Homework 2

Brandon Amaral, Monte Davityan, Nicholas Lombardo, Hongkai Lu

September 9, 2022

1 Chapter 6: Information Theory Summary

1.1 Entropy

Entropy is a measure of uncertainty for some probability distribution of a random variable.

If some information is sampled from a distribution with high entropy, then prediction of that sample will be difficult (ie. uncertainty will be high). If entropy of that distribution is at its lowest of 0, then every sample will be similar.

1.1.1 Entropy for Discrete Random Variable

Entropy of a random variable is:

$$H(X) = -\sum_{k=1}^{K} p(X=k)\log_2 p(X=k) = -E_X[\log p(x)]$$

The log base 2 is when the units are bits and log base e is when the units are nats. The **maximum entropy** discrete distribution is acheived for the uniform distribution when $p(x=k) = \frac{1}{K}$ where $H(X) = \log K$.

For binary random variables, $H(X) = -[\theta \log_2 \theta + (1-\theta)\log_2(1-\theta)]$ which is the binary entropy function

1.1.2 Cross Entropy

Cross entropy between two distributions p and q is

$$H(p,q) = -\sum_{k=1}^{K} p_k \log q_k$$

. **Shannon's source coding theorem** is when $q = p$, cross entropy is minimized such that $H(p,p) = H(p)$

1.1.3 Joint Entropy

Joint entropy of two random variables X and Y is

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

.

If X and Y are independent, $H(X,Y) = H(X) + H(Y)$.

If Y is a deterministic function of X, then $H(X,Y) = H(X)$ which implies that $H(X,Y) \geq max[H(X), H(Y)] \geq 0$. The intuition behind this means that entropy does not decrease when more information is added.

1.1.4 Conditional Entropy

Conditional entropy of Y given X is the uncertainty in Y after seeing X, averaged over all possible values of X:

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

If Y is deterministic on X, then $H(Y|X) = 0$ since knowing X results in Y so there is no uncertainty. In general, $H(Y|X) \leq H(Y)$, and if X and Y are independent, seeing X doesn't effect Y so $H(Y|X) = H(Y)$. Intuitively, on average if we first see data (condition on data) it never increases the uncertainty. The general **chain rule for entropy** is $H(X_1, X_2, ..., X_n) = \sum_{i=1}^{n} H(X_i|X_1, ..., X_{i-1}$

1.1.5 Perplexity

Perplexity can be seen as the measure of predictability and is perplexity$(p) = 2^{H(p)}$. If a distribution can perfectly predict outcomes it will have a perplexity of 1 (because the entropy will be 0 so uncertainty will be 0).

For an empirical distribution from data D, $p_D(x|D) = \frac{1}{N}\sum_{n=1}^{N} \delta_{x_n}(x)$, the perplexity (measure of predictability) is perplexity$(p_D, p) = 2^{H(p_D,p)}$

This is commonly used in statistical language modeling where the cross entropy is $H = -\frac{1}{N}\sum_{n=1}^{N} \log p(x_n)$ with perplexity$(p_D, p) = 2^H = 2^{-\frac{1}{N}\log(\prod_{n=1}^{N} p(x_n))} = \sqrt[N]{\prod_{n=1}^{N} \frac{1}{p(x_n)}}$ which is known as **exponentiated cross entropy**

1.1.6 Differential Entropy for Continuous Random Variable

If X is a continuous random variable, the **differential entropy** is $h(X) = -\int_X p(x) \log p(x) dx$

Discretization: In general computing that integral is difficult so an approximation can be found by discretizing the variables. One method is to bin the distribution into quantiles with one heuristic for the number of bins being: $B = N_D^{1/3} \frac{max(D) - min(D)}{3.5\sigma(D)}$ which doesn't scale well to multi- dimensionalities.

1.2 Relative Entropy (KL Divergence

1.2.1 Definition and Interpretation

If we ever want to find how similiar two distributions are we can use the Kullback- Leibler divergence (KL divergence) (also called information gain and relative entropy) among other measures.

