ nat。三條自保規則：(1) 報告熵、交叉熵、KL 時**永遠標明單位**；(2) 深度學習框架（PyTorch 的 `F.cross_entropy` 等）一律用自然對數、回傳 nat，資訊理論文獻慣用 bit；(3) 困惑度必須與單位配對：$\text{PPL} = 2^{H(\text{bit})} = e^{H(\text{nat})}$，錯配會得到毫無意義的數字。換算永遠是 $H_{\text{bit}} = H_{\text{nat}} / \ln 2$。

12.2　交叉熵與 KL 散度

熵回答「世界本身有多不確定」，前提是你知道真分布。但建模者的處境永遠是：資料來自 p，你手上只有模型 q。照 q 設計編碼、下注、決策，平均要付出多少代價？資訊理論給這筆帳兩個名字：總帳叫交叉熵，超收的部分叫 KL 散度。

定義 12.5　交叉熵

真分布 p 與模型分布 q 之間的**交叉熵**（cross-entropy）為

$$H(p, q) = -\sum_{x} p(x) \log q(x) = \mathbb{E}_{p}\big[-\log q(X)\big]:$$

「p 產生的結果，在 q 眼中的平均驚訝」。編碼讀法：照 q 設計碼本（給 x 長度 $-\log q(x)$ 的碼字），資料卻依 p 出現，實付平均碼長就是 $H(p, q)$。注意 $H(p, p) = H(p)$。

定義 12.6　KL 散度（相對熵）

Kullback–Leibler 散度（KL divergence，又稱相對熵 relative entropy）定義為

$$\mathrm{KL}(p \| q) = \sum_{x} p(x) \log \frac{p(x)}{q(x)} = \mathbb{E}_{p}\left[\log \frac{p(X)}{q(X)}\right],$$

約定 $0 \log \frac{0}{q(x)} = 0$；若存在 x 使 $p(x) > 0$ 而 $q(x) = 0$，則 $\mathrm{KL}(p\|q) = \infty$（模型宣稱不可能的事發生了，無限驚訝）。

命題 12.7　交叉熵分解

對任意離散分布 p, q：

$$H(p, q) = H(p) + \mathrm{KL}(p\|q).$$

證明

$$\mathrm{KL}(p\|q) = \sum_x p(x) \log p(x) - \sum_x p(x) \log q(x) = -H(p) + H(p, q). \qquad \blacksquare$$

也就是說：**實付碼長 = 理論下限 + 浪費**。要讓「浪費」名副其實，還需要它非負。

定理 12.8　Gibbs 不等式

對任意離散分布 p, q：$\mathrm{KL}(p\|q) \geq 0$，等號成立若且唯若 $p = q$。因此交叉熵永遠不小於熵：$H(p, q) \geq H(p)$，用錯分布必然多付。

證明　設 $S = \{x : p(x) > 0\}$。若有 $x \in S$ 使 $q(x) = 0$ 則 $\mathrm{KL} = \infty$，不等式自動成立。否則由 log 的凹性與 Jensen 不等式，

$$-\mathrm{KL}(p\|q) = \sum_{x \in S} p(x) \log \frac{q(x)}{p(x)} \leq \log \sum_{x \in S} p(x) \cdot \frac{q(x)}{p(x)} = \log \sum_{x \in S} q(x) \leq \log 1 = 0.$$

log 嚴格凹，第一個等號要求 $q(x)/p(x)$ 在 S 上為常數；第二個等號要求 q 在 S 之外無質量。合併得 $q = p$。（練習 12.9 給出只用 $\ln t \leq t - 1$ 的初等證明。）　$\blacksquare$

例 12.3　交叉熵、KL 與其不對稱

真分布 $p = (0.5,\, 0.5)$，模型 $q = (0.9,\, 0.1)$。(a) 以 nat 計算 $H(p)$、$H(p, q)$、$\mathrm{KL}(p\|q)$，驗證分解恆等式。(b) 計算反向的 $\mathrm{KL}(q\|p)$，確認不對稱。

解　(a) 熵：$H(p) = \ln 2 \approx 0.6931$ nat（$= 1$ bit）。交叉熵：

$$H(p, q) = -\big[0.5 \ln 0.9 + 0.5 \ln 0.1\big] = 0.5 \times 0.1054 + 0.5 \times 2.3026 \approx 1.2040 \text{ nat}.$$

KL 散度（由定義直接算）：

$$\mathrm{KL}(p\|q) = 0.5 \ln \frac{0.5}{0.9} + 0.5 \ln \frac{0.5}{0.1} = 0.5 \ln \tfrac{5}{9} + 0.5 \ln 5 = \ln \tfrac{5}{3} \approx 0.5108 \text{ nat}.$$

驗證：$H(p) + \mathrm{KL}(p\|q) = 0.6931 + 0.5108 = 1.2040 = H(p, q)$。✓　模型自信地押注第一個結果，真相卻是五五波——每個樣本平均多付 0.51 nat 的「驚訝稅」。

(b) 反向：

$$\mathrm{KL}(q\|p) = 0.9 \ln 1.8 + 0.1 \ln 0.2 \approx 0.5290 - 0.1609 = 0.3681 \text{ nat},$$

與 $\mathrm{KL}(p\|q) \approx 0.5108$ nat 不相等：**KL 不對稱**。$\mathrm{KL}(p\|q)$ 以真相 p 加權，重罰「真相常見、模型認為罕見」；反向以 q 加權，帳就不同——在 KL 的世界裡**自信的錯比保守的含糊更貴**。

核心觀念

$\mathrm{KL}(p\|q)$ 是「用 q 的碼本編碼 p 的資料所**浪費**的平均位元數」——模型到真相的資訊差距。因為 $H(p)$ 與模型無關，

$$\min_q H(p,q) \iff \min_q \mathrm{KL}(p\|q)\,:$$

最小化交叉熵就是**把模型分布拉向資料分布**。訓練時下降的 loss 曲線，其實是模型到真相的 KL 距離，加上一條壓不破的地板 $H(p)$——資料固有的隨機性（aleatoric uncertainty）。

