Entropy

Definition of Entropy: Let $\mathbf{X}$ be a discrete random vector with density $f(\mathbf{x}_i) = P(\mathbf{X} = \mathbf{x}_i)$. The entropy of $\mathbf{X}$ is defined as

$$H(\mathbf{X}) = -\sum f(\mathbf{x}_i) \log (f(\mathbf{x}_i)) = -E\{\log (f(\mathbf{X}))\} \tag{1}$$

Since $0 < f(\mathbf{x}_i) < 1$, it is clear that $H(\mathbf{X}) \geq 0$. Moreover, entropy can be interpreted as a measure of randomness or uncertainty. For example, if $\mathbf{X}$ can take only two values $\mathbf{a}$ and $\mathbf{b}$ with probabilities p and $(1-p)$, respectively. Then

$$H(\mathbf{X}) = -p \log(p) - (1-p) \log(1-p)$$

Differentiating with respect to p

$$-\log(p) - 1 + \log(1-p) + 1 = 0$$

$$\log\left(\frac{1-p}{p}\right) = 0$$

$$\frac{1-p}{p} = 1 \Rightarrow p = 1/2.$$

The most uncertain case is when $\mathbf{a}$ and $\mathbf{b}$ are equally likely. Notice that $H(\mathbf{X})$ doesn't depend on the particular values of $\mathbf{X}$, but in their probabilities.

Exercise 1: Suppose that $\mathbf{X}$ can take n possible values with probabilities $p_1, p_2, ..., p_n$. Show that in this case $H(\mathbf{X})$ is maximized when $p_i = 1/n$.

Definition of Differential Entropy: Let $\mathbf{X}$ be a continuous random vector with density $f(\mathbf{x})$. The differential entropy of $\mathbf{X}$ is defined as

$$H(\mathbf{X}) = -E\{\log(f(\mathbf{X}))\} = -\int \cdots \int f(\mathbf{x}) \log(f(\mathbf{x})) \, \mathbf{dx}$$

The definitions of entropy and differential entropy in terms of expected values are identical. But the behavior of H in the discrete and continuous cases are rather different.

1. **Entropy can be negative in the continuous case.**

   Notice that in the continuous case we no longer have $0 < f(\mathbf{x}) < 1$ and $H(\mathbf{X})$ can be negative. For example, if X is $\text{Unif}(0, a)$, $a > 0$, then

   $$H(X) = -\frac{1}{a}\int_0^a \log\left(\frac{1}{a}\right) dx = -\log\left(\frac{1}{a}\right) = \log(a) < 0 \text{ for all } 0 < a < 1.$$

   Notice that $H(X)$ increases with a and $H(X) \to -\infty$ when $a \to 0$.

2. **$H(\mathbf{X})$ is invariant under one-to-one transformations in the discrete case but not in the continuous case.**

   In fact, let

$$\mathbf{Y} = \mathbf{g}(\mathbf{X}) = \begin{pmatrix} g_1(\mathbf{X}) \\ g_2(\mathbf{X}) \\ \vdots \\ g_d(\mathbf{X}) \end{pmatrix} \tag{2}$$

   be a one-to-one transformation. Let

$$\mathbf{X} = \mathbf{g}^{-1}(\mathbf{Y}) = \mathbf{h}(\mathbf{Y}) = \begin{pmatrix} h_1(\mathbf{X}) \\ h_2(\mathbf{X}) \\ \vdots \\ h_d(\mathbf{X}) \end{pmatrix}$$

   If $\mathbf{X}$ is a discrete random vector with

$$p_i = P(\mathbf{X} = \mathbf{x}_i)$$

   then

$$\begin{aligned} H(\mathbf{Y}) &= -\sum P(\mathbf{Y} = \mathbf{y}_i) \log [P(\mathbf{Y} = \mathbf{y}_i)] \\ &= -\sum P(\mathbf{g}(\mathbf{X}) = \mathbf{g}(\mathbf{x_i})) \log [P(\mathbf{g}(\mathbf{X}) = \mathbf{g}(\mathbf{x_i}))] \\ &= -\sum P(\mathbf{X} = \mathbf{x}_i) \log [P(\mathbf{X} = \mathbf{x}_i)] \\ &= H(\mathbf{X}) \end{aligned}$$

