Supervised Learning :: CHEAT SHEET

Entropy

Entropy of Discreate Random Variables

Entropy of a discrete random variable X with domain $\mathcal{X}$ and pmf $p(x)$:

$$H(X) := H(p) = -\mathbb{E}[\log_2(p(X))] = -\sum_{x\in\mathcal{X}} p(x)\log_2 p(x)$$

Properties of discrete entropy:

- Entropy is non-negative, so $H(X) >= 0$.
- If one event has probability $p(x) = 1$, then $H(X) = 0$.
- Symmetry. Reordering values of $p(x)$ does not change entropy.
- Adding or removing an event with $p(x) = 0$ does not change entropy.
- $H(X)$ is continuous in probabilities $p(x)$.
- Entropy is additive for independent RVs.
- Entropy is maximal for a uniform distribution.

Differential Entropy of Continuous Random Variables

Differential entropy of a continuous random variable X with density function $f(x)$ and support $\mathcal{X}$:

$$h(X) := h(f) := -\mathbb{E}[\log(f(x))] = -\int_{\mathcal{X}} f(x)\log(f(x))dx$$

Properties of differential entropy:

- $h(f)$ can be negative.
- $h(f)$ is additive for independent RVs.
- $h(f)$ is maximized by the multivariate normal, if we restrict to all distributions with the same (co)variance, so $h(X) \le \frac{1}{2}\ln(2\pi e)^n|\Sigma|$.
- Translation-invariant, $h(X + a) = h(X)$.
- $h(AX) = h(X) + \log|A|$ for random vectors and matrix A.
- For a given variance, the continuous distribution that maximizes differential entropy is the Gaussian.

Joint and Continuous Entropy

Joint Entropy

Discrete:
Joint entropy of n discrete random variables $X_1, X_2, \ldots, X_n$:

$$H(X_1, X_2, \ldots, X_n) = -\sum_{x_1\in\mathcal{X}_1}\cdots\sum_{x_n\in\mathcal{X}_n} p(x_1, x_2, \ldots, x_n)\log_2(p(x_1, x_2, \ldots, x_n))$$

Continuous:
Joint differential entropy of a continuous random vector X with density function $f(x)$ and support $\mathcal{X}$:

$$h(X) = h(X_1, \ldots, X_n) = h(f) = -\int_{\mathcal{X}} f(x)\log(f(x))dx$$

Conditional Entropy

Discrete:
Conditional entropy of Y given X for $(X, Y) \sim p(x, y)$:

$$H(Y|X) = \mathbb{E}_X[H(Y|X = x)] = \sum_{x\in\mathcal{X}} p(x)H(Y|X = x)$$
$$= -\sum_{x\in\mathcal{X}} p(x)\sum_{y\in\mathcal{Y}} p(y|x)\log p(y|x) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log p(y|x)$$

Continuous:
Conditional entropy of Y given X (both continuous):

$$h(Y|X) = -\int f(x, y)\log f(x|y)dxdy.$$

Properties:

- $H(X, X) = H(X)$
- $H(X|X) = 0$
- $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$
- $H(X|Y) \le H(X)$
- If $H(X|Y) = 0$, then X is a function of Y

Chain rule for entropy:

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

n-Variable version:

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)$$

Cross-Entropy and Kullback-Leibler Divergence

Cross-entropy of two distributions p and q on the same domain $\mathcal{X}$:

Discrete:

$$H(p\|q) = \sum_{x\in\mathcal{X}} p(x)\log(\frac{1}{q(x)}) = -\sum_{x\in\mathcal{X}} p(x)\log(q(x)) = -\mathbb{E}_{X\sim p}[\log(q(X))]$$

Continuous:

$$H(p\|q) = \int p(x)\log(\frac{1}{q(x)})dx = -\int p(x)\log(q(x))dx = -\mathbb{E}_{X\sim p}[\log(q(X))]$$

Kullback-Leibler Divergence

Discrete:

$$D_{KL}(p\|q) = \mathbb{E}_p[\log\frac{p(X)}{q(X)}] = \sum_{x\in\mathcal{X}} p(x)\cdot\log\frac{p(x)}{q(x)}$$

Continuous:

$$D_{KL}(p\|q) = \mathbb{E}_p[\log\frac{p(X)}{q(X)}] = \int_{x\in\mathcal{X}} p(x)\cdot\log\frac{p(x)}{q(x)}$$

Relation

$$H(p\|q) = H(p) + D_{KL}(p\|q)$$

Mutual Information

Mutual information between X and Y :
Discrete:

$$I(X; Y) = \mathbb{E}_{p(x,y)}[\log\frac{p(X, Y)}{p(X)p(Y)}] = \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x, y)\log\frac{p(x, y)}{p(x)p(y)}$$
$$= H(X) - H(X|Y)$$

Continuous:

$$I(X; Y) = \int f(x, y)\log\frac{f(x, y)}{f(x)f(y)}dxdy$$

Properties:

- $I(X; Y) = H(X) - H(X|Y)$
- $I(X; Y) = H(Y) - H(Y|X)$
- $I(X; Y) = H(X) + H(Y) - H(X, Y)$
- $I(X; Y) = I(Y; X)$
- $I(X; X) = H(X)$
- $I(X; Y) \ge 0$, with equality if and only if X and Y are independent