Information Theory and Data Analysis

Homework #1 (2024-2025), Answers

Q1: The downward bias of entropy – near-worst-case scenario

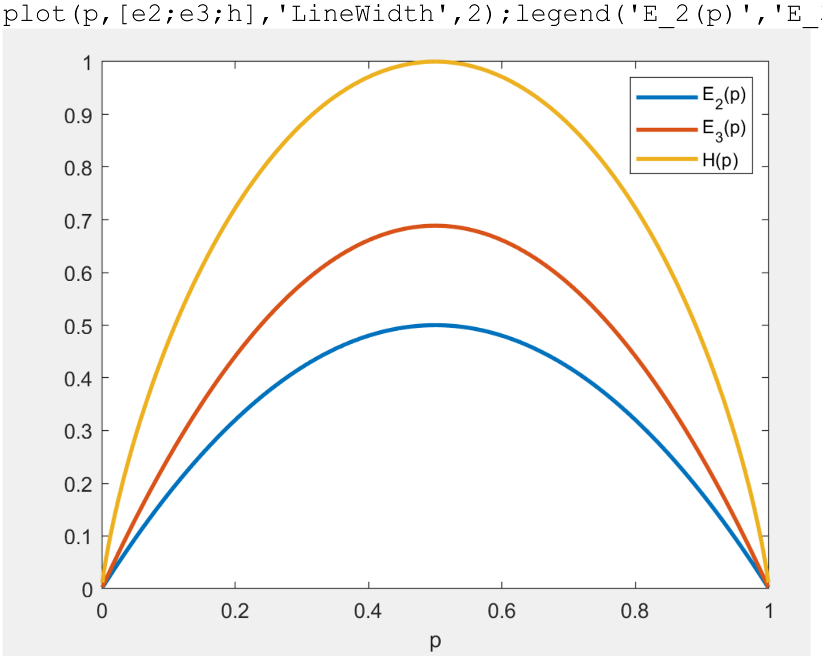
Consider estimating the entropy of a binary variable, whose distribution is defined by p*, where* p *is the probability of drawing a 0, and* $1-p$ *is the probability of drawing a 1. The true entropy is given by* $H(p)=-p\log p-(1-p)\log(1-p)$*. What is the expected value of the naïve ("plug-in") estimate of entropy, based estimating* p *from two samples? From 3 samples? Compare to* $H(p)$*.*

Two samples: With two samples, there are only four possible sets of observations: $\{0,0\}$, $\{0,1\}$, $\{1.0\}$, and $\{1,1\}$, with probabilities given, respectively, by p^2, $p(1-p)$, $(1-p)p$, and $(1-p)^2$. With either the first or the last draw, the experimental estimate of p is either 0 or 1, leading to a naïve estimate of entropy of 0. With the other two draws, the experimental estimate of p is $1/2$, so the plug-in estimate of the entropy (using $\log_2$) is 1. So the expected value of the plug-in estimate is $E_2(p)=p(1-p)+(1-p)p=2p(1-p)$.

Three samples: Three draws of the same token occur with probability $p^3+(1-p)^3$; the other draws, in which one symbol is drawn once and the other symbol is drawn twice, occur with total probability $3p^2(1-p)+3p(1-p)^2=3p(1-p)\big(p+(1-p)\big)=3p(1-p)$. In the latter case, the experimental estimate of p is $1/3$ and of $(1-p)$ is $2/3$, or vice-versa, so
$$E_3(p)=-3p(1-p)\left(\frac{1}{3}\log_2\frac{1}{3}+\frac{2}{3}\log_2\frac{2}{3}\right)=-p(1-p)\left(-\log_2 3-2\log_2\frac{3}{2}\right)=p(1-p)\log_2\frac{27}{4}\approx 2.75p(1-p).$$

All are symmetric about their maxima at $p=1/2$ and are concave down, and are zero at the extremes, they differ substantially. Plotted via this matlab script below:

```
p=[0.001:0.001:.999];e2=2*p.*(1-p);e3=p.*(1-p)*log2(27/4);h=-p.*log2(p)-(1-p).*log2(1-p);
plot(p,[e2;e3;h],'LineWidth',2);legend('E_2(p)','E_3(p)','H(p)');xlabel('p');
```