不對稱性不只是代數瑕疵，它決定了**最小化哪個方向**會得到行為迥異的模型。設真相 p 雙峰，而模型族 q 只能表達單峰：

- **前向 KL** $\mathrm{KL}(p\|q)$（以 p 加權）：凡 $p>0$ 而 $q\approx 0$ 之處代價爆炸，q 被迫鋪滿 p 的所有眾數（zero-avoiding），寧可攤平也不漏接——**覆蓋均值**（mean-seeking）。最大概似訓練最小化的正是它（12.3 節）。
- **反向 KL** $\mathrm{KL}(q\|p)$（以 q 加權）：q 只需避開「$p\approx 0$ 之處」（zero-forcing），可以放棄部分眾數、緊抱一峰——**抓眾數**（mode-seeking）。變分推論（variational inference，第 9 章）最小化反向 KL，這是它系統性低估後驗變異數的根本原因。

記憶口訣：**前向 KL 說「p 在哪，q 就得在哪」；反向 KL 說「q 在哪，p 最好在哪」**。讀論文看到 KL 項，第一件事永遠是確認：哪個分布在前、哪個在後。

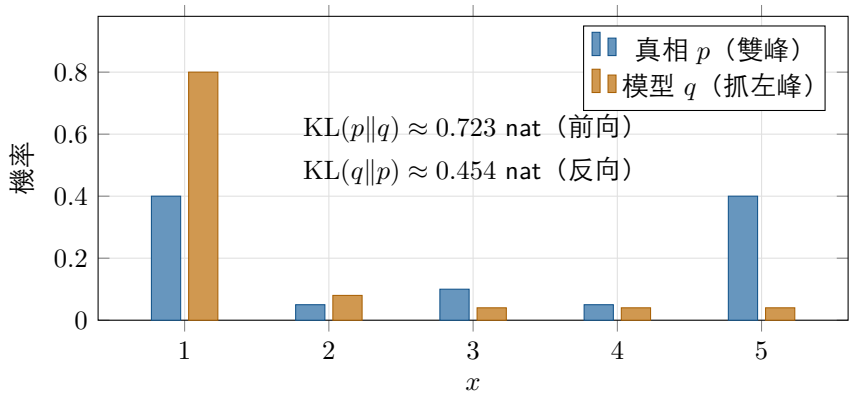


圖 12.2: KL 的不對稱：q 押注 p 的左峰、幾乎放棄右峰（$x=5$ 處 $p=0.4$ 而 $q=0.04$）。前向 KL 以 p 加權、重罰漏接；反向 KL 以 q 加權、幾乎看不見這個缺口——mode-seeking 的數值寫照。

常見陷阱

KL 不是距離（度量，metric）：(1) 不對稱（例 12.3）；(2) 不滿足三角不等式；(3) 可以無窮大——只要 q 在 p 的支撐上有零點。實務上模型輸出經 softmax 處處為正、或以 ε 平滑與 label smoothing 避開這顆地雷；需要對稱且有界的替代品時可用 Jensen–Shannon 散度（原始 GAN 理論分析的量）。口語說「KL 距離」無妨，寫證明時別把度量性質偷渡進來。

12.3　交叉熵損失：one-hot、最大概似與困惑度

現在把交叉熵接上你每天在跑的訓練迴圈：分類器的損失函數、語言模型的評估指標，全是上一節的量換了戰場。

one-hot 標籤下的坍縮。K 類分類的真標籤通常是 **one-hot**（獨熱）向量：正確類別 c 得機率 1、其餘為 0，即 $p = e_c$。代入交叉熵，總和只剩一項：

$$H(p, q) = -\sum_{j=1}^{K} p_j \log q_j = -\log q_c.$$

整個損失**只看模型分給正確答案多少機率**；錯誤類別之間如何瓜分剩餘機率，損失值一概不管（它們影響的是梯度，見練習 12.3）。這就是框架裡 cross-entropy loss 的真身：每筆樣本取 $-\log q$(正確類別)，對批次平均。圖 12.3 是這個場景的標準畫面。

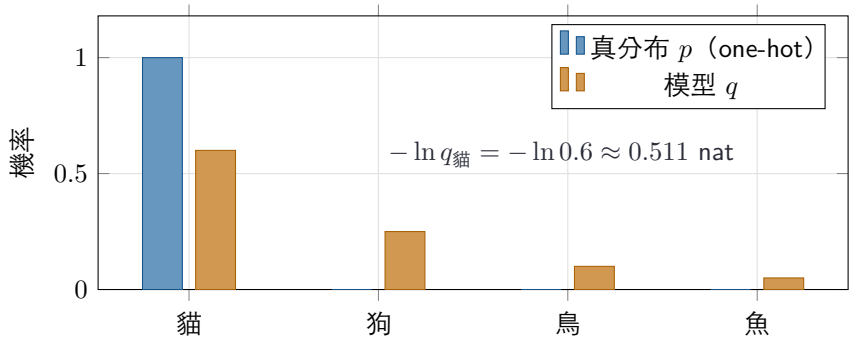


圖 12.3: one-hot 真標籤（貓）與模型輸出 $q = (0.6, 0.25, 0.10, 0.05)$：交叉熵坍縮成 $-\log q$(正確類別)，恰好懲罰「模型沒分給正確答案的機率」。

感受一下損失的階梯：好模型 $q_{\text{貓}} = 0.9$ 給 $-\ln 0.9 \approx 0.105$ nat，普通模型 $q_{\text{貓}} = 0.6$ 給 0.511 nat，糟模型 $q_{\text{貓}} = 0.05$ 給 $-\ln 0.05 \approx 2.996$ nat——從「自信且對」到「自信且錯」差近 30 倍。$-\log$ 的陡坡是刻意的：它重罰過度自信的錯誤。

