信息论基础：从不确定性到分类损失

机器学习 · 深度学习补充

学习目标

- 区分事件的信息量、分布的熵和条件熵，解释二分类熵何时最大
- 区分熵、交叉熵和 KL 散度，说明它们分别评价哪些分布
- 跟随示范解释交叉熵与 KL 的联系，以及真实标签如何决定二分类损失
- 说明模型输出、单样本损失与训练集平均目标之间的关系

1 信息怎样消除不确定性

从消息排除可能性到平均信息，再看已有线索后的剩余未知。

消息怎样帮助我们确定答案?

关于抽出的一张牌的消息	剩余可能的牌	起始数量 / 剩余数量
尚未收到消息	52	1
这张牌是红桃	13	4
这张牌不是黑桃 A	51	52/51

- 设一副 52 张牌中的每张牌等概率被抽到；收到消息前，52 张都有可能
- 同样从 52 张开始，“是红桃”将候选缩小到原来的 1/4，“不是黑桃 A”仍留下 51 张；前者更能帮助确定答案，因此提供的信息更多

连续缩小的比例相乘，信息量相加

从 8 个等可能答案中确定一个，连续三次将候选减半：

- 剩余数量依次为 $8 \to 4 \to 2 \to 1$，缩小倍数相乘为 $2 \times 2 \times 2 = 8$
- 每次都排除当前候选的一半，提供相同的信息；把三次判断合起来，信息量应为一次的三倍

对数将倍数的乘法变为加法：

$$\log_2(2 \times 2 \times 2) = 1 + 1 + 1 = 3$$

- 一次等可能的是 / 否判断提供 1 bit；三次这样的判断提供 3 bit
- 若消息将 N 个等可能选择缩小到 M 个，信息量为 $\log_2(N/M)$
- 红桃消息的概率为 $13/52 = 1/4$，其信息量为 $\log_2 4 = 2$ bit

负对数把事件概率变为信息量

在等可能候选中，消息概率越小，剩余候选占比越小。推广到一般概率分布，越难预料的结果发生时，提供的信息越大；用**信息量**（self-information）计量：

$$I_P(y) = -\log_2 P(y) = \log_2 \frac{1}{P(y)}, P(y) > 0$$

- y 为观察到的结果，$P(y)$ 为事先采用的概率；概率 $1/2,1/4,1/8$ 分别对应 $1,2,3$ bit
- 独立事件的联合概率相乘，信息量相加；这与前面的连续判断保持一致
- 确定事件的信息量为零；概率趋近零时，信息量趋向正无穷
- 使用自然对数 $\ln$ 时单位为 nat：$I_{\text{nat}} = I_{\text{bit}} \ln 2$。单位改变只缩放数值

常见答案应当更早被识别

答案	出现概率	是 / 否回答的编码	编码长度（bit）
A	1/2	0	1
B	1/4	10	2
C	1/8	110	3
D	1/8	111	3

- 依次问“是否 A？” “是否 B？” “是否 C？”，得到“是”即确定答案，三次都为“否”则确定为 D；0 表示是，1 表示否。例如，B 的编码 10 表示“不是 A，是 B”
- 常见的 A 很快被识别，少见的 C、D 需要更多回答；没有一个完整编码是另一个编码的开头；连续读到 0 或 10 等完整编码时，就能确定一个答案，无须另加分隔符

熵对每个结果的信息量求平均

若 8 个答案恰有 4 个 A、2 个 B、1 个 C、1 个 D，总码长为 14 bit，平均每个 1.75 bit。按概率计算同一平均值：

$$\frac{1}{2}\times 1+\frac{1}{4}\times 2+\frac{1}{8}\times 3+\frac{1}{8}\times 3=1.75\ \text{bit}$$

熵（entropy）用同样的方式，对每个结果的信息量按出现概率求平均：

$$H(P)=\sum_y P(y)\,I_P(y)=-\sum_y P(y)\log_2 P(y)$$

- 平均时，每个结果按出现概率计入：A 出现约一半，所以它的信息量权重为 1/2；C 出现约八分之一，权重为 1/8
- 本例码长恰好等于负对数概率，所以平均码长等于熵；一般整数码长未必相等，熵是可唯一解码的二进制编码平均长度的下界
- $P(y)=0$ 的项贡献为零；本段单位为 bit，换为 nat 后熵为 $1.75\ln 2\approx 1.2130$

