Information and Entropy

Data Mining

Edgar Roman-Rangel.
edgar.roman@itam.mx

Computer Science Department.
Instituto Tecnológico Autónomo de México, ITAM.

Outline

Definition

Remember:

Information

Mined or filtered data. Identification of relevant patterns within the data. Conveys meaning (purpose). Help us take decisions.

Information metric

How do we measure the amount of information in data?

In general terms, we consider that a variable is informative if it is not a constant. Often, the more it varies, the more informative it is.

Depending on the type of the variable, we could measure,

- Variance: for numeric data.
- Entropy: for categorical data.

We could also measure a metric of correlation between pairs of variables,

- Correlation index: between two numeric variables.
- Mutual information: between two categorical variables.
- ANOVA: between a numeric and a categorical variable.

Information Theory

Branch of applied mathematics that help us quantify the amount of information contained in a signal (data).

Intuition: learning that an unlikely event has happened is more informative than knowing that a likely event happened.

Example: the message "the sun rose this morning" is so uninformative, that it is unnecessary to communicate it, whereas the message "there was a solar eclipse this morning" is highly informative.

Information Theory

Claude Shannon, 1948.
"A Mathematical Theory of Communication".

Assumptions:

- Highly probable events contain little information.
- In the extreme case, events whose occurrence is certain, should not have information at all.
- Unlikely events, must have a lot of information.
- Independent events, must have additive information.
- The length of the message carrying the information, must be proportional to the amount of information.

Distribution of events

Very certain outcome

- One-hot distribution on possible outcomes.
- Low entropy.
- 1 event $\implies$ no information.

Very uncertain outcome

- Uniform distribution on possible outcomes.
- High entropy.
- 1 event $\implies$ lots of information.

Information

- a.k.a., information content, self-information, surprise, Shannon entropy.
- Let's think of an event as a random variable.
- The information (*surprise*) $I(x)$ for a random variable x, with probability $p(x)$, is defined as the inverse of its probability:

$$I(x) = \frac{1}{p(x)}.$$

- If an event is highly probable, then there is no surprise.

Chance and surprise $\Longrightarrow$ information

- If $p(x) = 1 \rightarrow I(x) = \frac{1}{1} = 1$. Not what we are looking for.
- Use $\log(\cdot)$ instead: $\log\left(\frac{1}{1}\right) = 0$.
  So, let's rephrase:

$$\begin{aligned} I(x) &= \log\left(\frac{1}{p(x)}\right), \\ &= -\log[p(x)]. \end{aligned}$$

- If $p(x) = 0 \rightarrow \log\left(\frac{1}{0}\right) = \log(1) - \log(0) =$ undefined, surprise of something that never happens.
- Often, $\ln(\cdot)$ is used (nats), or $\log_2(\cdot)$ (bits).

Outline

Entropy

Entropy: expected value of surprise.

- We can quantify the amount of information in a whole *pdf* using the concept of Entropy (Shannon Entropy).

$$\begin{aligned} H(x) &= \mathbb{E}[I(x)], \\ &= -\sum_x p(x) \log p(x). \end{aligned}$$

- $H(x)$ indicates the amount of information expected from an event sampled from that distribution.
- This is, the number of bits required to encode (communicate) the outcomes of such a distribution.

Example 1: weather 50%-50% chances

Example 2: weather 75%-25% chances

Example 3: weather 8 possible outcomes

even distribution

Example 4: weather 8 possible outcomes

uneven distribution: $[35, 35, 10, 10, 4, 4, 1, 1]\%$

Example 5: events as random variables

Fair coin

Ten realizations produced: H, H, T, H, T, T, T, H, T, H.

$$\begin{aligned} H(x) &= 0.5 \times \log_2\left(\frac{1}{0.5}\right) + 0.5 \times \log_2\left(\frac{1}{0.5}\right), \\ &= 2(0.5)(-\log_2(0.5)), \\ &= 1.0. \end{aligned}$$

Unfair coin

Ten realizations produced: H, H, T, H, H, H, H, H, H, H.

$$\begin{aligned} H(x) &= 0.9 \times \log_2\left(\frac{1}{0.9}\right) + 0.1 \times \log_2\left(\frac{1}{0.1}\right), \\ &= 0.47. \end{aligned}$$

Outline

Differential entropy

The entropy for a continuous random variable.

- Consider *pdf* instead of the *pmf*.

$$h(x) = -\int_{\mathcal{X}} P(x) \log P(x) dx.$$

Join entropy

The surprise of observing two random variables, i.e., a join distribution.

$$H(x,y) = -\sum_x \sum_y p(x,y) \log p(x,y).$$

Mutual information

Mutual information of two random variables x and y, is a measure of the mutual dependence between them.

- Amount of information about one random variable, obtained by observing the other one.
- How different is the join distribution $p(x, y)$ from the product of the marginal distributions $p(x)$ and $p(y)$.

$$I(x; y) = \sum_x \sum_y p(x, y) \log \left[\frac{p(x, y)}{p(x)p(y)} \right].$$

Cross entropy

Between two probability distributions $p(x)$ and $q(x)$, where both of them are defined over the same underlying set of possible values, $\mathcal{X} : \{x_1, \ldots, x_N\}$.

- Measures the average length, in bits, needed to inform (encode) an event using an assumed probability distribution q, rather than the true distribution p.

$$\begin{aligned} H(p, q) &= -\mathbb{E}_p \left[\log q\right], \\ &= -\sum_x p(x) \log q(x). \end{aligned}$$

Kullback-Leibler divergence

Divergence between two probability distributions.

- Measures how much one observed probability distribution p is different from a second reference probability distribution q.
- Expected excess of surprise from using q as a model of the data when the actual distribution is p.

$$D_{KL}(p||q) = \sum_x p(x) \log \left[\frac{p(x)}{q(x)} \right].$$

Jensen-Shannon divergence

D_{KL} is a distance but not a metric: it is not symmetric.

The Jensen-Shannon divergence is an alternative that is symmetric.

$$D_{JS}(p||q) = \frac{1}{2}D_{KL}(p||m) + \frac{1}{2}D_{KL}(q||m),$$

where, m is the element-wise average between the two probability distributions p and q,

$$m = \frac{1}{2}\left(p + q\right).$$

References

Claude Shannon (1948)

"A Mathematical Theory of Communication".

Bell System Technical Journal. 27(3):379–423.

Thomas M. Cover, and Joy A. Thomas (2006)

"Elements of Information Theory, 2nd Edition".

Wiley-Interscience.

David MacKay (2003)

"Information Theory, Inference and Learning Algorithms".

Cambridge University Press.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville (2016)

"Deep Learning".

MIT Press.

Q&A

Thank you!

edgar.roman@itam.mx