Information Theory with Applications, Math6397
Lecture Notes from September 09, 2014

taken by Tianxiao Jiang

Warm-up

Relative entropy for binary random variables

Let X, Y have alphabet $\mathbb{A} = \{0, 1\}$ and $\mathbb{P}_X(0) = p, \mathbb{Q}_Y(0) = q$, we abbreviate

$$d(p, q) = D(\mathbb{P}_X \| \mathbb{Q}_Y)$$

then with $h(p) = p \ln p + (1-p)\ln(1-p)$, we have

$$\begin{aligned} d(p,q) &= p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q} \\ &= -h(p) - p \ln q - (1-p) \ln (1-q) \\ &= \underbrace{-h(p)}_{\textit{binary entropy of } X} + \underbrace{s(p,q)}_{\textit{linearization of binary entropy of } X \textit{ at } q} \end{aligned}$$

1.5 Mutual information

1.5.1 Definition. Given two discrete random variables X and Y with alphabets $\mathbb{A}$ and $\mathbb{B}$, we define the *mutual information* to be

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

1.5.2 Remarks. From the entropy inequality, $I(X;Y) \geq 0$. Recall from additivity

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

So we get the symmetric expression

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

Moreover, we see that it can be rewritten as a divergence

$$\begin{aligned}
I(X;Y) &= -\sum_{a\in\mathbb{A}} \mathbb{P}_X(a)\ln\mathbb{P}_X(a) - \sum_{b\in\mathbb{B}} \mathbb{P}_Y(b)\ln\mathbb{P}_Y(b) + \sum_{(a,b)\in\mathbb{A}\times\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\ln\mathbb{P}_{X,Y}(a,b) \\
&= -\sum_{(a,b)\in\mathbb{A}\times\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\ln\mathbb{P}_X(a) - \sum_{(a,b)\in\mathbb{A}\times\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\ln\mathbb{P}_Y(b) + \sum_{(a,b)\in\mathbb{A}\times\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\ln\mathbb{P}_{X,Y}(a,b) \\
&= \sum_{(a,b)\in\mathbb{A}\times\mathbb{B}} \mathbb{P}_{X,Y}(a,b)\ln\frac{\mathbb{P}_{X,Y}(a,b)}{\mathbb{P}_X(a)\mathbb{P}_Y(b)} \\
&= D(\mathbb{P}_{X,Y}\|\mathbb{P}_X\mathbb{P}_Y)
\end{aligned}$$

1.6 Conditional mutual information

1.6.1 Definition. The mutual information between X and Y given the outcome $Z = c$ is

$$I(X;Y|Z=c) = \sum_{(a,b)\in\mathbb{A}\times\mathbb{B}} \mathbb{W}_{X,Y}(a,b|c)\ln\frac{\mathbb{W}_{X,Y}(a,b|c)}{\mathbb{V}_X(a|c)\mathbb{U}_Y(b|c)}$$

and the (averaged) conditional mutual information

$$I(X;Y|Z) = \sum_{c\in\mathbb{C}} \mathbb{P}_{\mathbb{Z}}(a) I(X;Y|Z=c)$$

Here, $\mathbb{W}_{X,Y}$ is the conditional probability for the joint distribution of X and Y given $Z = c$. $\mathbb{V}_X$ is conditional probability for X, $\mathbb{U}_Y$ is the conditional probability for Y and $\mathbb{P}_Z$ is the probability measure induced by Z .

1.7 Additivity of conditional mutual information

1.7.1 Theorem. *Given X, Y, Z as above, then $I(X;Y;Z) = I(X;Z) + I(X;Y|Z)$*

Proof. Use the defining equation with averaging

$$\begin{aligned}
I(X;Y|Z) &= \sum_{(a,b,c)\in\mathbb{A}\times\mathbb{B}\times\mathbb{C}} \mathbb{P}_{X,Y,Z}(a,b,c)\ln\frac{\mathbb{W}_{X,Y}(a,b|c)}{\mathbb{V}_X(a|c)\mathbb{U}_Y(b|c)} \\
&= H(X|Z) + H(Y|Z) - H(X,Y|Z)
\end{aligned}$$

by additivity we have $H(X,Y|Z) = H(Y|Z) + H(X|Y,Z)$, from

$$H(X,Y|Z=c) = H(Y|Z=c) + H(X|Y,Z=c)$$

and averaging over outcomes for Z.