两类等概率时，平均不确定性最大

一类概率 p	0	0.1	0.5	0.9	1
熵（nat）	0	0.3251	0.6931	0.3251	0

- 表中改用自然对数 ln，单位为 nat；二分类熵为 $-p\ln p-(1-p)\ln(1-p)$。一类几乎确定时，通常容易预料结果；两类各一半时平均不确定性最大，为 $\ln 2$ nat，即 1 bit
- 稀有结果发生时信息量大，但它出现得少，整个来源的平均熵仍可较小；对 K 个结果，等概率分布达到最大熵 $\ln K$ nat

最大熵原则保留已知信息，避免额外偏好

选择概率分布时，先满足已知事实；对尚无依据区分的结果，不凭空给其中某个结果更高概率。

- 只知道有 A、B、C、D 四个结果，且没有其他约束：等概率分布 $(1/4,1/4,1/4,1/4)$ 使熵最大
- 若已知 $P(A) = 1/2$，就必须保留这个事实；B、C、D 没有进一步区别时，剩余一半概率均分为 $(1/2,1/6,1/6,1/6)$

最大熵原则是在满足已知约束的候选分布中，选择熵最大的分布。

- 约束改变，可采用的分布和最大值也会改变；“熵最大”不意味着无条件采用均匀分布
- 该原则用于分布选择；分类训练将依据真实标签评价预测概率

已知一条回答后，还剩多少未知?

已知“是否 A”的回答	剩余答案与条件概率	剩余熵（bit）
是，概率 1/2	只有 A，概率 1	0
否，概率 1/2	B：1/2，C：1/4，D：1/4	1.5

- 仍采用 A、B、C、D 的概率 $(1/2,1/4,1/8,1/8)$；获知“否”后，B 的条件概率为 $(1/4)/(1/2) = 1/2$，C、D 同理
- “否”之后，B 需再问一次，C、D 需再问两次，剩余熵为 $(1/2) \times 1 + (1/4) \times 2 + (1/4) \times 2 = 1.5$ bit。因此平均剩余熵为 $(1/2) \times 0 + (1/2) \times 1.5 = 0.75$ bit

条件熵先在条件内平均，再对线索平均

令 Y 表示答案，X 表示已经知道的线索，$P_X(x)$ 为线索取值 x 的概率。**条件熵**是先算出每条可能线索下答案的熵，再按线索的出现概率求平均；它计量获知线索后平均还需要多少信息才能确定答案。

$$H(Y \mid X) = \sum_{x:P_X(x)>0} P_X(x) \left[-\sum_y P(y \mid x) \log_2 P(y \mid x) \right]$$

- 内层在固定线索 x 下，按 $P(y \mid x)$ 计算答案的熵；外层按 $P_X(x)$ 再平均，包含所有正概率的线索
- 前例从 $H(Y) = 1.75$ bit 降至 $H(Y \mid X) = 0.75$ bit；已知第一问的回答，平均少了 1 bit 未知
- 有限离散变量满足 $0 \le H(Y \mid X) \le H(Y)$。这是平均性质，某个特定条件的熵可以高于原来的熵
- 条件熵是一个量；条件最大熵建模是在约束下选择分布时最大化这一量，含义与当前“已知线索后测量剩余未知”不同

2 概率判断怎样改变预测代价

现实仍按 P 出现；用 Q 判断时，平均多付出了多少代价？

错配判断让常见的 A 需要更多回答

答案	实际概率 P	判断概率 Q	匹配 P 的编码	匹配 Q 的编码
A	1/2	1/8	0	110
B	1/4	1/8	10	111
C	1/8	1/4	110	10
D	1/8	1/2	111	0

- 按 Q 依次问“是否 D?” “是否 C?” “是否 A?”，得到“是”即确定答案，三次都为“否”则确定为 B；仍以 0 表示是、1 表示否。实际最常见的 A 从 1 次回答变为 3 次回答，对应码长从 1 bit 变为 3 bit
- 现实的出现频率没有因判断改变：用 Q 安排识别，评价代价时仍须按 P 求平均

交叉熵按现实频率平均模型的预测代价

前例用 Q 的编码时，A、B、C、D 的长度为 3、3、2、1 bit；平均代价为：

$$\frac{1}{2}\times 3+\frac{1}{4}\times 3+\frac{1}{8}\times 2+\frac{1}{8}\times 1=2.625\ \text{bit}$$

