

Información mutua y métodos contrastivos en el aprendizaje de representaciones

Doble Grado en Ingeniería Informática y Matemáticas

Francisco Javier Sáez Maldonado

9 de septiembre de 2021

Trabajo Fin de Grado

*E.T.S. de Ingenierías Informática y de Telecomunicación
Facultad de Ciencias*



**UNIVERSIDAD
DE GRANADA**

Teoría de la información

- Entropía

- Información mutua

- Cotas inferiores

Aprendizaje contrastivo

- Estimación del ruido contrastiva

- Contrastive predictive coding

- Pérdida usando tripletas

Marcos de trabajo

- SimCLR

- Bootstrap your own latent

Experimentación

- Objetivos

- Experimentos con SimCLR

- Experimentos con BYOL

Definición (Representación)

Sea $x \in \mathbb{R}^d$ un vector de entrada a un modelo de aprendizaje automático. Una representación $\tilde{x} \in \mathbb{R}^n$ es otro vector de menor dimensión que comparte información o características con x .

Definición (Divergencia Kullback-Leibler)

Sean P y Q dos distribuciones de probabilidad sobre el mismo espacio probabilístico, su divergencia de Kullback-Leibler $D_{KL}(Q \parallel P)$ mide la “diferencia” de Q a P

$$D_{KL}(P \parallel Q) = E_P \log \frac{P(x)}{Q(x)}.$$

La divergencia de Kullback-Leibler es siempre no negativa.

Teoría de la información

Sean X, Y variables aleatorias discretas, con imágenes \mathcal{X}, \mathcal{Y} .

Definición (Entropía y entropía relativa)

La entropía $H(X)$ de X se define como

$$H(X) = E_X \left[\log \frac{1}{P_X(X)} \right] = \sum_{x \in \mathcal{X}} P_X(x) \log \frac{1}{P_X(x)}.$$

Definición

La entropía condicionada $H(X | Y)$ se define como

$$H(X | Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} P_{XY}(x, y) \log \frac{P_Y(y)}{P_{XY}(x, y)}.$$

Información mutua

- $0 \leq H(X) \leq \log(|\mathcal{X}|)$
- $H(X|Y) \leq H(X)$

Definición (Información mutua)

Sean X, Z variables aleatorias. La información mutua entre ellas se expresa como

$$I(X, Z) = H(X) - H(X | Z).$$

Usando la definición y la divergencia de Kullback-Leibler entre las distribuciones P_X, P_Z asociadas a cada variable, se obtiene que

$$I(X, Z) = D_{KL}(P_{XZ} \parallel P_X P_Z)$$

Proposición (Cota inferior variacional)

Sean X, Z variables aleatorias y $Q_\theta(Z | X)$ una distribución de probabilidad arbitraria. Entonces,

$$I(X, Z) \geq H(Z) + E_{P_X} \left[E_{P_{X|Z}} [\log Q_\theta(Z | X)] \right]$$

En el contexto del aprendizaje automático, podemos considerar que Q_θ es una red neuronal y maximizar el lado derecho de la desigualdad para acercarnos al valor real de la información mutua.

Teorema (Representación Donsker-Varadhan)

La divergencia de Kullback-Leibler entre las distribuciones P y Q también puede expresarse como

$$D_{KL}(P \parallel Q) = \sup_T E_P[T] - \log E_Q \left[e^T \right],$$

donde el supremo se toma sobre todas las funciones $T : \Omega \rightarrow \mathbb{R}$ que hacen que la esperanza bajo P exista.

Corolario

Sea \mathcal{F} una clase de funciones $T : \Omega \rightarrow \mathbb{R}$ que satisfacen las condiciones del teorema anterior. Entonces:

$$I(P, Q) = D_{KL}(P \parallel Q) \geq \sup_{T \in \mathcal{F}} E_P[T] - \log E_Q \left[e^T \right]$$

Aprendizaje contrastivo

Estimación del ruido contrastiva - Problema

Consideramos:

- $X = \{x_1, \dots, x_{T_d}\}$ una muestra que suponemos extraída de una distribución $P_d \in \{P_m(\cdot; \theta)\}_\theta$.
- $Y = \{y_1, \dots, y_{T_n}\}$ una muestra de elementos idénticamente distribuidos, que asumimos extraída de una distribución de ruido conocida P_n .

Problema: Considerando el conjunto $U = X \cup Y = \{u_1, \dots, u_{T_d+T_n}\}$, ser capaces de discriminar entre elementos de U que fueron extraídos de P_d y elementos extraídos de P_n .

Trataremos de estimar el ratio P_d/P_n y usaremos este ratio para conocer propiedades sobre la distribución P_d .