On the other hand, if X is Unif$(0,1)$ and $Y = aX$, then

$$H(X) = \log(1) = 0 \quad \text{and} \quad H(Y) = \log(a)$$

Exercise 2: if $\mathbf{X}$ is a continuous random vector and $\mathbf{Y} = \mathbf{M}\mathbf{X}$, where $\mathbf{M}$ is an invertible constant matrix, then

$$H(\mathbf{Y}) = H(\mathbf{X}) + \log|\det \mathbf{M}|.$$

Exercise 3: Derive and analyze the entropy of the following random variables: (a) Binomial(n, p); (b) Negative Binomial (m, p); (c) Poisson(λ); (d) N$\left(\mu, \sigma^2\right)$; Gamma$(k, \lambda)$.

Definition of Mutual Information:

mutual information is a measure of the amount of information that the entries of a random vector have about each other. Mutual information is defined as follows:

$$D(\mathbf{X}) = \sum_{i=1}^{d} H(X_i) - H(\mathbf{X}) \tag{3}$$

If the entries $X_1, X_2, ..., X_d$ of the random vector $\mathbf{X}$ are independent then

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i)$$

$$\log(f_{\mathbf{X}}(\mathbf{x})) = \sum_{i=1}^{d} \log[f_i(x_i)]$$

$$H(\mathbf{X}) = \int \cdots \int \sum_{i=1}^{d} \log[f_i(x_i)] \left[\prod_{i=1}^{d} f_i(x_i)\right] dx_1 ... dx_d$$

$$= \sum_{i=1}^{d} \int \log[f_i(x_i)] f_i(x_i) \, dx_i$$

$$= \sum_{i=1}^{d} H(X_i)$$

Therefore, the mutual entropy (3) is the difference between the entropy that we would have if the entries of $\mathbf{X}$ were independent and the entropy of the actual joint distribution of $\mathbf{X}$. Intuition indicates that $D(\mathbf{X}) \geq 0$. The following discussion shows that this is case.

Exercise 4:

Let $\mathbf{X}$ be bivariate normal with means $\mathbf{m}$ and covariance matrix

$$\mathbf{V} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}.$$

Calculate the mutual information of $\mathbf{X}$.

Definition of Kullback-Leibler Distance: the Kullback–Leibler distance between two (multivariate) density distributions $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ is defined as follows

$$\delta(f_1, f_2) = E_{f_1}\left\{\log\left(\frac{f_1(\mathbf{X})}{f_2(\mathbf{X})}\right)\right\} = \int \cdots \int f_1(\mathbf{x}) \log\left(\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}\right) \mathbf{dx}$$

By Jensen's inequality

$$\begin{aligned} E_{f_1}\left\{\log\left(\frac{f_1(\mathbf{X})}{f_2(\mathbf{X})}\right)\right\} &= -E_{f_1}\left\{\log\left(\frac{f_2(\mathbf{X})}{f_1(\mathbf{X})}\right)\right\} \\ &\geq -\log\left[E_{f_1}\left\{\left(\frac{f_2(\mathbf{X})}{f_1(\mathbf{X})}\right)\right\}\right] \\ &= -\log\left[\int \cdots \int f_2(\mathbf{x})\, \mathbf{dx}\right] = -\log(1) = 0 \end{aligned}$$

Finally, we notice that the **mutual information** is simply the Kullback–Leibler distance between

$$f_1(\mathbf{x}) = f(\mathbf{x}) \quad \text{and} \quad f_2(\mathbf{x}) = \prod_{i=1}^{d} f_i(x_i).$$

In fact,

$$\delta(f_1, f_2) = \int \cdots \int f(\mathbf{x}) \log\left[\frac{f(\mathbf{x})}{\prod_{i=1}^{d} f_i(x_i)}\right] dx_1 ... dx_d$$

$$= \int \cdots \int f(\mathbf{x}) \log[f(\mathbf{x})]\, dx_1 ... dx_d - \int \cdots \int f(\mathbf{x}) \log\left[\prod_{i=1}^{d} f_i(x_i)\right] dx_1 ... dx_d$$

$$= -H(\mathbf{x}) - \sum_{i=1}^{n} \overbrace{\int f_i(x_i) \log[f_i(x_i)]\, dx_i}^{H(X_i)}$$

$$= \sum_{i=1}^{n} H(X_i) - H(\mathbf{x}).$$

Exercise 5: Let $\mathbf{X}_1$ and $\mathbf{X}_2$ be multivariate normal random vectors with means $\mathbf{m}_1$ and $\mathbf{m}_2$ and covariances $\mathbf{V}_1$ and $\mathbf{V}_2$, respectively. Calculate the Kullback-Leibler distance between $\mathbf{X}_1$ and $\mathbf{X}_2$.

