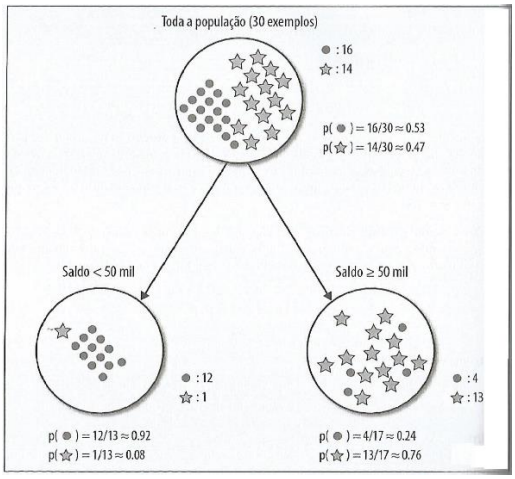
A entropia como medida de informação. Por Mario Marcondes Machado.



Foster Provost e Tom Fawcett (2016.p54)

Prezado leitor, a ideia que examinamos hoje vem do matemático e engenheiro eletrônico norte-americano Claude Elwood Shannon (1916 - 2001), que é considerado "o pai da teoria da informação".

Em um artigo de 1948 intitulado “Uma Teoria Matemática da Comunicação” *(A Mathematical Theory of Communication)*, Shannon propôs uma medida para a incerteza ou informação – a entropia.

Entre 1946 e 1953, Shannon integrou um time de renomados cientistas como John von Neumann e Norbert Wiener, tendo contribuído para a consolidação da teoria cibernética. Shannon também é creditado como fundador do computador digital quando, durante o mestrado no MIT em 1937, aos 21 anos, produziu aquela que seria considerada a dissertação mais importante do século.

Voltando à entropia, o termo vem do campo da termodinâmica, como medida do grau de irreversibilidade de um sistema. Quanto menor a chance deste voltar ao seu estado original, maior será sua entropia. Na Teoria da Informação, entretanto, entropia é uma forma de medir a quantidade da informação presente que flui no sistema. Isto, associado à ideia de que, quanto mais incerto é o resultado de um experimento aleatório, maior é a informação que se obtém ao observar a sua ocorrência. Em outras palavras, quanto menos informações sobre um sistema, maior será sua entropia.

Dentre as aplicações baseadas nos conceitos estabelecidos por Shannon, vamos tratar de uma muito utilizado atualmente em ciência de dados.

O critério do ganho de informação

De acordo com Provost e Fawcett (2016), no campo da modelagem preditiva de dados, a segmentação supervisionada traz a ideia de saber como segmentar a população (amostra) em grupos que diferem uns dos outros no que diz respeito a alguma quantidade de interesse (alvo). Nestes processos, um critério bastante utilizado é o critério do ganho de informação.

A ideia é saber se houve ganho ou perda de informação numa determinada segmentação e assim medir o desempenho do processo e aprimorá-lo, por exemplo, selecionando as variáveis (atributos) mais importantes e informativos das entidades descritas pelos dados. Este critério utiliza o conceito de entropia (ou pureza) proposto por Shannon, que funciona conforme segue:

AVISO: Os cálculos a seguir exigem mais do que uma calculadora do celular.

Ao segmentar um conjunto de clientes de uma organização financeira, por exemplo, em termos de decidir sobre uma concessão de crédito, considere que temos uma série de propriedades dos elementos do conjunto “clientes”. Estas propriedades corresponderão aos valores da variável alvo (e.g., cancelamento de crédito: *Sim, Não*). A entropia aqui, corresponde a quão misto (impuro) o segmento obtido é, com relação às propriedades de interesse. Assim, por exemplo, um segmento misto com muitos *cancelamentos de crédito* e muitos *não*

cancelamentos de crédito teria entropia alta (muito impuro). Matematicamente: a entropia de Shannon é definida como:

$$\textbf{entropia} = -\,p_1 \log_2(p_1) - p_2 \log_2(p_2) - \ldots$$

Onde cada p_i é a probabilidade (porcentagem relativa) da propriedade *i* dentro do conjunto, que varia de $p_i = 1$, quando todos os elementos do conjunto têm a propriedade *i*, e $p_i = 0$, quando nenhum elemento tem a propriedade *i*. Pode haver mais do que duas propriedades, por isso as reticências. O logaritmo é tomado na base 2.

