laSalle
Universitat Ramon Llull

Concepts of Entropy

Xavier Vilasís

Master Universitario en Data Science

Measures of Entropy

Shannon's Entropy

The gold standard in information theory. Measures the expected value of the information content.

$$H(X) = -\sum_i p(x_i) \log_2 p(x_i)$$

Gini Index

A measure of impurity often used in Decision Trees (CART). It approximates Shannon entropy linearly.

$$G = 1 - \sum_i p(x_i)^2$$

Rényi Entropy

Generalization of Shannon's entropy. The parameter α tunes sensitivity to rare events.

$$H_\alpha = \frac{1}{1-\alpha} \log \sum_i p_i^\alpha$$

Shannon's Source Coding Theorem

The Fundamental Limit of Data Compression

The Theorem

Shannon proved that the Entropy H(X)$ is the absolute mathematical limit for lossless compression. For an optimal code with average length L:

$$H(X) \leq L < H(X) + 1$$

You cannot compress data below the entropy without losing information.

Concrete Example (Huffman)
Let $X \in \{A, B, C, D\}$ with probabilities $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8})$.
Entropy: 1.75 bits.
Code: A: `0`, B: `10`, C: `110`, D: `111`.
Avg Length: $1(0.5) + 2(0.25) + 3(0.125) \times 2 = \mathbf{1.75}$.
Efficiency is 100%.

The Proof Mechanics

The proof relies on the **Asymptotic Equipartition Property (AEP)**, often called the "Law of Large Numbers" for Information Theory.

1. **The Typical Set ($A_\epsilon^{(n)}$):** For a large sequence length n, outcomes fall into two groups. The "Typical Set" contains nearly 100% of the probability mass.
2. **The Atypical Set:** Contains rare, weird sequences. We can essentially ignore them with negligible error.
3. **Counting Strategy:** We only need to assign unique binary codes to the sequences in the Typical Set.

Size of Typical Set:

$$|A_\epsilon^{(n)}| \approx 2^{nH(X)}$$

More Shannon Entropy

- Shannon's Entropy for continous distributions

$$S(X) = - < \ln f(x) > = - \int_{-\infty}^{\infty} dx \; f(x) \ln f(x)$$

- Variations on the logarithm type, but core idea is the same.
- Gaussian Random Variable

$$S(X) = \frac{1}{2} + \frac{1}{2} \ln \left(2\pi\sigma^2 \right) .$$

Maximum Entropy Principle

- The maximum entropy principle suggests that when we have incomplete knowledge, the probability distribution that best represents the state of knowledge is the one with the highest entropy
- The uniform distribution maximizes entropy because it is the least biased