為什麼是交叉熵？因為它就是最大概似。這不是修辭，是恆等式。

命題 12.9　一個目標，三個名字

設訓練樣本 $x_1, \dots, x_n \overset{\text{i.i.d.}}{\sim} p$，$q_\theta$ 為參數化模型，$\hat{p}$ 為樣本的經驗分布（$\hat{p}(x) = \frac{1}{n}\#\{i : x_i = x\}$）。則

$$\arg\max_{\theta} \prod_{i=1}^{n} q_\theta(x_i) = \arg\min_{\theta} H(\hat{p}, q_\theta) = \arg\min_{\theta} \mathrm{KL}(\hat{p} \| q_\theta).$$

證明　取對數、除以 n（單調變換，不改變極值點）：

$$-\frac{1}{n}\sum_{i=1}^{n} \log q_\theta(x_i) = -\sum_{x} \hat{p}(x) \log q_\theta(x) = H(\hat{p}, q_\theta) = \underbrace{H(\hat{p})}_{\text{與}\theta\text{ 無關}} + \mathrm{KL}(\hat{p} \| q_\theta). \qquad \blacksquare$$

當 $n \to \infty$ 時 $\hat{p} \to p$（大數法則，第 6 章），最大概似估計（第 7 章）漸近地就是在最小化「到真相的前向 KL」。**「交叉熵損失」「最大概似」「最小化 KL 散度」——三個名字，同一個訓練目標。**深度學習沒有發明新的損失，它把 1948 年的量放上了 GPU。

定義 12.10　困惑度

語言模型 q 對測試序列 $x_1,\dots,x_n$ 的每詞平均交叉熵為 $H_n=-\frac{1}{n}\sum_{i}\log_2 q(x_i\mid x_{<i})$（bit）。**困惑度**（perplexity）定義為

$$\mathrm{PPL}=2^{H_n}=\Big(\prod_{i=1}^{n} q(x_i\mid x_{<i})\Big)^{-1/n},$$

即逆幾何平均機率；H_n 以 nat 計則 $\mathrm{PPL}=e^{H_n}$。解讀：模型每步的不確定性**等效於在 PPL 個等可能選項中亂猜**（等效分支數）。

例 12.4　語言模型的困惑度

(a) 某語言模型每詞交叉熵 3.5 bit，求困惑度。(b) 詞彙表 50,000 詞、完全沒學過語言的均勻模型，困惑度是多少？

解　(a) $\mathrm{PPL}=2^{3.5}\approx 11.3$：每預測一個詞，平均像在 11 個等可能候選中挑選——遠小於詞彙表，這就是「學會語言統計結構」的量化意義。

(b) 均勻模型每詞交叉熵 $\log_2 50000\approx 15.61$ bit，$\mathrm{PPL}=50{,}000$：困惑度的天花板是詞彙量（均勻最大熵，定理 12.3）。

label smoothing：故意不給滿分的標籤。one-hot 目標有個隱憂：要讓 $-\log q_c\to 0$，模型得把 q_c 推向 1，即 logit 差距拉到無限——訓練後期梯度全花在放大自信上，容易過擬合且校準變差。**標籤平滑**（label smoothing）把目標換成 one-hot 與均勻分布的混合：

$$p^{\mathrm{ls}}=(1-\varepsilon)\,e_c+\frac{\varepsilon}{K}\mathbf{1},\qquad H(p^{\mathrm{ls}},q)=(1-\varepsilon)\big(-\log q_c\big)+\frac{\varepsilon}{K}\sum_{j}\big(-\log q_j\big).$$

例 12.5　label smoothing 的交叉熵

四類分類，真類別為第 1 類，$\varepsilon=0.1$，模型輸出 $q=(0.6,0.25,0.10,0.05)$。計算平滑後的目標、交叉熵與損失的理論下限。

解　平滑目標：正確類得 $1-0.1+0.1/4=0.925$、其餘三類各得 0.025，即 $p^{\mathrm{ls}}=(0.925,\,0.025,\,0.025,\,0.025)$。交叉熵（nat）：

$$H(p^{\mathrm{ls}},q)=-\big[0.925\ln 0.6+0.025\ln 0.25+0.025\ln 0.10+0.025\ln 0.05\big]\approx 0.640\ \text{nat},$$

比 one-hot 的 0.511 nat 高——多出的是「沒把 2.5% 機率留給每個錯誤類」的罰款。關鍵在下限：由 Gibbs 不等式（定理 12.8），$H(p^{\mathrm{ls}},q)\ge H(p^{\mathrm{ls}})$，等號在 $q=p^{\mathrm{ls}}$ 成立，且

$$H(p^{\mathrm{ls}})=-\big[0.925\ln 0.925+3\times 0.025\ln 0.025\big]\approx 0.0721+0.2767=0.349\ \text{nat}>0.$$

損失地板不再是零，最優解也不再是 $q_c\to 1$ 而是 $q=p^{\mathrm{ls}}$：模型被明文禁止過度自信（練習 12.10 掃描不同 ε）。

這如何驅動機器學習／ AI

資訊理論幾乎逐字供應現代機器學習的目標函數：

- **分類與語言建模**：預設損失就是 $H(p,q)$；最小化它 = 最大概似 = 最小化前向 KL（命題 12.9）。
- **VAE**：目標 ELBO（Evidence Lower BOund，證據下界）= $\mathbb{E}_{q(z|x)}[\log p(x|z)] - \mathrm{KL}\big(q(z|x)\|p(z)\big)$：重建項加「潛在變數向先驗靠攏」的 KL 正則。
- **變分推論**：整個框架就是 $\min_q \mathrm{KL}\big(q\|p(\cdot\mid x)\big)$——反向 KL，因此抓眾數、低估變異。
- **擴散模型**：訓練目標可寫成一串「反向去噪步驟 vs. 前向加噪後驗」之間的 KL 項之和。
- **知識蒸餾**（knowledge distillation）：學生以老師的軟化分布（帶溫度的 softmax）為目標最小化交叉熵。
- **label smoothing**：以非零熵的目標抑制過度自信、改善校準。

