The joint distribution is

$p(X,Y)$	$Y=0$	$Y=1$
$X=0$	$\frac{1}{8}$	$\frac{3}{8}$
$X=1$	$\frac{3}{8}$	$\frac{1}{8}$

so the joint entropy is given by

$$\mathbb{H}(X,Y) = -\left[\frac{1}{8}\log_2\frac{1}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{3}{8}\log_2\frac{3}{8} + \frac{1}{8}\log_2\frac{1}{8}\right] = 1.81 \text{ bits} \tag{6.8}$$

Clearly the marginal probabilities are uniform: $p(X=1) = p(X=0) = p(Y=0) = p(Y=1) = 0.5$, so $\mathbb{H}(X) = \mathbb{H}(Y) = 1$. Hence $\mathbb{H}(X,Y) = 1.81$ bits $< \mathbb{H}(X) + \mathbb{H}(Y) = 2$ bits. In fact, this upper bound on the joint entropy holds in general. If X and Y are independent, then $\mathbb{H}(X,Y) = \mathbb{H}(X) + \mathbb{H}(Y)$, so the bound is tight. This makes intuitive sense: when the parts are correlated in some way, it reduces the "degrees of freedom" of the system, and hence reduces the overall entropy.

What is the lower bound on $\mathbb{H}(X,Y)$? If Y is a deterministic function of X, then $\mathbb{H}(X,Y) = \mathbb{H}(X)$. So

$$\mathbb{H}(X,Y) \geq \max\{\mathbb{H}(X), \mathbb{H}(Y)\} \geq 0 \tag{6.9}$$

Intuitively this says combining variables together does not make the entropy go down: you cannot reduce uncertainty merely by adding more unknowns to the problem, you need to observe some data, a topic we discuss in Sec. 6.1.4.

We can extend the definition of joint entropy from two variables to n in the obvious way.

6.1.4 Conditional entropy

The **conditional entropy** of Y given X is the uncertainty we have in Y after seeing X, averaged over possible values for X:

$$\mathbb{H}(Y|X) \triangleq \mathbb{E}_{p(X)}\left[\mathbb{H}(p(Y|X))\right] \tag{6.10}$$

$$= \sum_x p(x)\,\mathbb{H}(p(Y|X=x)) = -\sum_x p(x)\sum_y p(y|x)\log p(y|x) \tag{6.11}$$

$$= -\sum_{x,y} p(x,y)\log p(y|x) = -\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)} \tag{6.12}$$

$$= -\sum_{x,y} p(x,y)\log p(x,y) + \sum_x p(x)\log\frac{1}{p(x)} \tag{6.13}$$

$$= \mathbb{H}(X,Y) - \mathbb{H}(X) \tag{6.14}$$

If Y is a deterministic function of X, then knowing X completely determines Y, so $\mathbb{H}(Y|X) = 0$. If X and Y are independent, knowing X tells us nothing about Y and $\mathbb{H}(Y|X) = \mathbb{H}(Y)$. Since $\mathbb{H}(X,Y) \leq \mathbb{H}(Y) + \mathbb{H}(X)$, we have

$$\mathbb{H}(Y|X) \leq \mathbb{H}(Y) \tag{6.15}$$

Summary of Comments on pml1.pdf

Page: 156

Author: petercerno Subject: Comment on Text Date: 02.02.21, 15:51:24

Should be -

Author: petercerno Subject: Sticky Note Date: 02.02.21, 15:52:05

Or can be + \sum_x p(x) log p(x)

6.2.2 Interpretation

We can rewrite the KL as follows:

$$\mathbb{KL}(p\|q) = \underbrace{\sum_{k=1}^{K} p_k \log p_k}_{-\mathbb{H}(p)} \underbrace{- \sum_{k=1}^{K} p_k \log q_k}_{\mathbb{H}(p,q)} \tag{6.30}$$

We recognize the first term as the negative entropy, and the second term as the cross entropy. Thus we can interpret the KL divergence as the "extra number of bits" you need to pay when compressing data samples from p using the incorrect distribution q as the basis of your coding scheme.

There are various other interpretations of KL divergence. See the sequel to this book, [Mur22], for more information.