Nesta altura, o leitor deve estar se perguntando, mas cadê o ganho de informação?

Vamos lá! A divisão em classes requer um atributo informativo (um limiar adequado). Por exemplo, saldo médio = 50 mil reais. Usando a entropia de Shannon, podemos definir o ganho de informação (GI) para medir se o atributo selecionado melhora (i.e., diminui a entropia) ao longo de toda a segmentação que ele cria. Assim, o ganho de informação é uma função do conjunto original (pai) e dos conjuntos (filhos) resultantes da divisão pelo atributo. Logo, a quantidade de informação que o atributo pode fornecer, depende de quão puro são os filhos em relação ao pai. A definição de ganho de informação (GI) é (Provost e Fawcett 2016):

$$\textbf{GI(pai, filhos)} = \textbf{entropia(pai)} - [p(c_1) \times \textbf{entropia}(c_1) + p(c_2) \times \textbf{entropia}\,(c_2) + \ldots]$$

Num problema de duas classes (● e ☆), onde o ponto representa *cancelamento de crédito* e estrela o *não cancelamento*, considere que o conjunto pai tem 30 elementos, contendo 16 ● e 14 ☆, assim:

$$\textbf{entropia(pai)} = -\,[p(\bullet) \times \log_2 p(\bullet) + p(\star) \times \log_2 p(\star)]$$

$$= -\,[(16/30) \times \log_2(16/30) + (14/30) \times \log_2(14/30)]$$

$$= -\,[0{,}53 \times (-\,0{,}9) + 0{,}47 \times (-\,1{,}1)]$$

$$= 0{,}99 \text{ (muito impura)}$$

A presente segmentação resultou em: (12 ● e 1 ☆) à esquerda (< 50mil), e (4 ● e 13 ☆) à direita (≥ 50mil). Usando o mesmo procedimento acima, a entropia dos filhos fica:

$$\textbf{entropia à esquerda (Saldo< 50mil)} = -\,[(12/13) \times -\log_2(12/13) + (4/13) \times -\log_2(4/13)]$$

$$= -\,[0{,}92 \times (-\,0{,}12) + 0{,}08 \times (-\,3{,}7)] = 0{,}39$$

$$\textbf{entropia à direita (Saldo≥ 50mil)} = -\,[(4/17) \times -\log_2(4/17) + (13/17) \times -\log_2(13/17)]$$

$$= -\,[0{,}24 \times (-\,2{,}1) + 0{,}76 \times (-\,0{,}39)] = 0{,}79$$

$$\textbf{entropia(filhos)} = (13/30) \times 0{,}39 + (17/30) \times 0{,}79 = 0{,}62$$

Portanto, o ganho de informação, GI = entropia(pai) – entropia(filhos) foi de:

$$\textbf{GI} = 0{,}99 - 0{,}62 = 0{,}37$$

Aproveito para recomendar um livro. Além de postagens da internet e outras fontes, este artigo se baseia no livro "Data Science para Negócios: o que você precisa saber sobre mineração de dados e pensamento analítico de dados", de Foster Provost e Tom Fawcett (2016).

Concluindo

Usando a entropia de Shannon, é possível avaliar o desempenho dos algoritmos de classificação e/ou recomendação tão frequentes hoje em dia.

Outro conceito estabelecido por Shannon é a capacidade de transmissão de um canal de comunicação, considerando a quantidade de *BITs (Binary Digit)* – expressão de uma unidade de informação cunhada por ele –, necessários para conter todos os valores ou significados de uma mensagem, mas isto fica para a próxima coluna. Até breve!