Dimensión efectiva en modelos de aprendizaje automático

Yackelin Sofía Giraldo Castaño
Bryan Steven Rodríguez Ruiz

Semillero Quantum Computing
Universidad Nacional de Colombia
Sede Medellín
2024-05

Contenido

Contenido

Ejemplos

Modelos económicos.

Número de componentes principales o factores significativos que capturan la mayor parte de la variabilidad de los datos en el modelo.

Este enfoque permite una simplificación significativa sin perder gran parte de la información.

Error de Generalización

Todos tienen en si la misma estructura:

$$\text{generalization error} \;\leq\; \text{estimate of error} \;+\; \text{complexity penalty}$$

Un set de entrenamiento

$$\mathcal{S}_n = \{(x_1, y_1), \ldots, (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$$

Minimiza el riesgo esperado - R(h):=$\mathrm{E}_{(x,y)\sim p}[\ell(h(x), y)]$

Minimiza el riesgo empírico - $\mathrm{R}_n(h) := \frac{1}{n}\sum_{i=1}^{n} \ell\left(h\left(x_i\right), y_i\right)$

Un límite superior en la cantidad:

$$\sup_{h\in\mathcal{H}} |R(h) - R_n(h)|$$

Matriz de Información de Fisher

Es sinónimo de capacidad del modelo.

La matriz de información de Fisher la podemos representar en la siguiente forma:

$$F(\theta) = \mathbb{E}_{(x,y)\sim p}\left[\frac{\partial}{\partial\theta}\log p(x,y;\theta)\frac{\partial}{\partial\theta}\log p(x,y;\theta)^{\top}\right] \in \mathbb{R}^{d\times d}$$

La matriz de información de Fisher es simidefinida positiva y por lo tanto sus valores propios no son negativos.

Ayuda a capturar la sensibilidad de la salida de un modelo de machine learning en relación con los movimientos en el espacio de parámetros, lo que resulta útil en la optimización del gradiente natural.

Contenido

Dimensión global efectiva

Definición

Sea $\kappa_{n,\gamma} = \frac{\gamma n}{2\pi \log n}$ y un modelo estadístico $\mathcal{M}_\Theta := \left\{p(\cdot,\cdot;\theta) : \theta \in \Theta \subset \mathbb{R}^d\right\}$,con respecto a $n \in \mathbb{N}$ y $\gamma \in \left(\frac{2\pi \log n}{n}, 1\right]$, se se define la dimensión efectiva global como:

$$d_{n,\gamma}\left(\mathcal{M}_\Theta\right) := \frac{2\log\left(\frac{1}{V_\Theta}\int_\Theta \sqrt{\det\left(\mathrm{id}_d + \kappa_{n,\gamma}\bar{F}(\theta)\right)}\mathrm{d}\theta\right)}{\log \kappa_{n,\gamma}} \qquad (1)$$

$$\bar{F}_{ij}(\theta) := d\frac{V_\Theta}{\int_\Theta \mathrm{tr}(F(\theta))\mathrm{d}\theta}F_{ij}(\theta)$$

$$V_\Theta := \int_\Theta \mathrm{d}\theta \in \mathbb{R}_+$$

Dimensión global efectiva

Continuidad

Sea $\kappa_{n,\gamma} = \frac{\gamma n}{2\pi \log n}, \gamma \in \left(\frac{2\pi \log n}{n}, 1\right]$, $n \in \mathbb{N}$ con $\mathcal{M}_\Theta$ y $\mathcal{M}'_\Theta$ dos modelos estadísticos, con matrices de información de Fisher F y F', respectivamente, se tiene:

$$\left|d_{n,\gamma}\left(\mathcal{M}_\Theta\right) - d_{n,\gamma}\left(\mathcal{M}'_\Theta\right)\right| \leq C_d \left(\frac{1}{\phi(F)} + \frac{1}{\phi(F')}\right) \text{máx}_{\theta\in\Theta} \left\|\sqrt{\bar{F}(\theta)} - \sqrt{\bar{F}'(\theta)}\right\| + \frac{2\psi(F)+2\psi(F')}{\log \kappa_{n,\gamma}}$$

$$\phi(F) := \frac{1}{V_\Theta}\int_\Theta \sqrt{\det(F(\theta))}\mathrm{d}\theta$$

$$\psi(F) = \text{máx}\left\{\log\left(\frac{1}{V_\Theta}\int_\Theta \sqrt{\det\left(\mathrm{id}_d + \bar{F}(\theta)\right)}\mathrm{d}\theta\right), -\log\left(\frac{1}{V_\Theta}\int_\Theta \sqrt{\det(\bar{F}(\theta))}\mathrm{d}\theta\right)\right\}$$