6.2.3 Example: KL divergence between two Gaussians

For example, one can show that the KL divergence between two multivariate Gaussian distributions is given by

$$\begin{aligned} &\mathbb{KL}(\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)\|\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)) \\ &= \frac{1}{2}\left[\operatorname{tr}(\boldsymbol{\Sigma}_2^{-1}\boldsymbol{\Sigma}_1) + (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_2^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) - D + \log\left(\frac{\det(\boldsymbol{\Sigma}_2)}{\det(\boldsymbol{\Sigma}_1)}\right)\right] \end{aligned} \tag{6.31}$$

In the scalar case, this becomes

$$\mathbb{KL}(\mathcal{N}(x|\mu_1, \sigma_1)\|\mathcal{N}(x|\mu_2, \sigma_2)) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2} \tag{6.32}$$

6.2.4 Non-negativity of KL

In this section, we prove that the KL divergence is always non-negative.

Theorem 6.2.1. *(Information inequality)* $\mathbb{KL}(p\|q) \geq 0$ *with equality iff* $p = q$.

Proof. We now prove the theorem following [CT06, p28]. Let $A = \{x : p(x) > 0\}$ be the support of $p(x)$. Using the convexity of the log function and Jensen's inequality (Sec. B.4.3), we have that

$$-\mathbb{KL}(p\|q) = -\sum_{x\in A} p(x)\log\frac{p(x)}{q(x)} = \sum_{x\in A} p(x)\log\frac{q(x)}{p(x)} \tag{6.33}$$

$$\leq \log\sum_{x\in A} p(x)\frac{q(x)}{p(x)} = \log\sum_{x\in A} q(x) \tag{6.34}$$

$$\leq \log\sum_{x\in\mathcal{X}} q(x) = \log 1 = 0 \tag{6.35}$$

Since $\log(x)$ is a strictly concave function ($-\log(x)$ is convex), we have equality in Eq. (6.34) iff $p(x) = cq(x)$ for some c that tracks the fraction of the whole space $\mathcal{X}$ contained in A. We have equality in Eq. (6.35) iff $\sum_{x\in A} q(x) = \sum_{x\in\mathcal{X}} q(x) = 1$, which implies $c = 1$. Hence $\mathbb{KL}(p\|q) = 0$ iff $p(x) = q(x)$ for all x. □

Page: 160

Author: petercerno Subject: Comment on Text Date: 02.02.21, 16:19:18

Nit: concavity?

6.3.4 Conditional mutual information

We can define the **conditional mutual information** in the obvious way

$$\mathbb{I}(X;Y|Z) \triangleq \mathbb{E}_{p(Z)}\left[\mathbb{I}(X;Y)|Z\right] \tag{6.53}$$
$$= \mathbb{E}_{p(x,y,z)}\left[\log \frac{p(x,y|z)}{p(x|z)p(y|z)}\right] \tag{6.54}$$
$$= \mathbb{H}(X|Z) + \mathbb{H}(Y|Z) - \mathbb{H}(X,Y|Z) \tag{6.55}$$
$$= \mathbb{H}(X|Z) - \mathbb{H}(X|Y,Z) = \mathbb{H}(Y|Z) - \mathbb{H}(Y|X,Z) \tag{6.56}$$
$$= \mathbb{H}(X,Z) + \mathbb{H}(Y,Z) - \mathbb{H}(Z) - \mathbb{H}(X,Y,Z) \tag{6.57}$$
$$= \mathbb{I}(Y;X,Z) - \mathbb{I}(Y;Z) \tag{6.58}$$

The last equation tells us that the conditional MI is the extra (residual) information that X tells us about Y, excluding what we already knew about Y given Z alone.

We can rewrite Eq. (6.58) as follows:

$$\mathbb{I}(Z,Y;X) = \mathbb{I}(Z;X) + \mathbb{I}(Y;X|Z) \tag{6.59}$$

Generalizing to N variables, we get the **chain rule for mutual information**:

$$\mathbb{I}(Z_1,\ldots,Z_N;X) = \sum_{n=1}^{N} \mathbb{I}(Z_n;X|Z_1,\ldots,Z_{n-1}) \tag{6.60}$$

6.3.5 Normalized mutual information

For some applications, it is useful to have a normalized measure of dependence, between 0 and 1. We now discuss one way to construct such a measure.