在深度學習裡最小化損失，幾乎總是在最小化兩個分布之間的散度——差別只在**哪兩個分布、哪個方向**。

12.4 互資訊：相依性的度量

熵度量單一變數的不確定性；下一個問題是：兩個變數**共享**多少資訊？氣壓計讀數透露多少明天的天氣？影像的內部表徵保留了多少標籤資訊？

定義 12.11　聯合熵與條件熵

設 (X,Y) 有聯合分布 $p(x,y)$。**聯合熵**（joint entropy）為

$$H(X,Y) = -\sum_{x,y} p(x,y)\log p(x,y),$$

即把數對 (X,Y) 視為單一隨機物件的熵。**條件熵**（conditional entropy）為

$$H(Y\mid X) = \sum_x p(x)\,H(Y\mid X=x) = -\sum_{x,y} p(x,y)\log p(y\mid x)\,:$$

知道 X 之後，對 Y 剩餘的平均不確定性。

命題 12.12　鏈式規則

$H(X,Y) = H(X) + H(Y\mid X) = H(Y) + H(X\mid Y)$。（證明見練習 12.5。）直觀：描述數對的最短方式，是先描述 X（花 $H(X)$），再描述已知 X 之下的 Y（花 $H(Y|X)$）。

定義 12.13　互資訊

X 與 Y 的**互資訊**（mutual information）定義為

$$I(X;Y) = \mathrm{KL}\big(p_{X,Y} \,\|\, p_X \otimes p_Y\big) = \sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)\,p(y)},$$

其中 $p_X \otimes p_Y$ 是兩個邊際相乘所得的「假想獨立版」聯合分布：互資訊度量**聯合分布離獨立有多遠**。

命題 12.14　互資訊的性質

1. **三種等價形式**：$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X,Y)$。
2. **非負性**：$I(X;Y) \geq 0$，且 $I(X;Y) = 0$ 若且唯若 X 與 Y 獨立。
3. **對稱性**：$I(X;Y) = I(Y;X)$。
4. **自互資訊**：$I(X;X) = H(X)$——變數與自己共享全部資訊，熵因此又稱「自資訊」。

證明　(1) 代入 $p(x,y) = p(x)p(y|x)$：

$$I(X;Y) = \sum_{x,y} p(x,y)\log\frac{p(y|x)}{p(y)} = H(Y) - H(Y|X),$$

對稱地得 $H(X) - H(X|Y)$；再以鏈式規則 $H(Y|X) = H(X,Y) - H(X)$ 代換得第三式。(2) 非負性是 Gibbs 不等式（定理 12.8）套在 $\mathrm{KL}(p_{X,Y} \| p_X \otimes p_Y)$ 上；等號條件即 $p(x,y) = p(x)p(y)$ 處處成立——獨立的定義（完整論證見練習 12.7）。(3) 定義對稱。(4) $H(X|X) = 0$。 ■

性質 (1)(2) 再送一個推論：$H(X|Y) = H(X) - I(X;Y) \leq H(X)$——**條件化平均不增熵**。注意「平均」二字：對特定觀察值 y，$H(X \mid Y = y)$ 仍可大於 $H(X)$，但對 y 加權平均後必定不增。

直觀理解

把 $H(X)$、$H(Y)$ 想成兩個部分重疊的圓的面積：交疊區是 $I(X;Y)$，X 圓扣掉交疊是 $H(X|Y)$，聯集是 $H(X,Y)$。所有恆等式都能從圖上讀出——「$I = H(X)+H(Y)-H(X,Y)$」正是第 1 章排容原理的資訊版。

例 12.6　互資訊手算：氣壓計與降雨

令 X 表示明天是否下雨（1 = 下雨）、Y 表示今晚氣壓（1 = 低壓），聯合分布為

	$Y=0$（高壓）	$Y=1$（低壓）	$p_X(x)$
$X=0$（不下雨）	0.4	0.1	0.5
$X=1$（下雨）	0.1	0.4	0.5
$p_Y(y)$	0.5	0.5	1

求 $H(X)$、$H(X|Y)$ 與 $I(X;Y)$（bit）。

解　邊際皆為 $(0.5, 0.5)$，故 $H(X) = 1$ bit。條件熵：已知 $Y = 0$ 時 $p(X = 0 \mid Y = 0) = 0.4/0.5 = 0.8$，

$$H(X \mid Y = 0) = -0.8 \log_2 0.8 - 0.2 \log_2 0.2 \approx 0.258 + 0.464 = 0.722 \text{ bit};$$

由對稱性 $H(X \mid Y = 1)$ 相同，故 $H(X|Y) = 0.722$ bit。互資訊（法一，恆等式）：

$$I(X;Y) = H(X) - H(X|Y) = 1 - 0.722 = 0.278 \text{ bit}.$$

法二（定義直接算）：四格的 $p(x,y)/\big(p(x)p(y)\big)$ 為對角 $0.4/0.25 = 1.6$、反對角 $0.1/0.25 = 0.4$，

$$I = 2 \times 0.4 \log_2 1.6 + 2 \times 0.1 \log_2 0.4 \approx 0.542 - 0.264 = 0.278 \text{ bit}.\checkmark$$

解讀：看一眼氣壓計，平均把「明天是否下雨」的不確定性從 1 bit 降到 0.72 bit——氣壓計為你買到 0.28 bit 的資訊。

相關係數看不見的相依，互資訊看得見：設 X 均勻分布於 $\{-1, 0, 1\}$、$Y = X^2$，則 $\mathrm{Cov}(X,Y) = \mathbb{E}[X^3] = 0$，零相關；但 Y 完全由 X 決定（$H(Y|X) = 0$），故 $I(X;Y) = H(Y) = -\frac{1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} \approx 0.918$ bit > 0。相關係數（第 5 章）只偵測線性關係；互資訊偵測**任何形式**的相依——$I(X;Y) = 0$ 才是「毫無關係（獨立）」的完整證書。