For discrete: $D_{KL}(p||q) = \sum_{k=1}^{K} p_k \log \frac{p_k}{q_k}$ For continuous: $D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$

An interpretation of this is "KL divergence as the extra number of bits you need to pay when compressing data samples if you use the incorrect distribution q as the basis of your coding scheme compared to the true distribution p"

1.2.2 Non- Negativity of KL

KL divergence is always non- negative.

The information inequality theorem states that $D_{KL}(p||q) \geq 0$ with equality if and only if $p = q$.

A corollary is the uniform distribution maximizes the entropy: $H(X) \leq \log |X|$ where $|X|$ is the number of states for X, with equality if and only if $p(x)$ is uniform

1.2.3 KL Divergence and MLE

If we want to find the distribution q that has minimal KL divergence to p so that q and p are as close as possible, then we seek:

$$q^* = \arg\min_q D_{KL}(p||q)$$

If we let p be the empirical distribution based on our data,

$$p_D(x) = \frac{1}{N_D} \sum_{n=1}^{N_D} \delta(x - x_n)$$

We can apply the sifting property of delta functions to get

$$D_{KL}(p_D||q) = -\frac{1}{N_D} \sum_n \log q(x_n) + C$$

Where C is the cross entropy objective, independent of q. "Thus we see that **minimizing KL divergence to the empirical distribution is equivalent to maximizing likelihood.**"

1.2.4 Forward vs Reverse KL

For the task of approximating a distribution p by using a simpler distribution q, we can minimize $D_{KL}(q||p)$ or $D_{KL}(p||q)$. We get two different behaviors depending on which we choose.

First is **forwards KL**, also called **inclusive KL**, defined by

$$D_{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$$

Minimizing this with respect to q is called **M-projection** or **moment projection**. Here, we must avoid any instance where $p(x) > 0$ but $q(x) = 0$ because $\log \frac{p(x)}{q(x)}$ will be infinite. Thus, we must force q to include all areas for which p has non-zero probability. We say q will be **zero-avoiding** or **mode-covering**, and q will tend to overestimate the support of p.

Second is **reverse KL**, also called **exclusive KL**, defined by

$$D_{KL}(q||p) = \int q(x) \log \frac{q(x)}{p(x)} dx$$

Minimizing this with respect to q is known as **I-projection** or **information projection**. Here, we avoid instances where $p(x) = 0$ but $q(x) > 0$ since $\log \frac{q(x)}{p(x)}$ will be infinite. Thus, we must force q to exclude all areas of space for which p has zero probability. We say will be **zero-forcing** or **mode-seeking**, and q will tend to underestimate the support of p.

1.3 Mutual Information

1.3.1 Definition and Interpretation

Mutual information comes from wanting to measure how dependent two random variables are on each other. It is defined as

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

This is always non- negative, even for continuous random variables, and when X and Y are independent, the mutual information is 0.

One interpretation of this is it is the information gained when we update the model from when the two variables are treated as independent, to when they are dependent. Another formulation is:

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

The interpretation here is the reduction in uncertainty of X when observing Y (and vice versa). This result incidentally gives us another proof that, on average, conditioning reduces entropy. Also, since $0 \leq I(X; Y) = H(X) - H(X|Y)$, we know that $H(X|Y) \leq H(X)$.

Another way to express this is $I(X; Y) = H(X, Y) - H(X|Y) - H(Y|X)$ and $I(X; Y) = H(X) + H(Y) - H(X, Y)$

1.3.2 Conditional Mutual Information

Conditional mutual information is $I(X;Y|Z) = E[I(X,Y)|Z] = I(Y;X,Z) - I(Y;Z)$. This can be generalized to N variables called the chain rule for mutual information:

$$I(Z_1, ..., Z_n; X) = \sum_{n=1}^{N} I(Z_n; X|Z_1, ..., Z_{n-1})$$

1.3.3 Mutual Information as a Generalized Correlation Coefficient

To generalize mutual information coefficient, we can suppose (x, y) are jointly Gaussian as follow:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}(0, \begin{pmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{pmatrix})$$

Then, we have

$$I(X,Y) = \frac{1}{2} log[1 - \rho^2]$$

There are three special cases when ρ take different values:

Case 1: If $\rho = 1$, $X = Y$ and $I(X,Y) = \infty$. Then, we will expect infinite amount of information of X when observing Y.

