CS4641B Machine Learning

Focus video: Entropy

Rodrigo Borela ▸ rborelav@gatech.edu

Entropy

- **Information:** Let X be a random variable with distribution $p(x) = p(X = x)$. Information is measured as:

$$h(x) = -\log_2 p(x)$$

- Average amount of information to encode a random variable X with respect to its distribution $p(x)$ is the entropy $H(x)$:

$$H(x) = E[h(x)] = \sum_x h(x)p(x) = -\sum_x p(x)\log_2 p(x)$$

Example

- X and Y are random variables
- $N = total\ number\ of\ trials$
- $n_{ij} = number\ of\ occurrence$

X = Throw a die

Y = Flip a coin

Joint probability distribution

X = dice roll

Y = coin flip

Y = coin flip	$x_{i=1} = 1$	$x_{i=2} = 2$	$x_{i=3} = 3$	$x_{i=4} = 4$	$x_{i=5} = 5$	$x_{i=6} = 6$	c_j
$y_{j=1} = head$	$n_{ij} = 3$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 5$	$n_{ij} = 1$	$n_{ij} = 5$	20
$y_{j=2} = tail$	$n_{ij} = 2$	$n_{ij} = 2$	$n_{ij} = 4$	$n_{ij} = 2$	$n_{ij} = 4$	$n_{ij} = 1$	15
c_i	5	6	6	7	5	6	$N = 35$

Joint probability distribution

X = dice roll

Y = coin flip

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$x = 6$	
$y = head$	$\frac{3}{35}$	$\frac{4}{35}$	$\frac{2}{35}$	$\frac{5}{35}$	$\frac{1}{35}$	$\frac{5}{35}$	$\frac{20}{35}$
$y = tail$	$\frac{2}{35}$	$\frac{2}{35}$	$\frac{4}{35}$	$\frac{2}{35}$	$\frac{4}{35}$	$\frac{1}{35}$	$\frac{15}{35}$
	$\frac{5}{35}$	$\frac{6}{35}$	$\frac{6}{35}$	$\frac{7}{35}$	$\frac{5}{35}$	$\frac{6}{35}$	$\frac{35}{35}$

Joint probability distribution $p(X = x, Y = y)$

Y = coin flip

X = dice roll

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$x = 6$	
$y = head$	0.09	0.11	0.06	0.14	0.03	0.14	0.57
$y = tail$	0.06	0.06	0.11	0.06	0.11	0.03	0.43
	0.14	0.17	0.17	0.20	0.14	0.17	1.0

$p(y = head)$ (0.57)

$p(Y = tail, X = 3)$ (0.11)

$p(X = 6)$ (0.17)

To get the conditional probabilities, we can use the product rule:

$$p(Y = tail, X = 3) = p(Y = tail|X = 3)p(X = 3)$$

$$p(Y = tail|X = 3) = \frac{p(Y = tail, X = 3)}{p(X = 3)} = \frac{0.11}{0.17} = 0.67$$

Joint Entropy

Coin flip

- $H(Y) = H\big(p(head), p(tail)\big) = H(0.57, 0.43) = 0.57 \times \log_2 \frac{1}{0.57} + 0.43 \times \log_2 \frac{1}{0.43} = 0.985\ bits$

Dice roll

- $H(X) = H(p(1), p(2), p(3), p(4), p(5), p(6)) = 0.14 \times \log_2 \frac{1}{0.14} + \cdots + 0.17 \times \log_2 \frac{1}{0.17} = 2.562\ bits$

Dice roll and coin flip

- $H(X, Y) = \sum_{x,y} p(X = x, Y = y) \cdot \log_2 \frac{1}{P(X=x,\, Y=y)}$
- $H(p(head, 1), p(head, 2), \ldots, p(tail, 6)) = 0.09 \times \log_2 \frac{1}{0.09} + 0.11 \times \log_2 \frac{1}{0.11} + \cdots + 0.03 \times \log_2 \frac{1}{0.03} =$ $3.435\ bits$

Conditional Entropy

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2 p(y|x)$$

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

Conditional probability distribution $p(Y = y|X = x)$

X = dice roll

Y = coin flip

	$x = 1$	$x = 2$	$x = 3$	$x = 4$	$x = 5$	$x = 6$
$y = head$	0.60	0.67	0.33	0.71	0.20	0.83
$y = tail$	0.40	0.33	0.67	0.28	0.80	0.17
	1.00	1.00	1.00	1.00	1.00	1.00