Maximum Entropy Distributions

Distribution Name	Probability density/mass function	Maximum Entropy Constraint	Support
Uniform (discrete)	f(k)=1b−a+1	None	{a,a+1,…,b−1,b}
Uniform (continuous)	f(x)=1b−a	None	[a,b]
Bernoulli	f(k)=pk(1−p)1−k	E(k)=p	{0,1}
Geometric	f(k)=(1−p)k−1p	E(k)=1p	N\{0}={1,2,3,…}
Exponential	f(x)=λexp(−λx)	E(x)=1λ	[0,∞)
Laplace	f(x)=12bexp(−|x−μ|b)	E(|x−μ|)=b	(−∞,∞)
Asymmetric Laplace	f(x)=λe−(x−m)λs_κs_κ+1/κ(s=sgn(x−m))	E((x−m)s_κs)=1/λ	(−∞,∞)
Pareto	f(x)=αxmαxα+1	E(ln(x))=1α+ln(xm)	[xm,∞)
Normal	f(x)=12πσ2exp(−(x−μ)22σ2)	E(x)=μ,E((x−μ)2)=σ2	(−∞,∞)
von Mises	f(θ)=12πI0(κ)exp(κcos(θ−μ))	E(cosθ)=I1(κ)I0(κ)cosμ,E(sinθ)=I1(κ)I0(κ)sinμ	[0,2π)
Rayleigh	f(x)=xσ2exp(−x22σ2)	E(x2)=2σ2,E(ln(x))=ln(2σ2)−γE2	[0,∞)
Beta	f(x)=xα−1(1−x)β−1B(α,β) for 0≤x≤1	E(ln(x))=ψ(α)−ψ(α+β) E(ln(1−x))=ψ(β)−ψ(α+β)	[0,1]
Cauchy	f(x)=1π(1+x2)	E(ln(1+x2))=2ln2	(−∞,∞)
Chi	f(x)=22k2Γ(k/2)xk−1exp(−x22)	E(x2)=k,E(ln(x))=12[ψ(k2)+ln(2)]	[0,∞)
Chi-squared	f(x)=12k2Γ(k/2)xk2−1exp(−x2)	E(x)=k,E(ln(x))=ψ(k2)+ln(2)	[0,∞)
Erlang	f(x)=λk(k−1)!xk−1exp(−λx)	E(x)=k/λ,E(ln(x))=ψ(k)−ln(λ)	[0,∞)
Gamma	f(x)=xk−1exp(−xθ)θkΓ(k)	E(x)=kθ,E(ln(x))=ψ(k)+ln(θ)	[0,∞)
Lognormal	f(x)=1σx2πexp(−(lnx−μ)22σ2)	E(ln(x))=μ,E((ln(x)−μ)2)=σ2	[0,∞)
Maxwell–Boltzmann	f(x)=1a32πx2exp(−x22a2)	E(x2)=3a2,E(ln(x))=1+ln(a2)−γE2	[0,∞)
Weibull	f(x)=kλkxk−1exp(−xkλk)	E(xk)=λk,E(ln(x))=ln(λ)−γEk	[0,∞)
Multivariate normal	fX(x→)=exp(−12(x→−μ→)TΣ−1(x→−μ→))(2π)N/2|Σ|1/2	E(x→)=μ→,E((x→−μ→)(x→−μ→)T)=Σ	Rn
Binomial	f(k)=(nk)pk(1−p)n−k	E(x)=μ,f∈n-generalized binomial distribution[p, q]	{0,…,n}
Poisson	f(k)=λkexp(−λ)k!	E(x)=λ,f∈∞-generalized binomial distribution[p, q]	NU{0}

Joint Entropy

The *joint entropy* of two discrete random variables *X* and *Y* is merely the entropy of their pairing: (*X*, *Y*). This implies that if *X* and *Y* are independent, then their joint entropy is the sum of their individual entropies.

$$h(X,Y) = -\int_{\mathcal{D}} dxdyf(x,y)\ln f(x,y)$$

Ex: Gaussian random variables with same sigma and zero mean

$$h(x,y) = 1 + \frac{1}{2}\ln\left(4\pi^2\sigma^2\sqrt{1-\rho^2}\right)$$

Conditional Entropy

The *conditional entropy* or *conditional uncertainty* of *X* given random variable *Y* (also called the *equivocation* of *X* about *Y*) is the average conditional entropy over *Y*

$$h(X \mid Y) = -\int_{\mathcal{D}} dxdy f(x,y) \ln f(x \mid y)$$

$$h(X \mid Y) = -\int_{\mathcal{D}} dxdy f(x,y) \ln \frac{f(x,y)}{f_Y(y)}$$

$$h(X \mid Y) = h(X,Y) - h(Y)$$

Example : the rain and the umbrella

The entropy of "carrying an umbrella" is high (I might or might not).

But given "it is raining" (Condition), the uncertainty drops near zero—I will carry it.

Kullback-Leiber divergence – Relative Entropy

The *Kullback–Leibler divergence* (or *information divergence*, *information gain*, or *relative entropy*) is a way of comparing two distributions: a "true" probability distribution *f(X)*, and an arbitrary probability distribution *g(X)*. It measures the difference in information when we assume *g(X)* is the distribution underlying some data, when, in reality, *f(X)* is the correct distribution