Case 2: If $\rho = 0$, X and Y are independent, and $I(X,Y) = \infty$. Then, we will expect nothing of X when observing Y.

Case 3: If $\rho = -1$, $X = -Y$ and $I(X,Y) = \infty$. Then, we will able to predict X to infinite precision when observing Y.

1.3.4 Normalized Mutual Information

The normalized mutual information is defined as follows:

$$NMI(X,Y) = \frac{I(X;Y)}{min(H(X), H(Y))} \leq 1$$

The range of the normalized mutual information is 0 to 1.

We have below Corollaries when NMI(X,Y) = 0 or 1:

(a): If $NMI(X,Y) = 0$, X and Y are independent since $I(X;Y) = 0$.

(b): If $NMI(X,Y) = 1$, X is a deterministic function of Y since $I(X;Y) = 1$ and knowing X gives us complete knowledge of Y.

1.3.5 Maximal Information Coefficient

The maximal information coefficient is defined as follows:

$$MIC(X,Y) = \max_{G} \frac{I((X,Y)|G)}{log||G||}$$

where G is the set of 2d grids; $(X,Y)|G$ is the represent of a discretization of the variables onto the grids. Also, $||G||$ is $min(G_x; G_y)$, where G_x is the number of grid cells in the x direction, and G_y is the number of grid cells in the y direction. The denominator is equivalent to the entropy of a uniform joint distribution; this ensures $0 \leq MIC \leq 1$. The **characteristic matrix** $M(k,l)$ is the maximum MI achievable by any grid of size (k,l), and it is normalized by log(min(k,l). Here, the MIC is the maximum entry of the matrix.

MIC is a measure of the strength of the linear or non-linear relationship between two variables X and Y; as such it is far more versatile than a correlation coefficient. An MIC of 0 indicates no relationship between the variables of any form; an MIC of 1 represents a noise-free relationship in any form, including exceptionally non-linear and even non-functional (multiple outputs per input) relationships. Murphy describes MIC as "a correlation coefficient for the 21st century."

1.3.6 Data Processing Inequality

Intuitively, when we create a new variable Z under processing the noisy observations of Y under an unknown variable X, we are not going to increase the amount of information for X. That is known as the data processing inequality.

According to Murphy, we have below formally theorem:

Theorem 6.3.1. Suppose $X \to Y \to X$ forms a Markov chin, so that $X \perp Z|Y$. Then $I(X;Y) \geq I(X;Z)$

1.3.7 Sufficient Statistics

An important result of Data Processing Inequality arises when we consider a chain $\theta \to D \to s(D)$, where

$$I(\theta; s(D)) \leq I(\theta; D)$$

If equality does hold, then we call $s(D)$ a sufficient statistic of the data D for inferring θ. Here we can reconstruct the data D just by knowing $s(D)$ without needing to know θ itself. A useful subcategory is the **minimal sufficient statistic** which does not add any extra information about θ, and thus $s(D)$ "maximally compresses" the data D without losing relevant predictive information.

1.3.8 Fano's Inequality

The Fano's inequality says that if we estimate the class label Y by observing the features X and applying a Markov chain $Y \to X \to \hat{Y}$, then we have

$$P_e \geq \frac{H(Y|X-1)}{\log |y|}$$

where $P_e = P(Y \neq \hat{Y}), H(Y|X)$ is the conditional entropy of Y given X.

2 Problem 6.1

Let X and Y be two random variables X and Y. The mutual information between X and Y is defined as follows:

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

Then, we have

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

By (6.1) to (6.15), we have $H(Y|X) = H(X,Y) - H(X)$ (A), so we can get $H(X|Y) = H(X,Y) - H(Y)$ (B) by switching X and Y. Then, we replace (A) and (B) in the above formula:

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

If we replace (A) in the above formula:

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

Then we have

$$H(X,Y) = H(Y|X) + H(X|Y) + I(X;Y).$$

3 Problem 6.2

Let p and q be probability distribution functions, and suppose $p(x) \approx q(x)$, i.e. $p(x) = \Delta(x) + q(x)$, for some $\Delta(x)$ small. Then, we have

$$\frac{p(x)}{q(x)} = 1 + \frac{\Delta(x)}{q(x)}.$$

Now, we have that

$$D_{KL}(p||q) = \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) = \sum_x (\Delta(x) + q(x)) \log\left(1 + \frac{\Delta(x)}{q(x)}\right)$$