這如何驅動機器學習／ AI

互資訊是「學到好表徵」的通用貨幣：

- **特徵選擇**：挑 $I(\text{特徵}; \text{標籤})$ 最大的特徵；決策樹分裂準則「資訊增益」（information gain）就是互資訊本尊。
- **對比學習**（contrastive learning）：InfoNCE 損失是互資訊的可計算下界——SimCLR、CLIP 都在最大化兩個視角表徵間的互資訊（InfoMax 原理）。
- **資訊瓶頸**（information bottleneck）：好表徵 Z 應壓縮輸入（$I(X;Z)$ 小）同時保留任務資訊（$I(Z;Y)$ 大）。
- **解糾纏表徵**（disentanglement）：VAE 變體以互資訊懲罰讓潛在維度各管一個生成因子。

12.5 Python 實作

理論之後，動手驗證。以下程式把本章每個量都算一遍：熵的基準值、交叉熵與 KL 的恆等式與不對稱性、one-hot 損失的坍縮、互資訊。

```
import numpy as np
```

```python

LN2 = np.log(2.0)

def entropy(p, unit="nat"):
    """熵 H(p) = -sum p log p。unit 可選 "nat" 或 "bit"。"""
    p = np.asarray(p, dtype=float)
    p = p[p > 0]                          # 約定 0 log 0 = 0
    h = -np.sum(p * np.log(p))
    return h / LN2 if unit == "bit" else h

def cross_entropy(p, q, unit="nat"):
    """交叉熵 H(p,q) = -sum p log q（p>0 之處需 q>0）。"""
    p, q = np.asarray(p, float), np.asarray(q, float)
    m = p > 0
    h = -np.sum(p[m] * np.log(q[m]))
    return h / LN2 if unit == "bit" else h

def kl(p, q, unit="nat"):
    """KL 散度 KL(p||q) = H(p,q) - H(p)，恆非負。"""
    return cross_entropy(p, q, unit) - entropy(p, unit)

# --- 1. 熵：越難預測，熵越高 ---
print(entropy([0.5, 0.5], "bit"))       # 1.0（公平硬幣）
print(entropy(np.ones(6) / 6, "bit"))   # 2.585 = log2(6)（公平骰子）
print(entropy([0.9, 0.1], "bit"))       # 0.469（偏硬幣：較可預測）

# --- 2. 交叉熵、KL 與恆等式 CE = H + KL（nat）---
p = np.array([0.5, 0.5])                # 真分布
q = np.array([0.9, 0.1])                # 模型
print(entropy(p), cross_entropy(p, q))  # 0.6931  1.2040
print(kl(p, q), kl(q, p))               # 0.5108  0.3681（不對稱！）
print(np.isclose(cross_entropy(p, q), entropy(p) + kl(p, q)))  # True

# --- 3. one-hot 標籤：交叉熵坍縮成 -log q(正確類) ---
q_model = np.array([0.60, 0.25, 0.10, 0.05])  # 貓、狗、鳥、魚
onehot = np.array([1.0, 0.0, 0.0, 0.0])       # 真標籤：貓
print(cross_entropy(onehot, q_model))   # 0.5108 = -ln(0.6)

# --- 4. 互資訊：I(X;Y) = KL(聯合 || 邊際乘積) ---
joint = np.array([[0.4, 0.1],
                  [0.1, 0.4]])          # 例 12.6 的聯合分布
px = joint.sum(axis=1, keepdims=True)
py = joint.sum(axis=0, keepdims=True)
```

```
print(kl(joint.ravel(), (px * py).ravel(), "bit"))  # 0.2781
```

輸出與本章手算全部吻合：恆等式在浮點精度內成立、兩個方向的 KL 明顯不同、互資訊 0.2781 bit 與例 12.6 一致。實務提醒：PyTorch 的 `F.cross_entropy` 接收 logits（內部自動 log-softmax）而非機率、回傳 nat；自行實作時 $0\log 0$ 與 $\log 0$ 的處理（如上面的遮罩）是最容易踩雷的兩處。

本章總結

- **驚訝度** $-\log p(x)$（底數 2 是 bit、ln 是 nat）；**熵** $H(p)=\mathbb{E}[-\log p(X)]$ 是期望驚訝，也是最短平均編碼長度（源編碼定理）。界限 $0\le H\le \log K$：退化取 0，均勻取 $\log K$（最大無知）。
- **交叉熵** $H(p,q)=\mathbb{E}_p[-\log q(X)]$ 用真相 p 為模型 q 打分；恆等式 $H(p,q)=H(p)+\mathrm{KL}(p\|q)$ 拆成「固有隨機性 + 模型的浪費」。
- **Gibbs 不等式**：$\mathrm{KL}(p\|q)\ge 0$，等號若且唯若 $p=q$。因 $H(p)$ 固定，最小化交叉熵 = 最小化 $\mathrm{KL}(p\|q)$ = 最大概似——一個目標，三個名字。
- **KL 不對稱**且非度量：前向覆蓋所有眾數（mean-seeking）、反向抓單一眾數（mode-seeking）；支撐不匹配時可為無窮。
- one-hot 使交叉熵坍縮為 $-\log q$(正確類別)；**label smoothing** 抬高損失地板、抑制過度自信；**困惑度** $\mathrm{PPL}=2^{H(\text{bit})}=e^{H(\text{nat})}$ 是「等效分支數」。
- **互資訊** $I(X;Y)=\mathrm{KL}(p_{X,Y}\|p_X\otimes p_Y)=H(X)-H(X|Y)\ge 0$：度量任意（含非線性）相依，$I=0$ 若且唯若獨立——零相關卻互資訊為正是可能的。
- 這些量定義了現代 ML 的目標：分類損失、VAE 的 ELBO、變分推論、擴散模型訓練、知識蒸餾、對比學習（InfoNCE）與資訊瓶頸。