$$D(f\|g) = \int_{\mathcal{D}} dx f(x) \ln \frac{f(x)}{g(x)}$$

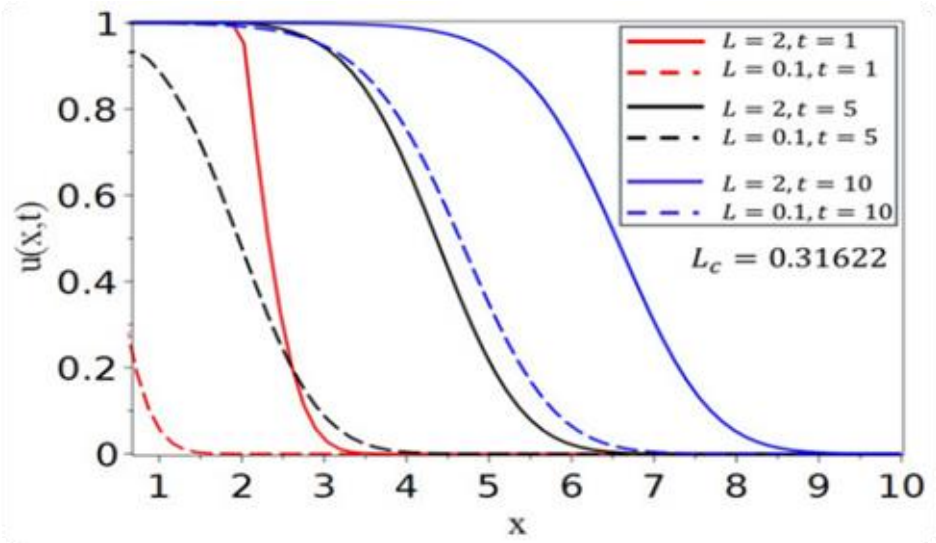
Ex: two gaussian variables with same mu and one with variance 1,

$$D(f\|g) \sim \ln(\ln(\sigma^2) + \frac{1}{\sigma^2} - 1)$$

Fisher Information

The Fisher information is a way of measuring the amount of information that an observable random variable *X* carries about an unknown parameter ϑ upon which the probability of *X* depends. Can be stated as the 'Curvature of Likelihood'. It may be related to the Kullback-Leiber divergence for two values of parameter ϑ.

$$\mathcal{I}(\vartheta) = \int_{\mathcal{D}} dx \left(\frac{\partial}{\partial \vartheta} \ln f(x;\vartheta) \right)^2 f(x;\vartheta)$$



Mutual Information

Mutual information measures the amount of information that can be obtained about one random variable by observing another.

