整理资料:

1. 邱锡鹏.神经网络与深度学习. https://nndl.github.io/ .
2. 信息论（1）——熵、互信息、相对熵 - 徐光宁的文章 - 知乎 https://zhuanlan.zhihu.com/p/36192699
3. 信息论（3）——联合熵，条件熵，熵的性质 - 徐光宁的文章 - 知乎 https://zhuanlan.zhihu.com/p/36385989
4. 为什么交叉熵（cross-entropy）可以用于计算代价？ - 微调的回答 - 知乎 https://www.zhihu.com/question/65288314/answer/244557337
5. 通俗理解条件熵 - 忆臻的文章 - 知乎 https://zhuanlan.zhihu.com/p/26551798

熵

自信息和熵

在信息论中，**熵**用来衡量一个随机事件的不确定性。假设对一个随机变量 X (取值集合为 $\mathcal{X}$，概率分布为 $p(x), x \in \mathcal{X}$) 进行编码，**自信息** (Self Information) $I(x)$ 是变量 $X = x$ 的**信息量**或**编码长度**，定义为

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

那么随机变量 X 的平均编码长度，即**熵**定义为

$$\begin{aligned} H(x) &= \mathbb{E}_X[I(x)] \\ &= \mathbb{E}_X[-\,log\,p(x)] \\ &= -\sum_{x \in \mathcal{X}} p(x)\,log\,p(x) \end{aligned}$$

其中，当 $p(x_i) = 0$ 时，我们定义 $0\,log\,0 = 0$，这与极限一致，$lim_{p \to 0+}\,p\,log\,p = 0$。

熵是一个随机变量的平均编码长度，即自信息的数学期望。熵越高，则随机变量的信息越多；熵越低，则信息越少。如果变量 X 当且仅当在 x 时 $p(x) = 1$，则熵为 0。也就是说，对于一个确定的信息，其熵为 0，信息量也为 0。如果其概率分布为一个均匀分布，则熵最大。

联合熵和条件熵

对于两个离散随机变量 X 和 Y，假设 X 取值集合为 $\mathcal{X}$；Y 取值集合为 $\mathcal{Y}$，其联合概率分布为 $p(x, y)$，则

X 和 Y 的**联合熵** (Joint Entropy) 为

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

X 和 Y 的**条件熵** (Conditional Entropy) 为

$$\begin{aligned} H(X|Y) &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y)\,log\,p(x|y) \\ &= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y)\,log\,\frac{p(x, y)}{p(y)} \end{aligned}$$

根据其定义，条件熵也可以写为

$$\begin{aligned}H(X|Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, \frac{p(x,y)}{p(y)} \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, p(x,y) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, p(y) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, p(x,y) + \sum_{y\in\mathcal{Y}} p(y)\, log\, p(y) \\ &= H(X,Y) - H(Y)\end{aligned}$$

补充：

条件熵 $H(Y|X)$ 表示在已知随机变量 X 的条件下随机变量 Y 的不确定性。其定义为 X 给定的条件下，Y 的条件概率分布的熵对 X 的数学期望：

$$\begin{aligned}H(Y|X) &= \sum_{x\in\mathcal{X}} p(x) H(Y|X=x) \\ &= -\sum_{x\in\mathcal{X}} p(x) \sum_{y\in\mathcal{Y}} p(y|x) \log p(y|x) \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y) \log p(y|x)\end{aligned}$$

我们可以理解为，条件熵的意思是，按变量 X 的每个取值对变量 Y 进行分类，然后在每个小类中计算变量 Y 的小熵，再对每个小熵乘以各个类别的概率，最后求和。

互信息

互信息 (Mutual Information) 是衡量已知一个变量时，另一个变量不确定性的减少程度。两个离散随机变量 X 和 Y 和互信息定义为

$$I(X;Y) = \sum_{x\in\mathcal{X}}\sum_{x\in\mathcal{Y}} p(x,y)\, log\, \frac{p(x,y)}{p(x)\, p(y)}$$

我们可以发现

$$\begin{aligned}I(X;Y) &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, \frac{p(x,y)}{p(x)\, p(y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, \frac{p(x|y)\, p(y)}{p(x)\, p(y)} \\ &= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, \frac{p(x|y)}{p(x)} \\ &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, p(x) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, p(x|y) \\ &= -\sum_{x\in\mathcal{X}} p(x)\, log\, p(x) + \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\, log\, p(x|y) \\ &= H(X) - H(X|Y)\end{aligned}$$

所以，互信息的一个性质为

$$\begin{aligned}I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X)\end{aligned}$$

如果 X 和 Y 相互独立，即 X 不对 Y 提供任何信息，反之亦然，因此它们的互信息为零。决策树内信息增益 (Information Gain) 的定义其实就是互信息。

交叉熵和散度

交叉熵

对应分布为 $p(x)$ 的随机变量，熵 $H(p)$ 表示其最优编码长度。**交叉熵** (Cross Entropy) 是按照概率分布 q 的最优编码对真实分布为 p 的信息进行编码的长度，定义为

$$\begin{aligned} H(p,q) &= \mathbb{E}_p[-log\, q(x)] \\ &= -\sum_x p(x)\, log\, q(x) \end{aligned}$$