12.6 練習題

建議先獨立完成再對照詳解。

練習 12.1 (基礎)　計算下列各量：(a) 公平四面骰的熵（bit）；(b) $p=(0.7,0.2,0.1)$ 的熵（bit）；(c) 將 (b) 與三結果均勻分布的熵比較，說明大小關係為何符合定理 12.3；(d) 把 (b) 的答案換算成 nat；(e) 機率 1/64 的事件發生時的驚訝度（bit）；(f) 某語言模型每詞交叉熵為 4.2 bit，求其困惑度；若另一框架以 nat 回報同一模型為 2.9110 nat，驗證兩者一致。

練習 12.2 (基礎)　設真分布 $p=(0.25,0.75)$、模型 $q=(0.5,0.5)$。(a) 求 $H(p)$（bit）。(b) 求 $H(p,q)$ 與 $\mathrm{KL}(p\|q)$（bit），並以數值驗證 $H(p,q)=H(p)+\mathrm{KL}(p\|q)$。(c) 從定義出發，證明該恆等式對一般離散分布成立。

練習 12.3 (進階)　三類分類問題，真類別為第 1 類。模型 A 輸出 $q^A = (0.7, 0.2, 0.1)$，模型 B 輸出 $q^B = (0.34, 0.33, 0.33)$。(a) 分別求交叉熵損失（nat）。(b) 對 $L = -\ln q_c$ 計算 $\mathrm{d}L/\,\mathrm{d}q_c$，說明為何「給正確類的機率越低，梯度越大」。(c) 設 logits 向量 z 經 softmax 得 $q_j = e^{z_j}/\sum_k e^{z_k}$，證明 $\partial L/\partial z_j = q_j - p_j$，其中 p 為 one-hot 目標。

練習 12.4 (進階)　(a) 設 $p = (0.5, 0.5)$、$q = (0.75, 0.25)$，計算 $\mathrm{KL}(p\|q)$ 與 $\mathrm{KL}(q\|p)$（nat），確認不相等。(b) 取 $r = (1, 0)$，計算 $\mathrm{KL}(p\|r)$ 與 $\mathrm{KL}(r\|p)$，說明支撐不匹配時會發生什麼。(c) 綜合 (a)(b)，各舉一個機器學習場景，說明前向與反向 KL 的實務差異。

練習 12.5 (進階)　(a) 證明鏈式規則 $H(X, Y) = H(X) + H(Y|X)$。(b) 由互資訊的定義證明 $I(X;Y) = H(X) + H(Y) - H(X, Y)$。(c) 推出 $H(X, Y) \le H(X) + H(Y)$ 與 $H(Y|X) \le H(Y)$，並指出等號成立的條件。

練習 12.6 (進階)　X, Y 的聯合分布如下表：

	$Y=0$	$Y=1$
$X=0$	$1/2$	$1/8$
$X=1$	$1/8$	$1/4$

(a) 求邊際分布 p_X 與 p_Y。(b) 求 $H(X)$、$H(Y)$ 與 $H(X, Y)$（bit）。(c) 用 (b) 與練習 12.5 的恆等式求 $I(X;Y)$。(d) 用定義 $I = \sum p \log_2 \frac{p(x,y)}{p_X(x)p_Y(y)}$ 直接計算，驗證兩法一致。

練習 12.7 (進階)　證明：$I(X;Y) = 0$ 若且唯若 X 與 Y 獨立。（提示：把互資訊寫成 KL 散度，動用定理 12.8 的等號條件。）

練習 12.8 (挑戰)　用 Lagrange 乘數法證明：在 K 個結果上，熵在均勻分布處取得唯一的最大值 $\log K$。請說明為何該駐點確為全域最大（凹性），並處理支撐落在邊界（某些 $p_i = 0$）的情形。

練習 12.9 (挑戰)　(a) 證明：對所有 $t > 0$，$\ln t \le t - 1$，等號成立若且唯若 $t = 1$。(b) 由此給出 Gibbs 不等式 $\mathrm{KL}(p\|q) \ge 0$ 的另一個證明（不使用 Jensen 不等式），並完整寫出等號成立的條件。

練習 12.10 (挑戰·程式)　用 NumPy：(a) 實作 `entropy`、`cross_entropy`、`kl` 三個函數（正確處理 $0 \log 0$），在隨機生成的分布對上數值驗證 $H(p, q) = H(p) + \mathrm{KL}(p\|q)$ 與 $\mathrm{KL}(p\|q) \ne \mathrm{KL}(q\|p)$。(b) 固定模型輸出 $q = (0.6, 0.25, 0.1, 0.05)$、真類別為第 1 類，對 label smoothing 參數 $\varepsilon \in \{0, 0.05, 0.1, 0.2\}$ 分別計算交叉熵。(c) 用「$I = \mathrm{KL}(\text{聯合} \,\|\, \text{邊際乘積})$」計算練習 12.6 的互資訊，驗證與手算一致。

12.7 練習題詳解

練習 12.1 詳解　(a) 四個結果等可能：$H = \log_2 4 = 2$ bit。
(b) 逐項計算：

$$H = -0.7\log_2 0.7 - 0.2\log_2 0.2 - 0.1\log_2 0.1 = 0.3602 + 0.4644 + 0.3322 \approx 1.157 \text{ bit}.$$

(c) 三結果均勻分布的熵為 $\log_2 3 \approx 1.585$ bit。$1.157 < 1.585$：偏斜的分布更可預測，熵必然更低——定理 12.3「均勻最大」的內容。
(d) $H_{\text{nat}} = 1.157 \times \ln 2 \approx 0.802$ nat。
(e) $I = -\log_2 \frac{1}{64} = 6$ bit：機率每減半，驚訝度加一位元。
(f) $\text{PPL} = 2^{4.2} \approx 18.4$：每步等效於在 18 個等可能選項中亂猜。單位驗證：$2.9110/\ln 2 \approx 4.20$ bit 一致，且 $e^{2.9110} \approx 18.4 = 2^{4.2}$：底數與單位配對，兩條路同一個答案。

練習 12.2 詳解　(a)

$$H(p) = -\tfrac{1}{4}\log_2 \tfrac{1}{4} - \tfrac{3}{4}\log_2 \tfrac{3}{4} = 0.25 \times 2 + 0.75 \times 0.4150 \approx 0.811 \text{ bit}.$$