$$I(X,Y) = \int_{\mathcal{D}} f(x,y) \ln \frac{f(x,y)}{f(x)f(y)}$$

$$I(X,Y) = H(X) + H(Y) - H(X,Y)$$

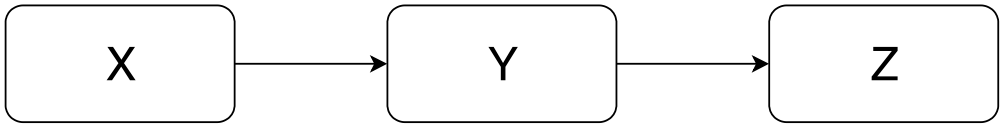
$$I(X,Y) = D(f(x,y) \| f(x)f(y))$$

Mutual Information

Ex : two gaussian random variables with zero mean and variance equal to 1

$$I(X,Y) = -\frac{1}{2}\ln(1-\rho^2)$$

Data processing inequality

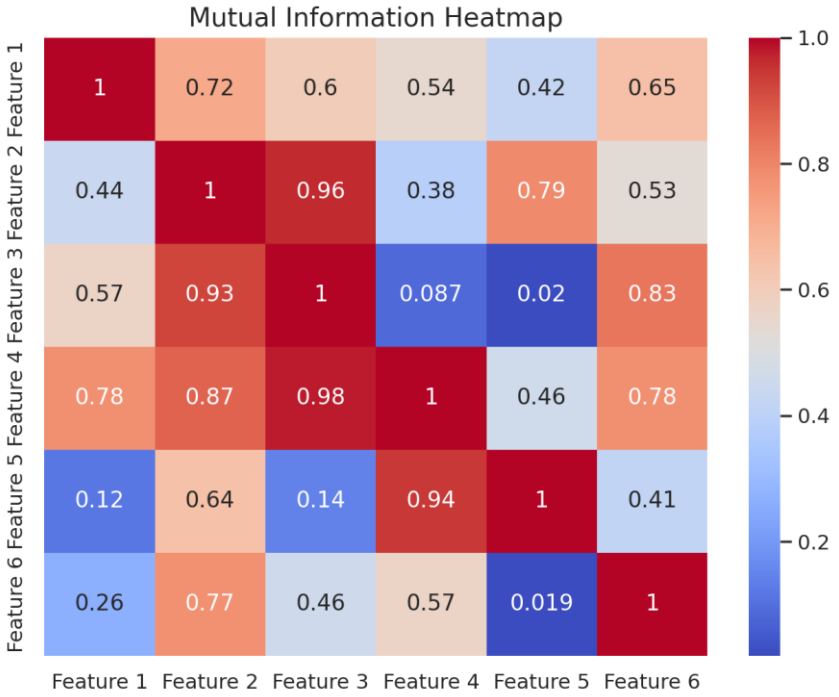


$$I(X,Z) \leq I(X;Y)$$

Mutual Information

- Measures how much information is shared between two variables.
- *M*easures the amount of information that can be obtained about one random variable by observing another.
- If X and Y are completely independent:

$$I(X;Y)=0$$



Cross-Entropy

the **cross entropy** between two probability distributions and over the same underlying set of events measures the average number of bits needed to identify an event drawn from the set, if a coding scheme is used that is optimized for an "artificial" probability distribution , rather than the "true" distribution.

$$H(f, g) = H(f) + D(f \| g)$$

$$H(f, g) = -\int_{\mathcal{D}} dx f(x) \ln g(x)$$

MI in Machine Learning

- **Feature Selection:** Mutual Information (MI) is used to identify features most relevant to the target variable.

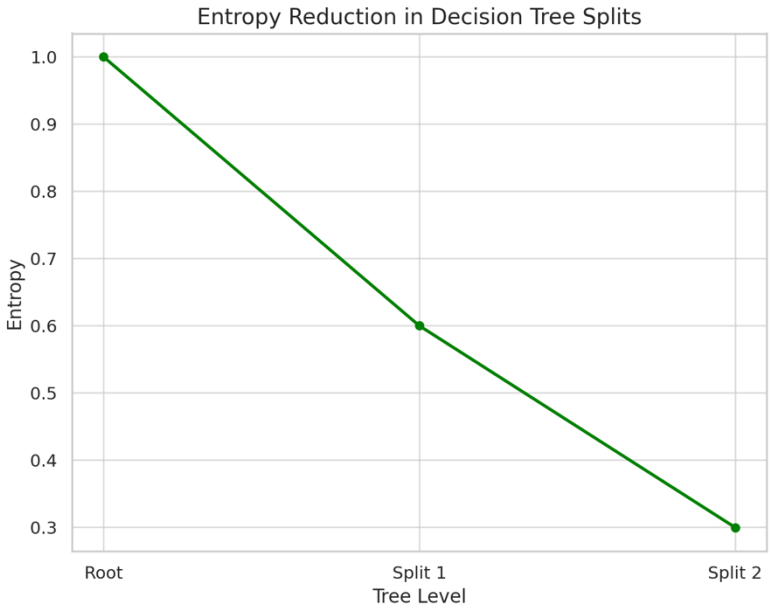
Predicting customer churn

MI(Age, Churn) = 0.1
MI(Usage Time, Churn) = 0.5

Use “Usage Time” as a predictive feature, since it shares higher mutual information with the target.

Entropy in Machine Learning

- **Decision Trees:** Entropy is used to calculate information gain, which helps to split data in a way that minimizes uncertainty.
- **Regularization:** Entropy can be used to control model complexity, reducing the risk of overfitting.