交叉熵（cross-entropy）按实际目标 P 的概率，对判断 Q 的负对数代价求平均：

$$H(P,Q)=-\sum_{y}P(y)\log_2 Q(y)$$

- $P(y)$ 决定结果 y 在平均中占多少权重；$Q(y)$ 越小，$-\log_2 Q(y)$ 越大，该结果的预测代价越高
- 两种评价都按同一个 P 求平均：熵使用 P 给出的负对数，交叉熵使用 Q 给出的负对数。前例分别为 1.75 与 2.625 bit
- 本例负对数恰好等于整数码长，一般表示理想信息代价。改用 $\ln$ 时单位为 nat；若 $P(y)>0$ 而 $Q(y)=0$，交叉熵为正无穷，$P(y)=0$ 的项贡献为零

KL 是概率不匹配增加的平均代价

对同一来源 P 的判断	平均代价（bit）	比采用 P 多出的代价（bit）
采用真实概率 P	1.75	0
采用错配概率 Q	2.625	0.875

- 采用真实概率 P 时，平均仍需 1.75 bit 来识别答案；采用 Q 后变为 2.625 bit，多出的 0.875 bit 来自概率判断的不匹配
- **KL 散度**（Kullback–Leibler divergence）把这一额外部分单独计量：$D_{\mathrm{KL}}(P\|Q) = H(P,Q) - H(P) = 0.875$ bit

逐个结果的代价差，按实际概率求平均

对于同一个结果 y，错配判断与匹配判断的代价差为：

$$-\log_2 Q(y) - [-\log_2 P(y)] = \log_2 \frac{P(y)}{Q(y)}$$

对这些差按现实概率 P 求平均，得到 KL 的公式：

$$D_{\mathrm{KL}}(P\|Q) = \sum_{y:P(y)>0} P(y) \log_2 \frac{P(y)}{Q(y)}$$

- 前例 A、B、C、D 的代价差为 $2, 1, -1, -2$ bit；加权得 $(1/2) \times 2 + (1/4) \times 1 - (1/8) \times 1 - (1/8) \times 2 = 0.875$。单项可为负，KL 整体非负
- 因此 $H(P, Q) = H(P) + D_{\mathrm{KL}}(P\|Q)$，交叉熵由来源自身的熵与错配多出的平均代价组成
- 固定目标 P 时，$H(P)$ 不随模型参数改变；降低交叉熵等价于降低此方向的 KL。KL 为零当且仅当 $P = Q$

平均权重决定 KL 的方向

方向：$P=(0.5,0.5)$、$Q=(0.1,0.9)$	计算（使用 ln）	KL/nat
$D_{\mathrm{KL}}(P\|\|Q)$	$0.5\ln(0.5/0.1)+0.5\ln(0.5/0.9)$	0.5108
$D_{\mathrm{KL}}(Q\|\|P)$	$0.1\ln(0.1/0.5)+0.9\ln(0.9/0.5)$	0.3681

- $D_{\mathrm{KL}}(P\|\|Q)$ 按 P 的频率评价 Q；$D_{\mathrm{KL}}(Q\|\|P)$ 改为按 Q 的频率评价 P。两个方向采用不同的平均权重；KL 一般不对称，不是满足距离公理的距离
- 若 $P(y)>0$ 而 $Q(y)=0$，模型把现实可能结果排除，代价为正无穷；$P(y)=0$ 的项贡献为零，有限值需 $Q(y)>0$ 对所有 $P(y)>0$ 的结果成立

3 从真实标签到分类训练

模型给真实类别的概率决定损失，训练平均评价所有样本。

真实标签选择模型要承担的预测代价

回到二分类：猫的标签为 $y=1$，狗为 $y=0$。模型根据输入 $\boldsymbol{x}$ 产生标量得分 o，sigmoid 函数将它映为猫的概率 $\hat{p}=\sigma(o)=1/(1+\exp(-o))$，狗的概率为 $1-\hat{p}$。