Then, using the Taylor expansion of $\log(1+x)$, we find that

$$\sum_x (\Delta(x) + q(x)) \log\left(1 + \frac{\Delta(x)}{q(x)}\right) = \sum_x (\Delta(x) + q(x)) \left(\frac{\Delta(x)}{q(x)} - \frac{(\Delta(x))^2}{2(q(x))^2} + \frac{(\Delta(x))^3}{3(q(x))^3} - \cdots (-1)^{n+1} \frac{(\Delta(x))^n}{n(q(x))^n} + \cdots\right).$$

Multiplying this out, we have

$$\begin{aligned} &\sum_x (\Delta(x) + q(x)) \left(\frac{\Delta(x)}{q(x)} - \frac{(\Delta(x))^2}{2(q(x))^2} + \frac{(\Delta(x))^3}{3(q(x))^3} - \cdots\right) \\ =&\sum_x \Delta(x) + \frac{(\Delta(x))^2}{2q(x)} - \frac{(\Delta(x))^3}{6(q(x))^2} + \frac{(\Delta(x))^4}{12(q(x))^3} - \cdots + (-1)^{n+1} \left(\frac{1}{n} - \frac{1}{n+1}\right) \frac{(\Delta(x))^{n+1}}{(q(x))^n} + \cdots. \end{aligned}$$

We see that

$$\begin{aligned} D_{KL}(p||q) &= \sum_x p(x) \log\left(\frac{p(x)}{q(x)}\right) \\ &= \sum_x \Delta(x) + \frac{(\Delta(x))^2}{2q(x)} - \frac{(\Delta(x))^3}{6(q(x))^2} + \frac{(\Delta(x))^4}{12(q(x))^3} - \cdots + (-1)^{n+1} \left(\frac{1}{n} - \frac{1}{n+1}\right) \frac{(\Delta(x))^{n+1}}{(q(x))^n} + \cdots. \end{aligned}$$

Now, notice that $\sum_x \Delta(x) = \sum_x p(x) - \sum_x q(x)$. Since p and q are probability distributions, we have $\sum_x p(x) = 1$ and $\sum_x q(x) = 1$, so we have $\sum_x \Delta(x) = 0$, i.e.

$$\begin{aligned} D_{KL}(p||q) &= \sum_x \frac{(\Delta(x))^2}{2q(x)} - \frac{(\Delta(x))^3}{6(q(x))^2} + \frac{(\Delta(x))^4}{12(q(x))^3} - \cdots + (-1)^{n+1} \left(\frac{1}{n} - \frac{1}{n+1}\right) \frac{(\Delta(x))^{n+1}}{(q(x))^n} + \cdots \\ &= \sum_x \sum_{n=0}^{\infty} (-1)^n \left(\frac{1}{n+1} - \frac{1}{n+2}\right) \frac{(\Delta(x))^{n+2}}{(q(x))^{n+1}}. \end{aligned}$$

Then, by the alternating series test, since the absolute value of the terms decrease monotonically and the limit of the terms tends to 0, the inner sum over n converges. Furthermore, by the alternating series estimation theorem, $\sum_x \frac{(\Delta(x))^2}{2q(x)}$ approximates $D_{KL}(p||q)$ with error less than $\frac{\Delta(x)^3}{6q(x)^2}$, i.e.

$$D_{KL}(p||q) \approx \sum_x \frac{(\Delta(x))^2}{2q(x)} = \sum_x \frac{(p(x)-q(x))^2}{q(x)} = \frac{1}{2}\chi^2.$$

4 Problem 6.4

Consider a factored approximation $q(x,y) = q(x)q(y)$ to a joint distribution $p(x,y)$. First we show that to minimize the forwards KL $D_{KL}(p||q)$, we should set $q(x) = p(x)$ and $q(y) = p(y)$. The forwards KL, $D_{KL}(p||q)$ is given by