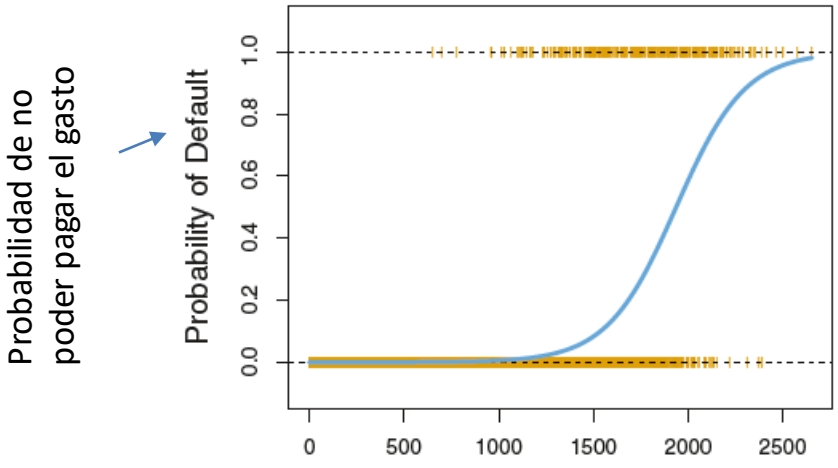


Entropy in Real-World Problems

- **Anomaly detection:** Entropy is used to detect rare events, by measuring deviations in entropy from normal patterns.
- Reinforcement learning, Clustering, Machine Learning models…

Regresión logística

- Cuando el output o predicción es cualitativo o una categoría (sí/no, 0/1, etc), hablamos de clasificación.
- En el caso de 2 categorías posibles, podemos usar una regresión logística, que dará la probabilidad de que Y pertenezca a una categoría determinada.
- Da un valor entre 0 y 1.
- (Para el ajuste se usa el método de *maximum likelihood)*.



- ¿Cómo pasar de una valor continuo obtenido en una regresión lineal a una probabilidad?

Measure	Min	Max	Name
$\Pr(Y = 1)$	0	1	"probability"
$\frac{\Pr(Y=1)}{1-\Pr(Y=1)}$	0	∞	"odds"
$\log\left(\frac{\Pr(Y=1)}{1-\Pr(Y=1)}\right)$	$-\infty$	∞	"log-odds" or "logit"

$$\log\left(\frac{\Pr(Y=1)}{\Pr(Y=0)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_r X_r$$

- Es decir, podemos obtener la probabilidad $\Pr(Y = 1)$ en función de una variable X como:

$$P(Y = 1) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} = \frac{1}{1 + e^{-\beta_0-\beta_1 x}}$$

- En el ejemplo de *Default* vs *Balance*:

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

Podemos usar los coeficientes obtenidos para predecir la probabilidad de no poder hacer frente al pago. Para alguien con un balance de 1000$:

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1{,}000}}{1 + e^{-10.6513+0.0055\times 1{,}000}} = 0.00576$$

La probabilidad es menor al 1%. Para alguien con balance de 2000$, será el 58.6%.

- Now we take p as the correct probability distribution for P(Y=1)=y, P(Y=0)=1-y and q as the estimated probability distribution ^y.

$$\hat{y} = \frac{1}{1 + e^{-\beta_0 - \beta_1 x}}$$

$$H(p, q) = -\sum_i p_i \ln q_i = -\sum_i y_i \ln \hat{y}_i + (1 - y_i) \ln(1 - \hat{y}_i)$$