Estimación del ruido contrastiva - Resolución

Asignando a cada elemento de U una *etiqueta* para poder aplicar la regresión logística

$$C_t(u_t) = \begin{cases} 1 & \text{if } u_t \in X \\ 0 & \text{if } u_t \in Y \end{cases} \implies \begin{cases} P(u \mid C = 1, \theta) = P_m(u; \theta) \\ P(u \mid C = 0) = P_n(u) \end{cases}$$

Sabemos además que, si $\nu = P(C = 0)/P(C = 1)$, las probabilidades posteriores son

$$P(C = 1|u; \theta) = \frac{P_m(u; \theta)}{P_m(u; \theta) + \nu P_n(u)} \quad , \quad P(C = 0|u; \theta) = \frac{\nu P_n(u)}{P_m(u; \theta) + \nu P_n(u)}.$$

Llamaremos $G(u; \theta)$ al logaritmo del ratio que queremos estimar

$$G(u; \theta) = \log \frac{P_m(u; \theta)}{P_n(u)}$$

Estimación del ruido contrastiva - Resolución (2)

Proposición

En las condiciones presentadas y llamando $h(u; \theta) := P(C = 1|u; \theta)$, se tiene que

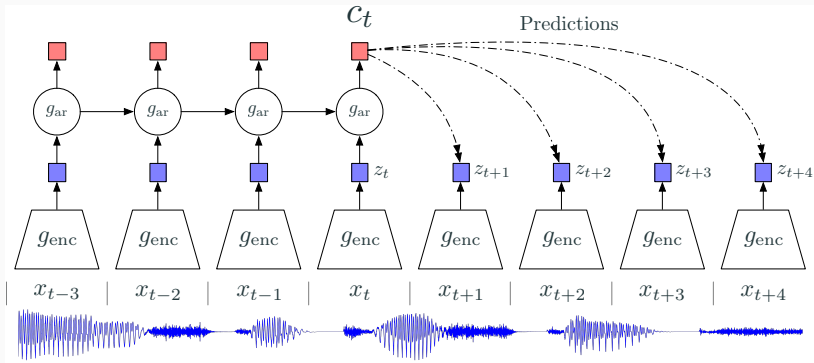
$$h(u; \theta) = r_\nu(G(u; \theta)), \quad \text{donde} \quad r_\nu(u) = \frac{1}{1 + \nu \exp(-u)}$$

es la función logística parametrizada por ν .

Puesto que las etiquetas C_t siguen una distribución de Bernoulli y son independientes, el logaritmo de la verosimilitud condicional tiene la forma

$$\begin{aligned} \ell(\theta) &= \sum_{t=1}^{T_d+T_n} C_t \log P(C_t = 1|u_t; \theta) + (1 - C_t) \log P(C_t = 0|u_t; \theta) \\ &= \sum_{t=1}^{T_d} \log[h(x_t; \theta)] + \sum_{t=1}^{T_n} \log[1 - h(y_t, \theta)] \end{aligned}$$

Contrastive predictive coding



- x_{t_j} es la entrada a la red en el instante de tiempo t_j .
- g_{enc} es un codificador que produce una representación $g_{enc}(x_t) = z_t$
- g_{ar} es un modelo autorregresivo que produce una representación que tiene en cuenta el contexto $g_{ar}(z_{\leq t}) = c_t$

Pérdida contrastiva y cota inferior contrastiva

Definición (Pérdida contrastiva)

Sea $X = \{x^*, x_1, \dots, x_{N-1}\}$ un conjunto de N ejemplos donde x^* ha sido extraído de la distribución conjunta $P(x, z)$ y le resto han sido extraídos del producto de las distribuciones marginales $P(x), P(z)$. Se define entonces la función de pérdida contrastiva como

$$\ell(\theta) = -E_X \left[\log \frac{h_\theta(x^*, z)}{\sum_{x \in X} h_\theta(x, z)} \right].$$

Proposición

En las mismas condiciones que en la definición anterior, se tiene que

$$I(x^*, z) \geq -\ell(\theta) + \log N$$

Funciones de pérdida usando tripletas



Original x



Ejemplo positivo x^+



Ejemplo negativo x^-

$$\|g(x) - g(x^+)\|_2 + \alpha < \|g(x) - g(x^-)\|_2.$$

$$\ell^\alpha(x, x^+, x^-) = \max\left(0, \|g(x) - g(x^+)\|_2^2 - \|g(x) - g(x^-)\|_2^2 + \alpha\right).$$

Definición

Dado un conjunto de tripletas, cada una con una imagen original, un ejemplo positivo y uno negativo $\mathcal{T} = \{(x_i, x_i^+, x_i^-)\}_{i \in \Lambda}$, una función de pérdida usando tripletas se define como:

$$\mathcal{L}(x_i, x_i^+, x_i^-) = \sum_{i \in \Lambda} \ell^\alpha(x_i, x_i^+, x_i^-).$$