- 图片实际是猫，就评价模型给猫的概率，代价为 $-\ln\hat{p}$
- 图片实际是狗，就评价模型给狗的概率，代价为 $-\ln(1-\hat{p})$

二分类交叉熵（binary cross-entropy，BCE）用一个式子表达两个分支：

$$\ell(y,\hat{p})=-[y\ln\hat{p}+(1-y)\ln(1-\hat{p})]$$

- 标签 $y=1$ 使第一项保留、第二项为零；$y=0$ 时相反，始终选真实类别的概率
- 按类别 0、1 排列，单样本目标为 $P_y=(1-y,y)$，模型分布为 $Q=(1-\hat{p},\hat{p})$；$\ell(y,\hat{p})=H(P_y,Q)$，因此前面的交叉熵在硬标签下成为上述损失
  - 有限得分的 sigmoid 输出满足 $0<\hat{p}<1$，两个标签分支的损失均有定义

自信地预测错误，会承担较大的损失

模型预测猫的概率 $\hat{p}$	图片实际为猫，$y=1$	图片实际为狗，$y=0$
0.9	0.1054	2.3026
0.5	0.6931	0.6931
0.1	2.3026	0.1054

- 实际为猫时，0.9 给真实类高概率，损失小；0.1 给真实类低概率，损失大。两个标签列都遵循同一规则：模型给真实类别的概率越低，损失越大
- 硬标签目标 $H(P_y)=0$，单样本交叉熵等于对应 KL；这只描述该样本的监督目标，不说明总体条件熵为零

训练平均评价所有样本的预测代价

若模型对两张图片都给出猫的概率 0.9，而它们的真实标签分别为猫、狗，则损失分别为 0.1054、2.3026，平均约为 1.2040。

- 只看第一张预测会觉得很好；第二张图的错误也须进入评价
- 给定 N 个训练样本，样本平均损失是参数 θ 的函数：

$$L(\theta) = \frac{1}{N}\sum_{i=1}^{N} \ell(y_i, \hat{p}_i), \hat{p}_i = Q_\theta(1 \mid \boldsymbol{x}_i)$$

- 数据与标签固定；改变参数会改变各输入的得分、概率和损失，训练寻找使这个平均目标降低的参数
- 训练平均损失下降说明当前数据上的目标改善，测试表现仍需独立评价

总体目标匹配给定输入后的标签分布

- 单样本目标 P_y 将全部权重放在这一次观察到的标签；总体条件分布 $P(\cdot \mid x)$ 描述固定输入 x 下各标签的真实概率，$\cdot$ 表示遍历标签取值
  - 若该输入下两类等概率，真实标签分布的熵为 $\ln 2$ nat；一次观察的硬标签目标熵仍为零
- 总体交叉熵按输入与标签的真实概率，平均模型 Q_θ 的负对数代价：

$$\mathcal{C}(\theta) = \mathbb{E}_{(X,Y)\sim P}[-\ln Q_\theta(Y \mid X)]$$

- $\mathbb{E}$ 表示总体平均；$L(\theta)$ 则对观察到的训练样本损失求平均，用来估计 $\mathcal{C}(\theta)$
- 固定输入 x 时，交叉熵分解为标签分布的熵与相应 KL；再对输入 X 求平均：

$$\mathcal{C}(\theta) = H(Y \mid X) + \mathbb{E}_X D_{\mathrm{KL}}(P(\cdot \mid X) \| Q_\theta(\cdot \mid X))$$

- 各输入的真实标签熵平均为 $H(Y \mid X)$；相应 KL 平均为右侧期望，表示模型增加的平均预测代价
- 固定总体分布且上述量有限时，条件熵项不随参数变化；降低总体交叉熵等价于降低平均条件 KL

本讲总结

- 信息量计量收到消息后对答案增加的确定性；熵按结果的出现概率平均信息量，二分类在两类等概率时达到最大值
- 条件熵对给定输入后的标签不确定性求平均；最大熵原则是在已知约束下选择分布
- 交叉熵用现实的出现频率平均模型的预测代价；KL 单独计量概率不匹配增加的平均代价。$H(P,Q) = H(P) + D_{\mathrm{KL}}(P\|Q)$
  - 目标固定且交叉熵有限时，降低交叉熵等价于降低所示方向的 KL；KL 非负、一般不对称。模型对目标正概率结果也给出正概率时，才有有限代价
- 二分类硬标签选择真实类别的负对数概率；训练把单样本损失汇入参数的平均目标
  - 硬标签目标熵为零不说明总体条件熵为零，训练目标改善仍需与测试评价区分