$p(Y = head|X = 6)$

$p(Y = tail|X = 3)$

$$H(Y|\, X = 1) = H\big(p(head|1), p(tail|1)\big) = H(0.60, 0.40) = 0.4 \times \log\frac{1}{0.4} + 0.6 \times \log\frac{1}{0.6} = 0.971\ bits$$

...

$$H(Y|\, X = 6) = H\big(p(head|6), p(tail|6)\big) = 0.83 \times \log\frac{1}{0.83} + 0.17 \times \log\frac{1}{0.17} = 0.658\ bits$$

$$H(Y|X) = \sum_x P(X = x) \cdot H(Y|X = x) = 0.14 \times 0.971 + \cdots + 0.17 \times 0.722 = \cdots$$

$p(X = 1)$ $H(Y|X = 1)$ $p(X = 6)$ $H(Y|X = 6)$

Mutual Information

- Mutual information: quantify the reduction in uncertainty in Y after seeing feature X

$$I(X,Y) = H(Y) - H(Y|X)$$

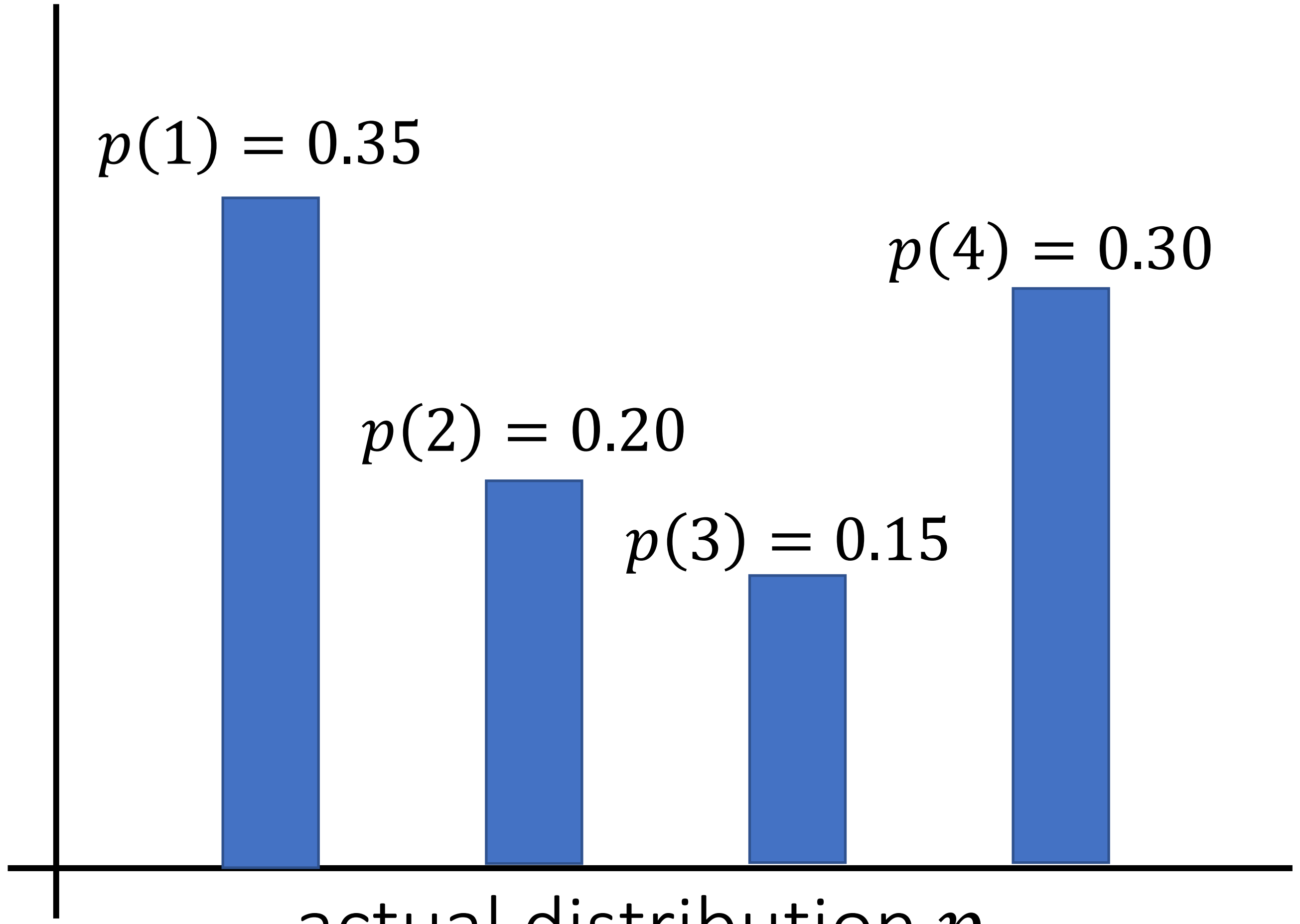
$$I(X,Y) = H(X) - H(X|Y)$$

(exercise to the reader)

