Machine learning

# Appendix 1

*Lecturer:Haim Permuter* *Scribe: Gal Rattner*

## I. Information Measures

### A. Entropy

**Definition 1 (Entropy)** The entropy of a discrete random variable $X$ with a finite alphabet is defined as

$$\begin{aligned} H(X) &\triangleq \mathbb{E}[-\log P_X(X)] \\ &= -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x). \end{aligned} \tag{1}$$

The entropy resembles the uncertainty of the random variable, and in information theory used to represent the number of bits required to describe the random variable, using base 2 logarithm.

**Example 1 (Entropy of a Bernoulli random variable)** Let $X \sim Bern(p)$, i.e., $\Pr\{X=1\}=p$ and $\Pr\{X=0\}=1-p$.

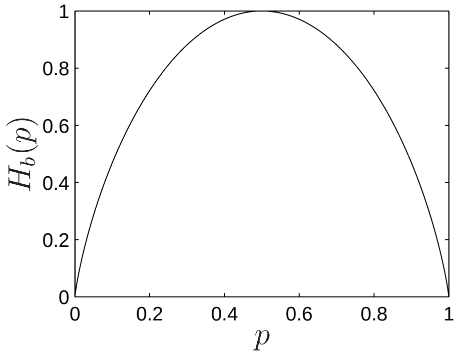


Figure 1. Binary Entropy $H_b(p)$. The figure depicts the entropy of a Bernoulli random variable with parameter $0 \le p \le 1$.

$$H(X) = -p\log p - (1-p)\log(1-p). \tag{2}$$

We denote the entropy of a Bernoulli random variable with parameter $p$ as $H_b(p)$. Fig. 1 depicts $H_b(p)$ as a function of $p$. Note that $H_b(0) = H_b(1) = 0$, the maximum is obtained at $p = 0.5$, $H_b(p) = H_b(1-p)$ and the curve is concave.

We can now refer to the general case of discrete random variable $X$, and note that the next assumptions stand:

(a) $H(X) \geq 0$.

(b) $H(X) = 0 \Longleftrightarrow X$ is a constant (deterministic).

(c) $H(X)$ curve is concave.

(d) $H(X) \leq \log|\mathcal{X}|$ where $\mathcal{X}$ is the alphabet of $X$.

Where (a) follows directly from $0 \leq P(x) \leq 1$ and therefore $\log P(x) \leq 0$, (b) follows from the fact that for constant $x$, $P(x) \in \{0, 1\}$ either $P(x) = 0$ or $\log P(x) = 0$ and (d) can be proved using the divergence definition as presented forward in this paper. For full proofs see lecture 1 of Information theory course [2] or T. Cover's book [3].

**Example 2 (Mean Length)** Let $X$ be a random variable with $\mathcal{X} = \{1, 2, 3, 4\}$, and

$$X = \begin{cases} 1 & \text{p(x) = 1/2} \\ 2 & \text{p(x) = 1/4} \\ 3 & \text{p(x) = 1/8} \\ 4 & \text{p(x) = 1/8} \end{cases}$$

We can now consider a compressed coding method, different than the standard code, i.e.

| $x$ | $p(x)$ | standard code | compressed code |
|---|---|---|---|
| 1 | $\frac{1}{2}$ | 00 | 0 |
| 2 | $\frac{1}{4}$ | 01 | 10 |
| 3 | $\frac{1}{8}$ | 10 | 110 |
| 4 | $\frac{1}{8}$ | 11 | 111 |

Note that the expected code length obtained by the compressed coding equals $H(X)$, i.e.

$$\begin{aligned} \mathbb{E}\left[l(X)\right] &= \frac{1}{2}\cdot 1 + \frac{1}{4}\cdot 2 + \frac{1}{8}\cdot 3 + \frac{1}{8}\cdot 3 \\ &= 1\frac{3}{4}[bits] \end{aligned}$$

$$
\begin{aligned}
&= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} \\
&= H(X). \qquad (3)
\end{aligned}
$$

### B. Divergence

**Definition 2 (Kullback-Leibler divergence)** Consider two different probability mass functions $p(x)$ and $q(x)$, over the same alphabet $\mathcal{X}$. The *Kullback-Leibler divergence* also known as *relative entropy* is defined as

$$
\begin{aligned}
D(P_X||Q_X) &\triangleq \sum_x P(x)\log\frac{P(x)}{Q(x)} \\
&= \mathbb{E}_P\left[\log\frac{P(X)}{Q(X)}\right]. \qquad (4)
\end{aligned}
$$

Note that the next assumptions stand for the KL-divergence:

(a) $D(P_X||Q_X)$ is convex in the pair $(P, Q)$

(b) $D(P_X||Q_X) \geq 0$.