Q2. Differential entropy of a multivariate Gaussian

Information Theory and Data Analysis

Recall that the multivariate Gaussian distribution for a variable $\vec{x}$ (a column vector of length n) with mean zero and covariance matrix $\langle \vec{x} \bullet \vec{x}^T \rangle = V$ is given by $p_V(\vec{x}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det V}} \exp\left(-\frac{\vec{x}^T V^{-1}\vec{x}}{2}\right)$. What is the differential entropy $(\log_e)$ *of $p_V(\vec{x})$?*

$$-\int p_V(\vec{x})\log(p_V(\vec{x}))\,d\vec{x} = -\int p_V(\vec{x})\left(\log\frac{1}{(2\pi)^{n/2}\sqrt{\det V}} - \frac{\vec{x}^T V^{-1}\vec{x}}{2}\right)d\vec{x}$$

$$= -\int p_V(\vec{x})\left(\log\frac{1}{(2\pi)^{n/2}\sqrt{\det V}}\right)d\vec{x} + \int p_V(\vec{x})\left(\frac{\vec{x}^T V^{-1}\vec{x}}{2}\right)d\vec{x}$$

$$= -\log\frac{1}{(2\pi)^{n/2}\sqrt{\det V}}\int p_V(\vec{x})d\vec{x} + \int p_V(\vec{x})\left(\frac{\vec{x}^T V^{-1}\vec{x}}{2}\right)d\vec{x}$$

$$= \left(\frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det V\right) + \int p_V(\vec{x})\left(\frac{\vec{x}^T V^{-1}\vec{x}}{2}\right)d\vec{x}$$

For second term, note that, since $\vec{x}^T V^{-1}\vec{x}$ is a scalar, $\vec{x}^T V^{-1}\vec{x} = tr\left(\vec{x}^T V^{-1}\vec{x}\right) = tr\left(V^{-1}\vec{x}\vec{x}^T\right)$. The expected value of $\vec{x}\vec{x}^T$ is the covariance matrix V. So $\int p_V(\vec{x})\left(\frac{\vec{x}^T V^{-1}\vec{x}}{2}\right)d\vec{x} = \frac{1}{2}\langle trV^{-1}\vec{x}\vec{x}^T\rangle = \frac{1}{2}\langle trI_n\rangle = \frac{n}{2}$, so

$$-\int p_V(\vec{x})\log(p_V(\vec{x}))\,d\vec{x} = \frac{n}{2}\log(2\pi) + \frac{1}{2}\log\det V + \frac{n}{2} = \frac{1}{2}\left(n + n\log 2\pi + \log\det V\right).$$

B. Use this result to show that if $\vec{x}$ is a column vector of length n_X drawn from a Gaussian with mean zero and covariance matrix $\langle \vec{x} \bullet \vec{x}^T \rangle = V_X$, with differential entropy H_X, and $\vec{y}$ is a column vector of length n_Y drawn independently from a Gaussian with mean zero and covariance matrix $\langle \vec{y} \bullet \vec{y}^T \rangle = V_Y$, with differential entropy H_Y then the joint distribution of $\vec{x}$ and $\vec{y}$ has differential entropy $H_{X,Y} = H_X + H_Y$.

Since $\vec{x}$ and $\vec{y}$ are independent, the covariance matrix of $\begin{pmatrix}\vec{x}\\ \vec{y}\end{pmatrix}$ is a block-diagonal matrix $\begin{pmatrix}V_X & 0\\ 0 & V_Y\end{pmatrix}$, whose determinant is $\det V_{X,Y} = \det V_X \det V_Y$. And clearly $n_{X,Y} = n_X + n_Y$. So

$$H_{X,Y} = \frac{1}{2}\left(n_{X,Y} + n_{X,Y}\log 2\pi + \log\det V_{X,Y}\right) = \frac{1}{2}\left((n_X + n_Y) + (n_X + n_Y)\log 2\pi + \log(\det V_X \det V_Y)\right)$$

$$= \frac{1}{2}\left(n_X + n_X\log 2\pi + \log\det V_X\right) + \frac{1}{2}\left(n_Y + n_Y\log 2\pi + \log\det V_Y\right) = H_X + H_Y$$