Información mutua y métodos contrastivos en el aprendizaje de representaciones

Doble Grado en Ingeniería Informática y Matemáticas

Francisco Javier Sáez Maldonado

10 de septiembre de 2021

Trabajo Fin de Grado

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

UNIVERSIDAD DE GRANADA

Índice

Motivación

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.

Teoría de la información

Divergencia Kullback-Leibler

Definición (Divergencia Kullback-Leibler)

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

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

La divergencia de Kullback-Leibler es siempre no negativa.

Entropía

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

Definición (Entropía)

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

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

Entropía

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

Definición (Entropía)

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 (Entropía relativa)

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

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

Información mutua

Propiedades:

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

Información mutua

Propiedades:

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

Definición (Información mutua)

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

$$I(X, Z) = H(X) - H(X \mid Z).$$

También se puede expresar como

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

Cotas inferiores de la información mutua

Proposición (Cota inferior variacional)

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

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

En el contexto del aprendizaje automático, podemos considerar que Q_θ es una red neuronal y maximizar la cota inferior usando *backpropagation*.

Teorema (Representación Donsker-Varadhan)

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

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

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

Teorema (Representación Donsker-Varadhan)

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

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

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

Corolario

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

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

Aprendizaje contrastivo

Estimación del ruido contrastiva - Problema

Consideramos:

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

Estimación del ruido contrastiva - Problema

Consideramos:

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

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

Estimación del ruido contrastiva - Problema

Consideramos:

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

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

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

Proposición

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

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

es la función logística parametrizada por $\nu = P(C = 0)/P(C = 1)$.

Estimación del ruido contrastiva - Resolución

Proposición

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

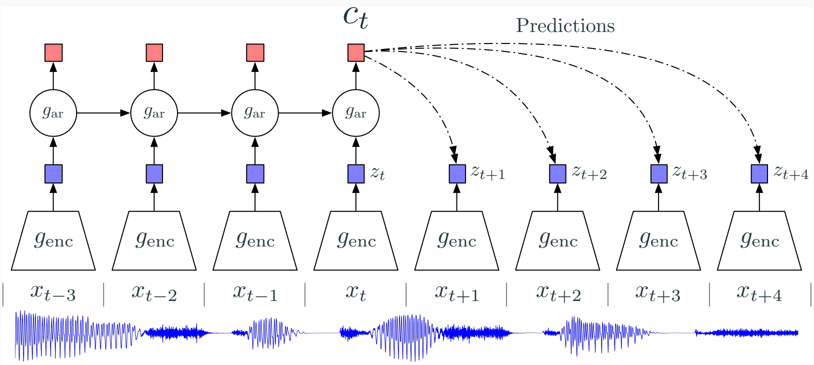
$$h(u;\theta) = r_\nu(G(u;\theta)), \quad donde \quad r_\nu(u) = \frac{1}{1 + \nu exp(-u)}$$

es la función logística parametrizada por $\nu = P(C = 0)/P(C = 1)$.

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

$$\ell(\theta) = \sum_{t=1}^{T_d} \log[h(x_t;\theta)] + \sum_{t=1}^{T_n} \log[1 - h(y_t,\theta)]$$

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 $X = \{x^*, x_1, \ldots, x_{N-1}\}$ un conjunto de N ejemplos donde x^* ha sido extraido 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

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

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 extraido 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^-

Nos interesa tener un margen $\alpha \in \mathbb{R}$

$$\left\|g(x) - g(x^+)\right\|_2 + \alpha < \left\|g(x) - g(x^-)\right\|_2 ,$$

lo que nos lleva a la pérdida de una tripleta individual

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

Funciones de pérdida usando tripletas

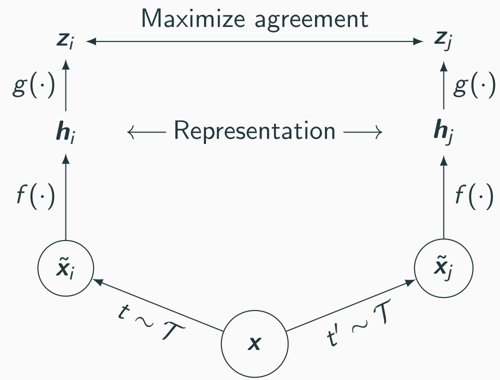
Definición

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:*

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

Podemos generalizar esta función de pérdida a un caso más eficiente y, a continuación, se puede probar que se está generalizando la pérdida contrastiva.

Nuevos marcos de trabajo



$$\ell_{i,j} = -\log \frac{exp(sim(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} I_{k\neq i} exp(sim(z_i, z_k)/\tau)},$$

Bootstrap your own latent

titulo

titulo

titulo

titulo

titulo