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

by the rules for logarithms and since $q(x,y) = q(x)q(y)$. Then, since $p(x,y) \geq 0$ and $q(x) \geq 0$, we can switch the order of summation on the last term, so that we have

$$\sum_x \sum_y p(x,y)\log\left(p(x,y)\right) - \sum_x \sum_y p(x,y)\log(q(x)) - \sum_y \sum_x p(x,y)\log(q(y)).$$

Now, factoring out both $\log(q(x))$ and $\log(q(y))$ from their respective inner summations, we have

$$\sum_x \sum_y p(x,y)\log\left(p(x,y)\right) - \sum_x \log(q(x)) \sum_y p(x,y) - \sum_y \log(q(y)) \sum_x p(x,y).$$

Then, we notice that $p(x) = \sum_y p(x,y)$ and $p(y) = \sum_x p(x,y)$, so that we have

$$\sum_x \sum_y p(x,y)\log\left(p(x,y)\right) - \sum_x p(x)\log(q(x)) - \sum_y p(y)\log(q(y)).$$

The last two terms are the cross entropies between the marginal distributions of p and q, which are minimized exactly when the marginal distributions are equal to each other, so to minimize the forwards KL $D_{KL}(p||q)$, we set $q(x) = p(x)$ and $q(y) = p(y)$.

Now, we consider the following joint distribution, where the rows represent y and the columns x.

	1	2	3	4
1	1/8	1/8	0	0
2	1/8	1/8	0	0
3	0	0	1/4	0
4	0	0	0	1/4

to show that the reverse KL $D_{KL}(q||p)$ for this p has three distinct minima.

Let q be a separable distribution as before, where $q(x,y) = q(x)q(y)$. Minimizing the reverse KL will "force q to exclude all the areas of space for which p has zero probability." Notice that when x=4, the only non-zero value for p(x,y) is when y=4. Also notice the same for x=3 and y=3. Finally, notice that when x=1 or x=2, the only non-zero values for p(x,y) are y=1 or y=2. So, the reverse KL $D_{KL}(q||p)$ for the above distribution p has three distinct minima centered around these three modes:

1. $q_1(x) = q_1(y) = (0,0,0,1)$
2. $q_2(x) = q_2(y) = (0,0,1,0)$

3. $q_3(x) = q_3(y) = (1/2, 1/2, 0, 0)$

Now, we evaluate $D_{KL}(q||p)$ for each of the three minima:

1. $D_{KL}(q_1||p_1) = \sum_X \sum_Y q_1(x)q_1(y) \log_2 \frac{q_1(x)q_1(y)}{p(x,y)} = 0 + 0 + ... + 0 + (1)(1) \log_2 \frac{(1)(1)}{1/4} = 2$

2. $D_{KL}(q_2||p_2) = \sum_X \sum_Y q_2(x)q_2(y) \log_2 \frac{q_2(x)q_2(y)}{p(x,y)} = 0 + 0 + ... + 0 + (1)(1) \log_2 \frac{(1)(1)}{1/4} + 0 + ... + 0 = 2$

3. $D_{KL}(q_3||p_3) = \sum_X \sum_Y q_3(x)q_3(y) \log_2 \frac{q_3(x)q_3(y)}{p(x,y)} = (1/2)(1/2) \log_2 \frac{(1/2)(1/2)}{1/8} + (1/2)(1/2) \log_2 \frac{(1/2)(1/2)}{1/8} + 0 + 0 + (1/2)(1/2) \log_2 \frac{(1/2)(1/2)}{1/8} + (1/2)(1/2) \log_2 \frac{(1/2)(1/2)}{1/8} + 0 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 1$

On the other hand, if we set $q(x, y) = p(x)p(y)$, the value of $D_{KL}(q||p)$ is infinite. This is because setting $q(x, y) = p(x)p(y)$ fails to "force zeros," or avoid instances where both $q(x, y) > 0$ and $p(x, y) = 0$, and so $\log \frac{q(x,y)}{p(x,y)} = \infty$ several times in our summation, like when $x = 4$ and $y = 1$, for example. This demonstrates that setting $q(x, y)$ to the marginal distributions over x and y can be a good approximation for $q(x, y)$ in mode-covering Forwards KL but not in mode-seeking Reverse KL.