Inserting this expression for $H(X,Y|Z)$, we have

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

□

1.7.2 Corollary.

$$\begin{aligned} I(X;Y_1,Y_2,\cdots,Y_n) = I(X;Y_1) &+ I(X;Y_2|Y_1) \\ &+ I(X;Y_3|Y_2,Y_1) \\ &+ \cdots \\ &+ I(X;Y_n|Y_1,\cdots,Y_{n-1}) \end{aligned}$$

1.8 Inequalities for mutual information

We have upper and lower bounds.

1.8.1 Theorem. *With X, Y as before,*

$$0 \leq I(X;Y) \leq \min\{H(X), H(Y)\}$$

and equality on LHS holds iff *one determines the other with probability one.*

Proof. LHS inequality follows from

$$I(X;Y) = D(\mathbb{P}_{X,Y} \| \mathbb{P}_X \mathbb{P}_Y) \geq 0$$

also cases of equality.
RHS inequality follows from

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \longleftarrow \textit{(due to the symmetric property of X and Y)} \end{aligned}$$

with $H(X|Y) \geq 0, H(Y|X) \geq 0$
as well as cases of equality □

1.8.2 Moral. Mutual information measures how much X, Y determines each other.

1.8.3 Example. Match two kinds of modalities of data by maximizing the mutual information.

1.9 Mutual information and Markov chains

Let $\{X_j\}_{j=1}^n$ be a Markov chain, *i.e.* for $x \in \mathbb{A}^n$

$$\mathbb{P}_{X_1,X_2,\cdots,X_n}(x) = \mathbb{P}_{X_1}(x_1)\mathbb{M}_1(x_2|x_1)\cdots\mathbb{M}_{n-1}(x_n|x_{n-1})$$

with conditional probability measures $\mathbb{M}_j(\bullet|x)$ for transition from state $x \in \mathbb{A}$ in *j-th* step.
Markov chains are characterized by the property that for all *j*, $\{X_1, X_2 \cdots X_{j-1}, X_{j+1}\}$ is independent given X_j, *i.e.*

$$\mathbb{W}_{X_1,X_2\cdots X_{j-1},X_{j+1}}(x_1,x_2\cdots x_{j-1},x_{j+1}|x_j) = \mathbb{V}_{X_1,X_2\cdots X_{j-1}}(x_1,x_2\cdots x_{j-1}|x_j)\underbrace{\mathbb{U}_{X_{j+1}}(x_{j+1}|x_j)}_{\mathbb{M}_j(x_{j+1}|x_j)}$$

1.9.1 Theorem. *A sequence of random variables* $\{X_j\}_{j=1}^n$ *is a Markov chain iff for all* $j \in \{2, 3 \cdots n-1\}$, $I(X_1, X_2 \cdots X_{j-1}; X_{j+1}|X_j) = 0$.

We need to show that equality in inequality $I(X;Z|Y) \geq 0$ holds *iff* X, Z are independent given Y.
This is because

$$0 \leq I(X;Z|Y=b) = \sum_{(a,c)\in\mathbb{A}\times\mathbb{C}} \mathbb{W}_{X,Z}(a,c|b) \ln \frac{\mathbb{W}_{X,Z}(a,c|b)}{\mathbb{V}_X(a|b)\mathbb{U}_Z(c|b)}$$

so if $I(X;Z|Y) = 0$, terms must vanish for each b, which means

$$\mathbb{W}_{X,Z}(a,c|b) = \mathbb{V}_X(a|b)\mathbb{U}_Z(c|b)$$

independence of X, Z gives outcomes of Y.
Proving the converse is straightforward.