(b) 模型均勻，每個結果的碼長都是 $-\log_2 0.5 = 1$ bit，無論真相如何：$H(p,q) = 1$ bit。KL 直接算：

$$\text{KL}(p\|q) = 0.25\log_2\frac{0.25}{0.5} + 0.75\log_2\frac{0.75}{0.5} = -0.25 + 0.75 \times 0.5850 \approx 0.189 \text{ bit}.$$

驗證：$H(p) + \text{KL}(p\|q) = 0.811 + 0.189 = 1.000 = H(p,q)$。✓
(c) 設和取遍 $p(x) > 0$ 的 x（其餘項按約定為零），

$$\text{KL}(p\|q) = \sum_x p(x)\log\frac{p(x)}{q(x)} = \sum_x p(x)\log p(x) - \sum_x p(x)\log q(x) = -H(p) + H(p,q). \qquad \blacksquare$$

練習 12.3 詳解　(a) one-hot 下損失只看正確類的機率：$L_A = -\ln 0.7 \approx 0.357$ nat、$L_B = -\ln 0.34 \approx 1.079$ nat。B 的損失約為 A 的三倍——即使 B「排序」也對了（0.34 是最大分量）：交叉熵看的是機率本身，不是排名。
(b) $\dfrac{\mathrm{d}L}{\mathrm{d}q_c} = -\dfrac{1}{q_c}$：在 $q_c = 0.7$ 處為 -1.43、$q_c = 0.1$ 處為 -10、$q_c \to 0$ 時發散。模型給正確答案的機率越低（錯得越自信），損失曲面越陡、修正越猛。
(c) $L = -\ln q_c = -z_c + \ln\sum_k e^{z_k}$，對任意 j 微分：

$$\frac{\partial L}{\partial z_j} = -\mathbf{1}\{j = c\} + \frac{e^{z_j}}{\sum_k e^{z_k}} = q_j - p_j. \qquad \blacksquare$$

梯度恰是**預測分布減目標分布**：正確類 $q_c - 1 < 0$（logit 抬高）、錯誤類 $q_j > 0$（logit 壓低）。

練習 12.4 詳解　(a) 前向：

$$\mathrm{KL}(p\|q) = 0.5\ln\frac{0.5}{0.75} + 0.5\ln\frac{0.5}{0.25} = 0.5\big(\ln\tfrac{2}{3} + \ln 2\big) = 0.5\ln\tfrac{4}{3} \approx 0.144 \text{ nat}.$$

反向：

$$\mathrm{KL}(q\|p) = 0.75\ln\tfrac{3}{2} - 0.25\ln 2 \approx 0.304 - 0.173 = 0.131 \text{ nat}.$$

$0.144 \neq 0.131$：不對稱。✓

(b) $\mathrm{KL}(p\|r)$：在 $x=2$ 處 $p=0.5>0$ 而 $r=0$，該項為 $0.5\ln(0.5/0)=\infty$，故 $\mathrm{KL}(p\|r)=\infty$。反方向：

$$\mathrm{KL}(r\|p) = 1\times\ln\frac{1}{0.5} + 0 = \ln 2 \approx 0.693 \text{ nat}$$

（第二項按約定 $0\log 0 = 0$）。同一對分布，一個方向無窮、另一方向 0.693 nat——不對稱可以極端到這個程度，根源是支撐不匹配。

(c) 前向 KL（最大概似訓練）對「模型漏掉真相會出現的模式」施以無限重罰，MLE 訓練的生成模型因此傾向覆蓋所有資料模式、甚至過度平滑；反向 KL（變分推論）允許棄守部分眾數，變分近似後驗因此偏窄、低估不確定性（圖 12.2）。

練習 12.5 詳解　(a) 利用 $p(x,y)=p(x)\,p(y|x)$：

$$H(X,Y) = -\sum_{x,y} p(x,y)\big[\log p(x) + \log p(y|x)\big] = -\sum_{x} p(x)\log p(x) - \sum_{x,y} p(x,y)\log p(y|x),$$

即 $H(X)+H(Y|X)$（第一項用了 $\sum_y p(x,y)=p(x)$）。對稱論證給出 $H(Y)+H(X|Y)$。 ■

(b) 展開定義中的對數：

$$I(X;Y) = \sum_{x,y} p(x,y)\big[\log p(x,y) - \log p(x) - \log p(y)\big] = -H(X,Y) + H(X) + H(Y). \qquad \blacksquare$$

(c) 由 (b) 與 $I(X;Y)\geq 0$（定理 12.8 套用於 $\mathrm{KL}(p_{X,Y}\|p_X\otimes p_Y)$）：

$$H(X,Y) = H(X) + H(Y) - I(X;Y) \leq H(X) + H(Y),$$

再以 (a) 代換得 $H(Y|X)\leq H(Y)$。兩式等號皆成立若且唯若 $I(X;Y)=0$，即獨立（練習 12.7）。

練習 12.6 詳解　(a) 列和與行和：

$$p_X = \big(\tfrac{1}{2}+\tfrac{1}{8},\ \tfrac{1}{8}+\tfrac{1}{4}\big) = \big(\tfrac{5}{8},\tfrac{3}{8}\big), \qquad p_Y = \big(\tfrac{5}{8},\tfrac{3}{8}\big).$$

(b) 邊際熵（兩者相同）：

$$H(X) = H(Y) = -\tfrac{5}{8}\log_2\tfrac{5}{8} - \tfrac{3}{8}\log_2\tfrac{3}{8} = 0.625\times 0.6781 + 0.375\times 1.4150 \approx 0.954 \text{ bit}.$$

聯合熵（四格驚訝度為 $1,3,3,2$ bit）：

$$H(X,Y)=\tfrac12(1)+\tfrac18(3)+\tfrac18(3)+\tfrac14(2)=1.75\text{ bit}.$$

(c) $I(X;Y)=H(X)+H(Y)-H(X,Y)=0.954+0.954-1.75\approx0.159$ bit。
(d) 四格的比值 $p(x,y)/\big(p_X(x)p_Y(y)\big)$ 為

$$\frac{1/2}{25/64}=\frac{32}{25},\qquad \frac{1/8}{15/64}=\frac{8}{15}\ (\times2),\qquad \frac{1/4}{9/64}=\frac{16}{9},$$