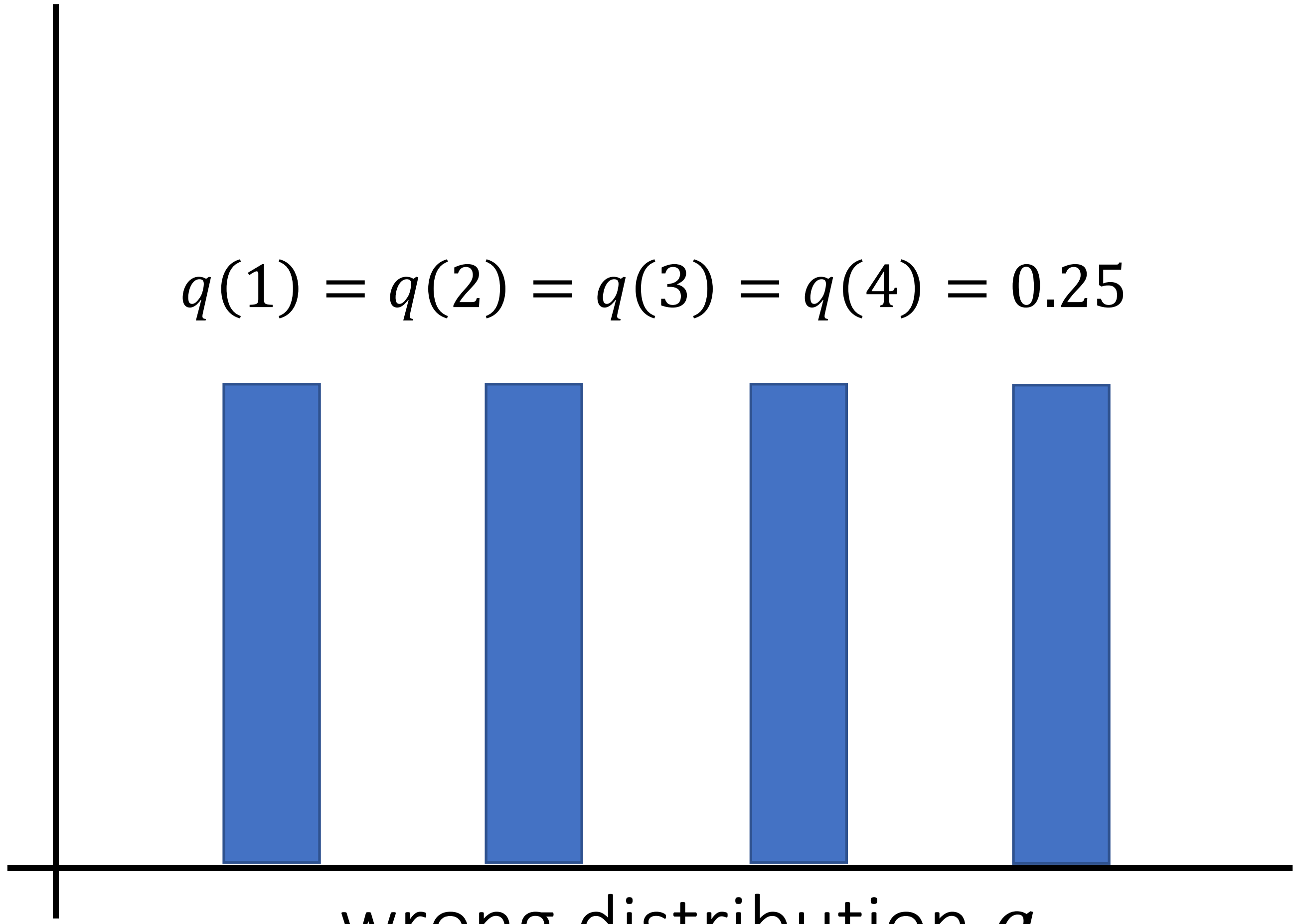
Cross Entropy

- Cross Entropy: The expected number of bits when a wrong distribution q is assumed while the data actually follows a distribution p