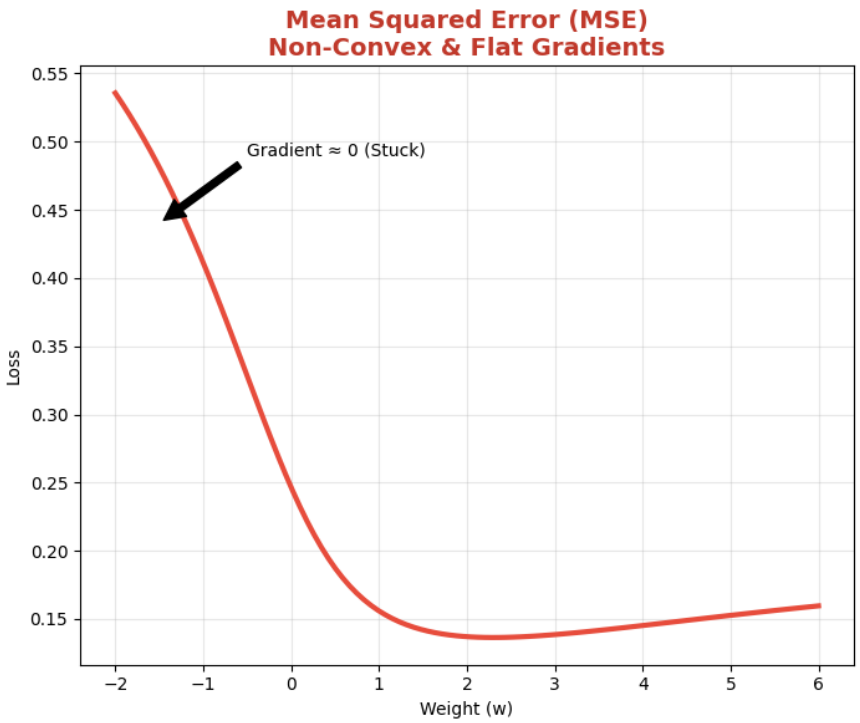
- We use the cross entropy as cost function.

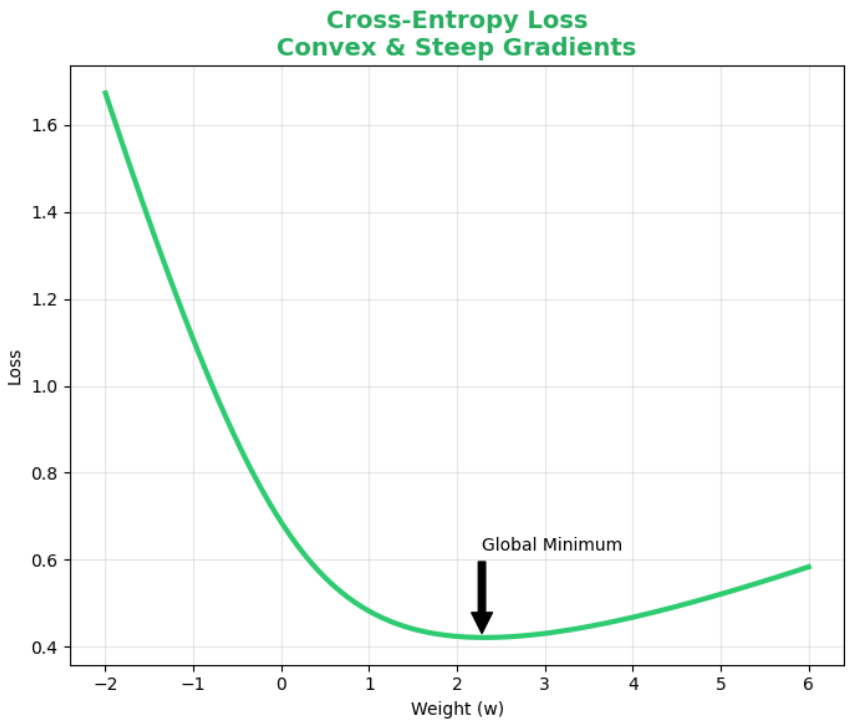
MSE gradient vanishes when $\hat{y}_i$ gets close to 0 or 1

Cross Entropy gradient is proportional to error

$$\nabla \text{MSE} \propto (y - \hat{y})\hat{y}(1 - \hat{y})x$$

$$\nabla \text{CE} \propto (y - \hat{y})x$$


Mean Squared Error (MSE)
Non-Convex & Flat Gradients
Gradient ≈ 0 (Stuck)
Loss
Weight (w)
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
−2
−1
0
1
2
3
4
5
6


Cross-Entropy Loss
Convex & Steep Gradients
Global Minimum
Loss
Weight (w)
1.6
1.4
1.2
1.0
0.8
0.6
0.4
−2
−1
0
1
2
3
4
5
6

laSalle
Universitat Ramon Llull

Concepts of Entropy

Xavier Vilasís

Master Universitario en Data Science