First, note that

$$\mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) \leq \mathbb{H}(X) \tag{6.61}$$
$$= \mathbb{H}(Y) - \mathbb{H}(Y|X) \leq \mathbb{H}(Y) \tag{6.62}$$

so

$$0 \leq \mathbb{I}(X;Y) \leq \min(\mathbb{H}(X), \mathbb{H}(Y)) \tag{6.63}$$

Therefore we can define the **normalized mutual information** as follows:

$$NMI(X,Y) = \frac{\mathbb{I}(X;Y)}{\min(\mathbb{H}(X), \mathbb{H}(Y))} \leq 1 \tag{6.64}$$

This normalized mutual information ranges from 0 to 1. When $NMI(X,Y) = 0$, $\mathbb{I}(X;Y) = 0$ so X and Y are independent. Without loss of generality assume X has the higher entropy: $NMI(X,Y) = 1 \implies \mathbb{I}(X;Y) = \mathbb{H}(X) - \mathbb{H}(X|Y) = \mathbb{H}(X) \implies \mathbb{H}(X|Y) = 0$ and so X is a deterministic function of Y.

Page: 165

Author: petercerno Subject: Comment on Text Date: 02.02.21, 16:52:51

lower

Author: petercerno Subject: Sticky Note Date: 02.02.21, 16:54:01

X cannot be a deterministic function of Y and at the same time have higher entropy than Y

An example of a sufficient statistic is the data itself, $s(\mathcal{D}) = \mathcal{D}$, but this is not very useful, since it doesn't summarize the data at all. Hence we define a **minimal sufficient statistic** $s(\mathcal{D})$ as one which is sufficient, and which contains no extra information about θ; thus $s(\mathcal{D})$ maximally compresses the data $\mathcal{D}$ without losing information which is relevant to predicting θ. More formally, we say s is a *minimal* sufficient statistic for $\mathcal{D}$ if for all sufficient statistics $s'(\mathcal{D})$ there is some function f such that $s(\mathcal{D}) = f(s'(\mathcal{D}))$. We can summarize the situation as follows:

$$\theta \to s(\mathcal{D}) \to s'(\mathcal{D}) \to \mathcal{D} \tag{6.78}$$

Here $s'(\mathcal{D})$ takes $s(\mathcal{D})$ and adds redundant information to it, thus creating a one-to-many mapping.

For example, a minimal sufficient statistic for a set of N Bernoulli trials is simply N and $N_1 = \sum_n \mathbb{I}(X_n = 1)$, i.e., the number of successes. In other words, we don't need to keep track of the entire sequence of heads and tails and their ordering, we only need to keep track of the total number of heads and tails. Similarlt, for inferring the mean of a Gaussian distribution with known variance we only need to know the empirical mean and number of samples.

6.3.9 Fano's inequality

A common method for **feature selection** is to pick input features X_d which have high mutual information with the response variable Y. Below we justify why this is a reasonable thing to do. In particular, we state a result, known as **Fano's inequality**, which bounds the probability of misclassification (for any method) in terms of the mutual information between the features X and the class label Y.

Theorem 6.3.2. *(Fano's inequality) Consider an estimator $\hat{Y} = f(X)$ such that $Y \to X \to \hat{Y}$ forms a Markov chain. Let E be the event $\hat{Y} \neq Y$, indicating that an error occured, and let $P_e = P(Y \neq \hat{Y})$ be the probability of error. Then we have*

$$\mathbb{H}(Y|X) \le \mathbb{H}\left(Y|\hat{Y}\right) \le \mathbb{H}(E) + P_e \log|\mathcal{Y}| \tag{6.79}$$

Since $\mathbb{H}(E) \le 1$, as we saw in Fig. 6.1, we can weaken this result to get

$$1 + P_e \log|\mathcal{Y}| \ge \mathbb{H}(Y|X) \tag{6.80}$$

and hence

$$P_e \ge \frac{\mathbb{H}(Y|X) - 1}{\log|\mathcal{Y}|} \tag{6.81}$$

Thus minimizing $\mathbb{H}(Y|X)$ (which can be done by maximizing $\mathbb{I}(X;Y)$) will also minimize the lower bound on P_e.

Proof. (From [CT06, p38].) Using the chain rule for entropy, we have

$$\mathbb{H}\left(E, Y|\hat{Y}\right) = \mathbb{H}\left(Y|\hat{Y}\right) + \underbrace{\mathbb{H}\left(E|Y, \hat{Y}\right)}_{=0} \tag{6.82}$$

$$= \mathbb{H}\left(E|\hat{Y}\right) + \mathbb{H}\left(Y|E, \hat{Y}\right) \tag{6.83}$$

Page: 169

Author: petercerno Subject: Comment on Text Date: 03.02.21, 09:44:31
Similarly