(c) $D(P_X||Q_X) = 0 \iff p_x = q_x$.

Where (a) is shown in [2], (b) follows from Jensen's inequality and the fact that $\sum_x P(x) = 1$, and (c) from strict concavity of the logarithm function. For full proofs see lecture 2 of Information theory course [2] or T. Cover's book [3]. Consider the task of binary classification using a single neuron, where $y \in \{0, 1\}$. In such case, we can think of $y$ and $a(x)$ as two different probability functions over the class of $x$, denoted as $c_x$. We can now use the notation $y = P\{Class_x = 1\} = p(c_x)$ for the true classification probability, and $a(x) = P\{Cl\hat{a}ss_x = 1\} = q(\hat{c_x})$ for the estimated classification probability from the net. Recall that the cross entropy cost for a single neuron was defined as

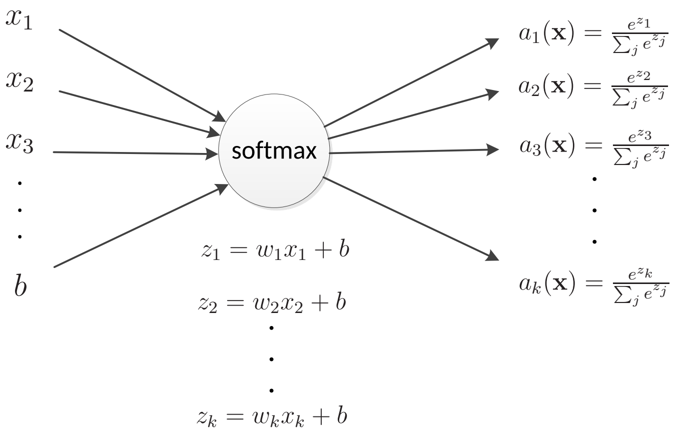
$$
\begin{aligned}
C(x, y) &= -y\log a(x) - \overline{y}\log\overline{a}(x) \\
&= -y\log\sigma(z) - \overline{y}\log\overline{\sigma}(z). \qquad (5)
\end{aligned}
$$

Therefore using the definition of the two probability mass functions above, we get

$$
-\sum_{c_x} p(c_x)\log q(c_x) = -\sum_{c_x} p(c_x)\log\frac{q(c_x)}{p(c_x)} \cdot p(c_x)
$$

$$
\begin{aligned}
&= \sum_{c_x} p(c_x) \log \frac{q(c_x)}{p(c_x)} - \sum_{c_x} p(c_x) \log p(c_x) \\
&= D(P||Q) + H(P). \qquad (6)
\end{aligned}
$$

So far we applied the cross entropy cost to the case of classifying with two possible classes. Now consider the task of classification where the number of possible classes is larger than two, i.e. $y \in \{1, \dots, K\},\ K > 2$. We would like the net to produce a probability vector $\mathbf{a}(\mathbf{x})$, with $i = 1, \dots, K$ elements all summed to 1, s.t. $\forall i\ a_i(\mathbf{x})$ will represent the probability that the input $\mathbf{x}$ is of class $i$. The softmax regression layer is commonly used as an output layer in such casses, it is presented in Figure (2) and described with more details in lecture 3. The cross entropy cost in the multiple classes case is given by



$x_1$
$x_2$
$x_3$
$b$
softmax
$z_1 = w_1 x_1 + b$
$z_2 = w_2 x_2 + b$
$z_k = w_k x_k + b$
$a_1(\mathbf{x}) = \frac{e^{z_1}}{\sum_j e^{z_j}}$
$a_2(\mathbf{x}) = \frac{e^{z2}}{\sum_j e^{z_j}}$
$a_3(\mathbf{x}) = \frac{e^{z3}}{\sum_j e^{z_j}}$
$a_k(\mathbf{x}) = \frac{e^{z_k}}{\sum_j e^{z_j}}$

Figure 2. The softmax layer.

$$
\begin{aligned}
C(\mathbf{x}, y) &= -\sum_{i=1}^{K} \mathbb{1}\{y = i\} \log a_i(\mathbf{x}) \\
&= -\log a_{i=y}(\mathbf{x}), \qquad (7)
\end{aligned}
$$

where each element in the output vector is in range $[0, 1]$.

For further read please see Information Theory course's lectures [2].

## References

[1] M. Nielsen *Neural Networks and Deep Learning, Chap. 3*. http://neuralnetworksanddeeplearning.com/chap3.html. January 2016.

[2] H. Permuter *Introduction to Information Theory course*. http://www.ee.bgu.ac.il/ haimp/it/index.html

[3] T. M. Cover and J. A. Thomas *Elements of Information Theory, Chap. 1*.