在给定 p 的情况下，如果 q 和 p 越接近，交叉熵越小；如果 q 和 p 越远，交叉熵就越大。

KL 散度

KL散度 (Kullback-Leibler Divergence)，也叫**KL距离**或**相对熵** (Relative Entropy)，是用概率分布 q 来近似 p 时所造成的信息损失量。KL散度是按照概率分布 q 的最优编码对真实分布为 p 的信息进行编码，其平均编码长度 $H(p,q)$ 和 p 的最优平均编码长度 $H(p)$ 之间的差异。对于离散概率分布 p 和 q，从 q 到 p 的KL散度定义为

$$\begin{aligned} D_{KL}(p||q) &= H(p,q) - H(p) \\ &= \sum_x p(x) \log \frac{p(x)}{q(x)} \end{aligned}$$

其中，为了保证连续性，定义 $0 \log \frac{0}{0} = 0$，$0 \log \frac{0}{q} = 0$。

KL散度可以是衡量两个概率分布之间的距离。KL散度总是非负的，$D_{KL}(p||q) \geq 0$。只有当 $p = q$ 时，$D_{KL}(p||q) = 0$。如果两个分布越接近，KL散度越小；如果两个分布越远，KL散度就越大。<u>但KL散度并不是一个真正的度量或距离，一是KL散度不满足距离的对称性，二是KL散度不满足距离的三角不等式性质。</u>

通过KL散度，我们再看互信息可以发现：

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)\, p(y)} = D_{KL}(p(x,y)||p(x)\, p(y))$$

由此，<u>我们知道互信息本质上是描述了联合分布 $p(x,y)$，与两个边缘分布之积 $p(x)\, p(y)$ 的差异程度</u>。如果差异为 0，表示两个随机变量独立。

补充

熵的性质：最值性

即当随机变量 X 服从均匀分布时，熵取得最大值。我们可以利用KL散度的非负性进行证明。

不妨设 $u(x) = \frac{1}{|\mathcal{X}|}$ ，其中 $|\mathcal{X}|$ 表示为随机变量的取值集合的势（即集合的元素个数），则对于任意的 $p(x)$，它们的KL散度

$$
\begin{aligned}
D_{KL}(p||u) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} \\
&= \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log u(x) \\
&= \sum_{x\in\mathcal{X}} p(x)\log p(x) - \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{|\mathcal{X}|} \\
&= \log|\mathcal{X}|\sum_{x\in\mathcal{X}} p(x) - \left(-\sum_{x\in\mathcal{X}} p(x)\log p(x)\right) \\
&= \log|\mathcal{X}| - H(x) \geq 0
\end{aligned}
$$

于是，$H(x) \leq \log|\mathcal{X}|$。当随机变量 X 服从均匀分布时，等号成立。

为什么交叉熵可以用于计算代价？

原因：KL散度可以被用于计算代价，而<u>在特定情况下最小化KL散度等价于最小化交叉熵。而交叉熵的运算更简单，所以用交叉熵来当作代价</u>。当然，不同领域有不同解释，更传统的机器学习的说法是似然函数的最大化就是交叉熵。

- **机器如何 “学习”？**

  机器学习的过程就是希望在训练数据上**模型学到的分布** $P(model)$ 和**真实数据的分布** $P(real)$ 越接近越好。怎么最小化两个分布之间的不同呢？**用默认的方法，使其KL散度最小。**

  但我们没有真实数据的分布，那么只能退而求其次， 希望**模型学到的分布**和**训练数据的分布** $P(training)$ 尽量相同，也就是把训练数据当作模型和真实数据之间的代理人。

  假设训练数据是从总体独立同分布采样而来，那么我们可以利用最小化训练数据的经验误差来降低模型的泛化误差。简单说：

  - 最终目的是希望学到的模型的分布与真实分布一致：$P(model) \simeq P(real)$
  - 但真实分布是不可知的，我们只好假设训练数据是从真实数据中独立同分布采样而来：$P(training) \simeq P(real)$
  - 退而求其次，我们希望学到的模型分布至少和训练数据的分布一致 $P(model) \simeq P(training)$

  由此，非常理想化的看法是，如果**模型（左）**能够学到**训练数据（中）**的分布，那么应该近似学到了**真实数据（右）**的分布：$P(model) \simeq P(training) \simeq P(real)$

- **为什么交叉熵可以用作代价？**

  最小化模型分布 $P(model)$ 与训练数据上的分布 $P(training)$ 的差异等价于最小化这两个分布间的KL散度，也就是最小化 $D_{KL}(P(training)||P(model))$。

  我们知道，$D_{KL}(p||q) = H(p,q) - H(p)$，那么

  $$D_{KL}(P(training)||P(model)) = H(P(training), P(model)) - H(P(training))$$

  因为训练数据的分布是给定的，所以当 $H(P(training))$ 固定不变时，最小化KL散度等价于最小化交叉熵。因此，交叉熵可以用于计算“学习模型的分布”与“训练数据分布”之间的不同。当交叉熵最低时 (等于训练数据分布的熵)，我们学到了“最好的模型”。

  但是，完美学到了训练数据分布往往意味着过拟合，因为训练数据不等于真实数据，我们只是假设它们是相似的，而一般还要假设存在一个高斯分布的误差，是模型的泛化误差下限。