A General Setting for the EM Algorithm

Recall that

$$\underbrace{\mathbf{Z}}_{\text{complete data}} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} = \begin{pmatrix} \text{inclomplete data} \\ \text{augmented data} \end{pmatrix}$$

Then

$$\mathbf{Y} = \mathbf{T}_0 \left(\mathbf{Z} \right)$$

where the function $\mathbf{T}_0$ is the projection on the first k coordinates.

More generally, consider the case where

$$\mathbf{Y} = \mathbf{t}_0 \left(\mathbf{Z} \right),$$

where $\mathbf{t}_0$ is a function, $\mathbf{t}_0 : R^p \rightarrow R^k$, with $k < p$. For example we could have

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \\ Z_5 \end{pmatrix}$$

and

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 + Z_3 \\ Z_4 + Z_5 \end{pmatrix} = \mathbf{t}_0 \left(\mathbf{Z} \right)$$

Discrete Case

In this case we have

$$l\left(\theta|\mathbf{y}\right) = \log f_{\mathbf{Y}}\left(\mathbf{y};\theta\right) \qquad \text{Incomplete log-likelihood}$$

$$l\left(\theta|\mathbf{z}\right) = \log f_{\mathbf{Z}}\left(\mathbf{z};\theta\right) \qquad \text{Complete log-likelihood}$$

$$\tilde{l}\left(\theta|\mathbf{y},\theta^{(k)}\right) = E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log f_{\mathbf{Z}}\left(\mathbf{z};\theta\right)\right\} \qquad \text{E–Step}$$

$$= \sum \cdots \sum \log\left[f_{\mathbf{Z}}\left(\mathbf{z};\theta\right)\right] h_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left(\mathbf{z};\theta^{(k)}\right)$$

$$= \sum \cdots \sum \log\left[f_{\mathbf{Z}}\left(\mathbf{z};\theta\right)\right] \frac{f\left(\mathbf{z};\theta^{(k)}\right)}{f\left(\mathbf{y};\theta^{(k)}\right)}$$

$$h_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left(\mathbf{z};\theta^{(k)}\right) = \frac{f_Z\left(\mathbf{z};\theta^{(k)}\right)}{f_Y\left(\mathbf{y};\theta^{(k)}\right)}$$

$$\tilde{l}\left(\theta|\mathbf{y},\theta^{(k)}\right)=\sum\cdots\sum\log\left[f_{\mathbf{Z}}\left(\mathbf{z};\theta\right)\right]h_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left(\mathbf{z};\theta^{(k)}\right)$$

$$=\sum\cdots\sum\log\left[f_{\mathbf{Z}}\left(\mathbf{z};\theta\right)\right]\frac{f\left(\mathbf{z};\theta^{(k)}\right)}{f\left(\mathbf{y};\theta^{(k)}\right)}$$

Example. Let

$$\mathbf{Z}=\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \end{pmatrix}$$

where Z_1, Z_2, Z_3 are independent, $Z_1 \sim$Poisson(λ), $Z_2 \sim$Poisson(λ) and $Z_3 \sim$Poisson(δ).

Suppose that we observe

$$\mathbf{Y}=\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}=\begin{pmatrix} Z_1+Z_2 \\ Z_2+Z_3 \end{pmatrix}\sim\begin{pmatrix} \text{Poisson}(2\lambda) \\ \text{Poisson}(\lambda+\delta) \end{pmatrix}$$

Notice that Y_1 and $Y_2$are not independent. We can use the EM algorithm to find the MLE for λ and δ.