代入：

$$I=\tfrac12\log_2\tfrac{32}{25}+2\times\tfrac18\log_2\tfrac{8}{15}+\tfrac14\log_2\tfrac{16}{9}\approx0.1781-0.1134-0.1134+0.2075=0.159\text{ bit}.\checkmark$$

練習 12.7 詳解　互資訊即 $I(X;Y)=\mathrm{KL}\big(p_{X,Y}\,\|\,p_X\otimes p_Y\big)$，其中 $p_X\otimes p_Y$ 是合法分布（各格非負，總和 $\big(\sum_x p_X\big)\big(\sum_y p_Y\big)=1$）。
（⇐）若獨立，$p(x,y)=p_X(x)p_Y(y)$ 處處成立，每項的對數為 $\log1=0$，故 $I=0$。
（⇒）若 $I=0$，由定理 12.8（Gibbs）的等號條件，KL 為零僅當兩分布完全相等：

$$p(x,y)=p_X(x)\,p_Y(y)\quad\text{對所有}(x,y),$$

正是獨立的定義。 ■

練習 12.8 詳解　以 nat 計（結論對任何底數成立）。目標：在 $\sum_i p_i=1$、$p_i>0$ 的內部區域最大化 $H(p)=-\sum_i p_i\ln p_i$。
駐點。Lagrange 函數

$$\mathcal{L}(p,\lambda)=-\sum_{i=1}^{K}p_i\ln p_i+\lambda\Big(\sum_{i=1}^{K}p_i-1\Big),\qquad \frac{\partial\mathcal{L}}{\partial p_i}=-\ln p_i-1+\lambda=0\implies p_i=e^{\lambda-1}.$$

右邊與 i 無關：所有 p_i 相等，由約束得 $p_i=1/K$（均勻分布），此時 $H=\ln K$。
全域性（凹性）。$\varphi(t)=-t\ln t$ 的二階導 $\varphi''(t)=-1/t<0$，故每項嚴格凹，H 為凹函數之和；可行域（機率單體）是凸集。凹函數在凸集內部的駐點即全域最大，且由嚴格凹性唯一。
邊界。若某些 $p_i=0$，問題退化為 $K'<K$ 個結果上的同型問題，最大值 $\ln K'<\ln K$，不可能超越內部解。故均勻分布是唯一的最大熵分布，$H_{\max}=\ln K$ nat $=\log_2 K$ bit，與定理 12.3 的 Jensen 證明一致。 ■

練習 12.9 詳解　(a) 令 $g(t)=\ln t-t+1$（$t>0$）。$g'(t)=\frac1t-1$：$t<1$ 時遞增、$t>1$ 時遞減，故 $t=1$ 是唯一極大點且 $g(1)=0$。於是 $\ln t\le t-1$，等號若且唯若 $t=1$。 ■

(b) 設 $S=\{x:p(x)>0\}$。若 S 上某處 $q(x)=0$，則 $\mathrm{KL}(p\|q)=\infty\geq 0$；以下設 $q>0$ 於 S。對每個 $x\in S$ 取 $t=q(x)/p(x)$ 套用 (a)：

$$-\mathrm{KL}(p\|q)=\sum_{x\in S}p(x)\ln\frac{q(x)}{p(x)}\leq\sum_{x\in S}p(x)\Big(\frac{q(x)}{p(x)}-1\Big)=\sum_{x\in S}q(x)-1\leq 0,$$

最後一步因 $\sum_{x\in S}q(x)\leq 1$。故 $\mathrm{KL}(p\|q)\geq 0$。等號需兩處同時取等：(i) 每個 $x\in S$ 上 $q(x)/p(x)=1$；(ii) $\sum_{x\in S}q(x)=1$，即 q 在 S 外無質量——合併得 $q=p$ 處處成立。 ■

練習 12.10 詳解　`entropy`、`cross_entropy`、`kl` 三個函數沿用 12.5 節的實作（該版本已以遮罩正確處理 $0\log 0$），以下列出三個驗證段：

```
import numpy as np

rng = np.random.default_rng(0)

# (a) 隨機分布上驗證恆等式與不對稱性
for _ in range(3):
    p = rng.dirichlet(np.ones(5))
    q = rng.dirichlet(np.ones(5))
    assert np.isclose(cross_entropy(p, q),
                      entropy(p) + kl(p, q))
    print(round(kl(p, q), 4), round(kl(q, p), 4))

# (b) label smoothing：目標混入均勻分布
q_model = np.array([0.60, 0.25, 0.10, 0.05])
K = len(q_model)
for eps in [0.0, 0.05, 0.1, 0.2]:
    target = (1 - eps) * np.eye(K)[0] + eps / K
    print(eps, round(cross_entropy(target, q_model), 4))

# (c) 互資訊 = KL(聯合 || 邊際乘積)
joint = np.array([[0.5, 0.125], [0.125, 0.25]])
px = joint.sum(1, keepdims=True)
py = joint.sum(0, keepdims=True)
print(round(kl(joint.ravel(), (px * py).ravel(), "bit"), 4))
```

(a) 每輪 `assert` 都通過：恆等式在浮點精度內成立；印出的兩欄（前向、反向 KL）每輪皆不同，展示不對稱性。(b) 輸出 $0.5108, 0.5752, 0.6396, 0.7684$ nat：ε 越大、損失越高（$\varepsilon=0.1$ 與例 12.5 的手算一致）——不是變差，是損失地板從 0 抬升到 $H(p^{\mathrm{ls}})$。(c) 輸出 0.1589 bit，與練習 12.6 的手算 0.159 bit 一致。

下一章：校準——讓模型輸出的 0.83，真的意味著 83% 。