$$H(p,q) = -\sum_{x \in X} p(x) \log_2 q(x) = H(p) + KL[p][q]$$



actual distribution p



wrong distribution q

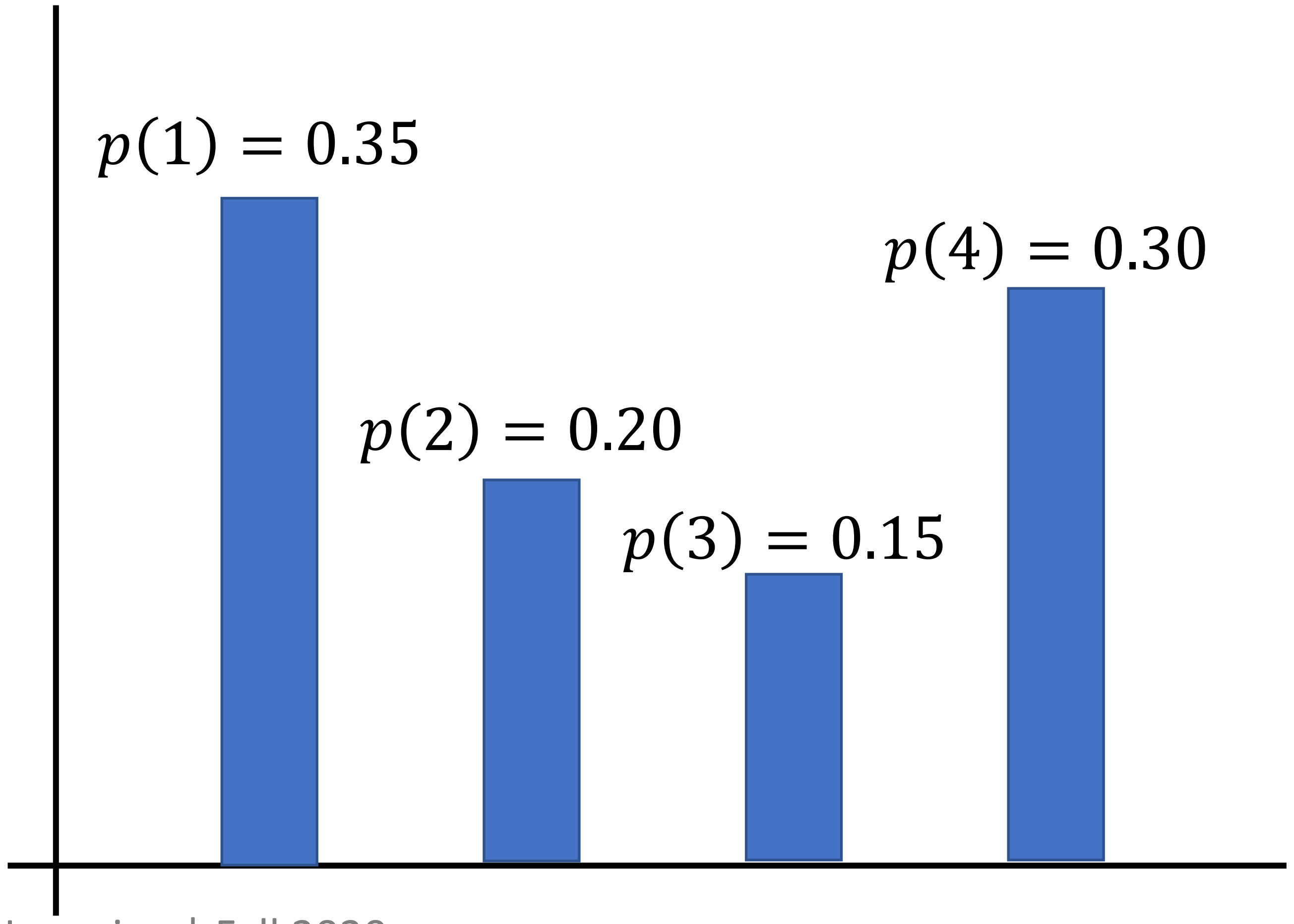
Cross Entropy

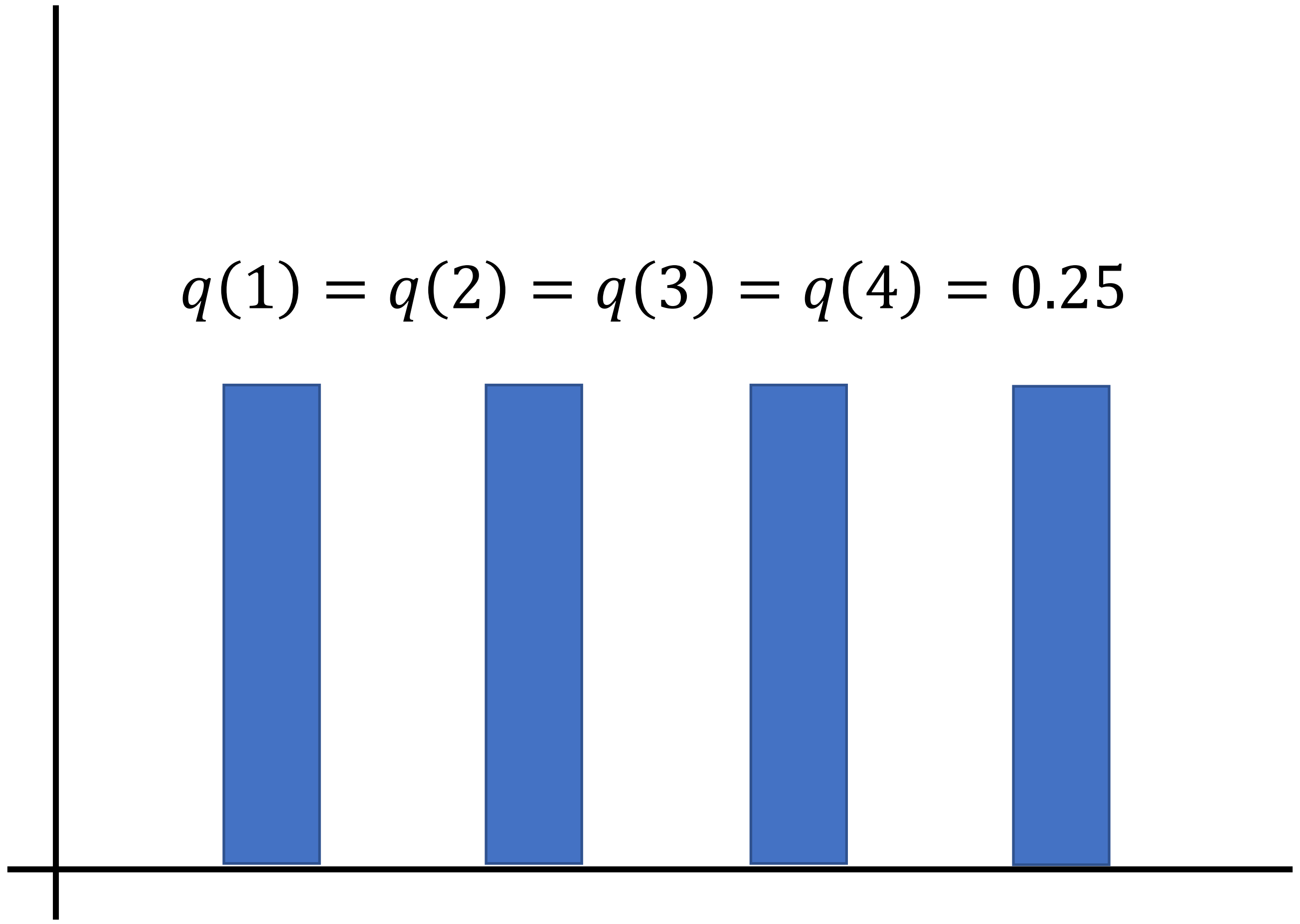
$$H(p,q) = -(0.35 \times \log_2 0.25 + \boxed{0.20} \times \log_2 \boxed{0.25} + \cdots + 0.30 \times \log_2 0.25) = 2.0\ bits$$

$p(2)$ $q(2)$

$$H(p) = -(\boxed{0.35} \times \log_2 \boxed{0.35} + 0.20 \times \log_2 0.20 + \cdots + 0.30 \times \log_2 0.30) = 1.926\ bits$$

$p(1)$ $p(1)$





Kullback-Leibler Divergence

- Another useful information theoretic quantity measures the difference between two distributions
- $KL[p][q] = \sum_{x_i} p(x_i) \log \frac{p(x_i)}{q(x_i)} = \sum_{x_i} \underbrace{p(x_i) \log \frac{1}{q(x)}}_{\text{Cross-entropy}} - H(p) = H(p, q) - H(p)$

- Excess cost in bits paid by encoding according to q instead of p

$$KL[p][q] = H(p, q) - H(p) = 2.0 - 1.926 = 0.074\ bits$$