The complete data log-likelihood (for n independent observations) is

$$l\left(\lambda,\delta|\mathbf{Z}_1,\mathbf{Z}_2,\mathbf{Z}_3\right)=K-2\lambda+\log\left(\lambda\right)\frac{1}{n}\sum\left(Z_{1i}+Z_{2i}\right)-\delta+\log\left(\delta\right)\frac{1}{n}\sum Z_{3i}$$

and

$$\tilde{l}\left(\lambda, \delta | \mathbf{Y}_1, \mathbf{Y}_2, \lambda^{(k)}, \delta^{(k)}\right) = K - 2\lambda + \log(\lambda)\left(\frac{1}{n}\sum Y_{1i}\right) - \delta$$

$$+ \frac{\delta^{(k)}}{\lambda^{(k)} + \delta^{(k)}}\left(\frac{1}{n}\sum Y_{2i}\right)\log(\delta)$$

$$= K - 2\lambda + \log(\lambda)\overline{Y}_1 - \delta + \frac{\delta^{(k)}}{\lambda^{(k)} + \delta^{(k)}}\overline{Y}_2 \log(\delta)$$

NOTE: We used the fact that $Z_{3i}|Y_{2i} \sim Bin\left(Y_{2i}, \delta/\left(\lambda + \delta\right)\right)$ and so $E\left(Z_{3i}|Y_{2i}\right) = Y_{2i}\delta/\left(\lambda + \delta\right)$.

For the M-step we differentiate $\tilde{l}$ with respect to λ and δ and set the derivatives equal to zero:

$$-2 + \frac{\overline{Y}_1}{\lambda} = 0 \Longrightarrow \hat{\lambda}^{(k+1)} = \frac{\overline{Y}_1}{2}$$

$$-1 + \frac{\overline{Y}_2\delta^{(k)}}{\lambda^{(k)} + \delta^{(k)}}\frac{1}{\delta} = 0 \Longrightarrow \hat{\delta}^{(k+1)} = \frac{\overline{Y}_2\delta^{(k)}}{\lambda^{(k)} + \delta^{(k)}}$$

Setting

$$\frac{\overline{Y}_2\delta}{\overline{Y}_1/2 + \delta} = \delta \Longrightarrow \hat{\delta} = \overline{Y}_2 - \frac{\overline{Y}_1}{2}$$

The ML problem has the unique solution $\hat{\lambda} = \overline{Y}_1/2$, $\hat{\delta} = \overline{Y}_2 - \overline{Y}_1/2$, provided $\overline{Y}_2 > \overline{Y}_1/2 > 0$.

Continuous Case

In this case we must complete the transformation by adding another transformation

$$\mathbf{X} = \mathbf{t}_1(\mathbf{Z}) \qquad \text{(complete the transformation)}$$

such that

$$\mathbf{T} = \begin{pmatrix} \mathbf{Y} \\ \mathbf{X} \end{pmatrix} = \begin{pmatrix} \mathbf{t}_0(\mathbf{Z}) \\ \mathbf{t}_1(\mathbf{Z}) \end{pmatrix}$$

$$= \mathbf{t}(\mathbf{Z}) \quad \text{is a one-to-one}$$

Then

$$l(\theta|\mathbf{t}_0) = \log f_{\mathbf{T}_0}(\mathbf{t}_0;\theta) \qquad \text{Incomplete log-likelihood}$$

$$l(\theta|\mathbf{t}_0,\mathbf{t}_1) = \log f(\mathbf{t}_0,\mathbf{t}_1;\theta) \qquad \text{Complete log-likelihood}$$

$$\tilde{l}\left(\theta|\mathbf{t}_0,\theta^{(k)}\right) = E_{\mathbf{T}_1|\mathbf{t}_0,\theta^{(k)}}\{\log f(\mathbf{t}_0,\mathbf{T}_1;\theta)\} \qquad \text{E–Step}$$

$$= \int \cdots \int \log f(\mathbf{t}_0,\mathbf{t}_1;\theta) \frac{f\left(\mathbf{t}_0,\mathbf{t}_1;\theta^{(k)}\right)}{f\left(\mathbf{t}_0;\theta^{(k)}\right)} \mathbf{dt}_1$$

The Ascent Property of the EM Algorithm

We will explicitly consider the discrete case. Derivations for the continuous case are identical.

