Entropie croisée

Introduction à l'apprentissage automatique
Master Sciences Cognitives
Aix Marseille Université

Alexis Nasr

Surprise

- Soit e un des événements pourvant survenir à l'issue d'une expérience.
- A quel point est on surpris d'apprendre que e a eu lieu ?
- On est d'autant plus surpris que la probabilité de e est faible.
- On définit la **surprise** correspondant à la probabilité p de la manière suivante :

$$S(p) = -\log_2(p)$$

$$S(1) = 0 \qquad S(0) = \infty \qquad S(0.5) = 1$$

Entropie

- Soit X une variable aléatoire discrète de distribution p : $P(X = i) = p_i$.
- On appelle **entropie** de X la moyenne des surprises des valeurs de $S(p_i)$ [1] :

$$\begin{aligned} H(X) &= \sum_{i=1}^{n} p_i S(p_i) \\ &= -\sum_{i=1}^{n} p_i \log_2(p_i) \end{aligned}$$

- L'entropie est exprimée en *bits*, elle correspond à la taille minimale du message qu'il faut pour transmettre la valeur de la variable aléatoire.
- Si $p_1 = p_2 = 0.5$, alors $H(p) = 1$, il faut un bit pour transmettre l'information.

1. On considère que $0 \times \log_2(0) = 0$

Codage

- Exemple : on lance une pièce équilibrée, il faut un bit pour transmettre le résultat du lancer (par exemple *pile* $\rightarrow 0$ et *face* $\rightarrow 1$)
- Exemple : on lance un dé à 8 faces, $p_1 = p_2 = \ldots = p_8 = \frac{1}{8}$

$$
\begin{aligned}
H(X) &= -\sum_{i=1}^{8} \frac{1}{8} \log_2(\frac{1}{8}) \\
&= -\sum_{i=1}^{8} \frac{1}{8} (\log_2(1) - \log_2(8)) \\
&= -8 \times \frac{1}{8}(0 - 3) \\
&= 3
\end{aligned}
$$

- Il faut trois bits pour transmettre l'information :

1	2	3	4	5	6	7	8
000	001	010	011	100	101	110	111

Codage

- Toujours huit valeurs mais une distribution p non uniforme :

i	1	2	3	4	5	6	7	8
p_i	0.35	0.35	0.10	0.10	0.04	0.04	0.01	0.01

- $H(p) = 2.23$ bits
- Mais pour transmettre 8 valeurs, il faut 3 bits!!
- On change le code

i	1	2	3	4	5	6	7	8
p_i	0.35	0.35	0.10	0.10	0.04	0.04	0.01	0.01
ancien code	000	001	010	011	100	101	110	111
nouveau code	00	01	100	101	1100	1101	11100	11101

- longueur moyenne du message :
  $(2 \times 0.35 \times 2) + (2 \times 0.1 \times 3) + (2 \times 0.04 \times 4) + (2 \times 0.01 \times 5) = 2.42$
- On ne pourra pas faire mieux que 2.23, atteint pour le code optimal.

Codage

- Si on change les probabilités et que l'on garde le même code, ça peut être la catastrophe !

i	1	2	3	4	5	6	7	8
p_i	0.35	0.35	0.10	0.10	0.04	0.04	0.01	0.01
q_i	0.01	0.01	0.04	0.04	0.10	0.10	0.35	0.35
code	00	01	100	101	1100	1101	11100	11101

- longueur moyenne du message : 4.58 bits !!
- Le code fait l'hypothèse implicite sur la distribution de probabilité, qu'une valeur codée sur 3 bits a un probabilité de $\frac{1}{8}$
- Il y a deux distributions de probabilités
  - La *vraie* distribution : p
  - La distribution *prédite* : q

Entropie croisée

- On définit l'entropie croisée de la façon suivante :

$$H(p,q) = -\sum_i p_i \log_2(q_i)$$

- Si $p = q$, alors $H(p,q) = H(p)$
- Sinon $H(p,q) > H(p)$
- La différence $H(p,q) - H(p)$ est appelée divergence de Kullback-Liebler :

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

- Elle mesure la différence entre les distributions de probabilité p et q.

Application à un problème de classification

- Soit un problème de classification à 7 classes

classe	1	2	3	4	5	6	7
vraie distribution	0	1	0	0	0	0	0
distribution prédite	0.02	0.30	0.45	0	0.25	0.05	0

- Entropie croisée :

$$H(p,q) = -\sum_{i=1}^{7} p_i \log_2(q_i) = -\log_2(0.3) = 1.7369$$

- plus la valeur de q_2 est élevée, plus l'entropie croisée est faible