First define the "improvement" or "increment" functions

$$d(\theta) = l\left(\theta^{(k)}|\mathbf{y}\right) - l(\theta|\mathbf{y}) \tag{4}$$

$$\widetilde{d}(\theta) = \tilde{l}\left(\theta^{(k)}|\mathbf{y}, \theta^{(k)}\right) - \tilde{l}\left(\theta|\mathbf{y}, \theta^{(k)}\right) \tag{5}$$

for the **function we maximize** in the M step, $\tilde{l}\left(\theta|\mathbf{y}, \theta^{(k)}\right)$, and the "target function" which we actually wish to maximize, $l(\theta|\mathbf{y})$, about the current value $\theta^{(k)}$. We have the following

Lemma

$$d(\theta) \leq \widetilde{d}(\theta) \qquad \text{for all} \quad \theta,$$

or equivalently

$$\tilde{l}\left(\theta|\mathbf{y}, \theta^{(k)}\right) - l(\theta|\mathbf{y}) \leq \tilde{l}\left(\theta^{(k)}|\mathbf{y}, \theta^{(k)}\right) - l\left(\theta^{(k)}|\mathbf{y}\right), \qquad \text{for all } \theta. \tag{6}$$

Proof:

$$\tilde{l}\left(\theta|\mathbf{y},\theta^{(k)}\right) - l\left(\theta|\mathbf{y}\right) = E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log f\left(\mathbf{Z};\theta\right)\right\} - \log f\left(\mathbf{y};\theta\right)$$

$$= E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log f\left(\mathbf{Z};\theta\right) - \log f\left(\mathbf{y};\theta\right)\right\}$$

$$= E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log \frac{f\left(\mathbf{Z};\theta\right)}{f\left(\mathbf{y};\theta\right)}\right\}$$

$$= E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log f\left(\mathbf{Z}|\mathbf{y},\theta\right)\right\}$$

$$\leq E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log f\left(\mathbf{Z}|\mathbf{y},\theta^{(k)}\right)\right\} \qquad \text{by the Entropy Inequality}$$

Notice that we have strict inequality unless $f\left(\mathbf{Z}|\mathbf{y},\theta^{(k)}\right) = f\left(\mathbf{Z}|\mathbf{y},\theta\right)$ for some $\theta \neq \theta^{(k)}$. Hence,

$$\tilde{l}\left(\theta|\mathbf{y},\theta^{(k)}\right) - l\left(\theta|\mathbf{y}\right) \leq E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log \frac{f\left(\mathbf{Z};\theta^{(k)}\right)}{f\left(\mathbf{y};\theta^{(k)}\right)}\right\}$$

$$= E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log f\left(\mathbf{Z};\theta^{(k)}\right)\right\} - E_{\mathbf{Z}|\mathbf{y},\theta^{(k)}}\left\{\log f\left(\mathbf{y};\theta^{(k)}\right)\right\}$$

$$= \tilde{l}\left(\theta^{(k)}|\mathbf{y},\theta^{(k)}\right) - l\left(\theta^{(k)}|\mathbf{y}\right),$$

proving the desired inequality.

The Ascending Property of the EM Algorithm

Theorem: Let $\theta^{(k+1)}$ and $\theta^{(k)}$ be two consecutive steps in the EM algorithm. That is

$$\tilde{l}\left(\theta^{(k+1)}|\mathbf{y},\theta^{(k)}\right) \geq \tilde{l}\left(\theta|\mathbf{y},\theta^{(k)}\right), \qquad \text{for all } \theta.$$

Then

$$l\left(\theta^{(k+1)}|\mathbf{y}\right) \geq l\left(\theta^{(k)}|\mathbf{y}\right).$$

Proof.

$$\begin{aligned}
l\left(\theta^{(k+1)}|\mathbf{y}\right) &= \tilde{l}\left(\theta^{(k+1)}|\mathbf{y},\theta^{(k)}\right) - \left[\tilde{l}\left(\theta^{(k+1)}|\mathbf{y},\theta^{(k)}\right) - l\left(\theta^{(k+1)}|\mathbf{y}\right)\right] \\
&\geq \tilde{l}\left(\theta^{(k+1)}|\mathbf{y},\theta^{(k)}\right) - \left[\tilde{l}\left(\theta^{(k)}|\mathbf{y},\theta^{(k)}\right) - l\left(\theta^{(k)}|\mathbf{y}\right)\right] \qquad \text{by (6)} \\
&\geq \tilde{l}\left(\theta^{(k)}|\mathbf{y},\theta^{(k)}\right) + \left[l\left(\theta^{(k)}|\mathbf{y}\right) - \tilde{l}\left(\theta^{(k)}|\mathbf{y},\theta^{(k)}\right)\right] \qquad \text{by definition of } \theta^{(k+1)} \\
&= l\left(\theta^{(k)}|\mathbf{y}\right).
\end{aligned}$$