Dimensión global efectiva

Estabilidad computacional

Para lograr obtener una fórmula que se pueda computar se hará una modificación a la formula 2. Se definen

$$z(\theta) = \frac{1}{2}\sum_{i=1}^{d} \log\left(1 + \kappa_{n,\gamma}\lambda_i(\bar{F}(\theta))\right) =: z(\theta)$$

$$\zeta = \max_{\theta\in\Theta} z(\theta)$$

Tal que se tiene una formula estable para la dimensión global efectiva

$$d_{n,\gamma}\left(\mathcal{M}_\Theta\right) = \frac{2\zeta}{\log\kappa_{n,\gamma}} + \frac{2}{\log\kappa_{n,\gamma}}\log\left(\frac{1}{V_\Theta}\int_\Theta \exp(z(\theta)-\zeta)\mathrm{d}\theta\right)$$

Dimensión local efectiva

Definición

Sea $\mathcal{B}_{\varepsilon}(\theta^{\star}) := \{\theta \in \Theta : \|\theta - \theta^{\star}\| \leq \varepsilon\}$ una bola centrada en θ^{*} de tamaño ε , $\kappa_{n,\gamma} = \frac{\gamma n}{2\pi \log n}$ y un modelo estadístico $\mathcal{M}_{\Theta} := \left\{p(\cdot,\cdot;\theta) : \theta \in \Theta \subset \mathbb{R}^{d}\right\}$, alrededor de $\theta^{*} \in \Theta$ con respecto a $n \in \mathbb{N}$ y $\gamma \in \left(\frac{2\pi \log n}{n}, 1\right]$ y $\varepsilon > 1/\sqrt{n}$ se define la dimensión local efectiva como:

$$d_{n,\gamma}\left(\mathcal{M}_{\mathcal{B}_{c}(\theta^{*})}\right) = \frac{2\log\left(\frac{1}{V_{c}}\int_{\mathcal{B}_{c}(\theta^{*})}\sqrt{\det\left(\mathrm{id}_{d} + \kappa_{n,\gamma}\bar{F}(\theta)\right)}\mathrm{d}\theta\right)}{\log \kappa_{n,\gamma}} \quad (2)$$

$$\bar{F}_{ij}(\theta) := d\frac{V_{\varepsilon}}{\int_{\mathcal{B}_{\varepsilon}(\theta^{*})}\mathrm{tr}(F(\theta))\mathrm{d}\theta}F_{ij}(\theta)$$

$$V_{\varepsilon} := \int_{\mathcal{B}_{\varepsilon}(\theta^{*})}\mathrm{d}\theta = \frac{\pi^{d/2}\varepsilon^{d}}{\Gamma(d/2+1)} \in \mathbb{R}_{+}$$

Generalización del error y dimensión local efectiva

Lo que mostrará será la relación que hay entre la dimensión local efectiva y la generalización del error de un modelo. Se evidencia como la dimensión local efectiva es una medida de la capacidad del modelo pues nos permite acotar la generalización del error del mismo.

Consideraciones generales

Sea $\Theta = [-1, 1]^d$ y considerando un modelo estadísitico $\mathcal{M}_\Theta := \{p(\cdot, \cdot; \theta) : \theta \in \Theta\}$ el cual satisface que la función de probabilidad es M_1 Lipschitz continua tal que $\bar{F}(\theta)$ tiene rango completo para todo $\theta \in \Theta$, y $\left\|\nabla_\theta \log \bar{F}(\theta)\right\| \leq \Lambda$ para algún $\Lambda \geq 0$ y todo $\theta \in \Theta$. Además, sea $\ell : \mathrm{P}(\mathcal{Y}) \times \mathrm{P}(\mathcal{Y}) \to [-B/2, B/2]$ para $B > 0$ una función de pérdida Lipschitz continua en el primer argumento con constante M_2. Entonces, existe una constante c_d tal que $\theta^\star \in \Theta, n \in \mathbb{N}, \gamma \in \left(\frac{2\pi \log n}{n}, 1\right]$, y $\varepsilon > 1/\sqrt{n}$.

Generalización del error y dimensión local efectiva

Sea $\mathcal{N}^{\mathcal{B}_\varepsilon(\theta^*)}(r)$ el número de cajas con lado de longitud r requeridas para cubrir la bola $\mathcal{B}_\varepsilon\left(\theta^\star\right)$. Bajo las consideraciones generales, antes mencionadas, se tiene que para cualquier $\xi \in (0,1)$

Lema 1

$$\mathbb{P}\left(\sup_{\theta\in\mathcal{B}_c(\theta^*)} |R(\theta) - R_n(\theta)| \geq \xi\right) \leq 2\mathcal{N}^{\mathcal{B}_c(\theta^*)}\left(\frac{\xi}{4M}\right)\exp\left(-\frac{n\xi^2}{2B^2}\right)$$

Lema 2

$$\mathcal{N}^{\mathcal{B}_\varepsilon(\theta^*)}\left(\frac{\varepsilon}{\sqrt{\kappa_{n,\gamma}}}\right) \leq c_d(1+\varepsilon\Lambda)^d \cdot \kappa_{n,\gamma}^{\frac{d_{n,\gamma,\varepsilon}}{2}}$$

Generalización del error y dimensión local efectiva

Teorema

Bajo las consideraciones generales,mencionadas anteriormente y usando los Lemas 1 y 2

$$\mathbb{P}\left(\sup_{\theta\in\mathcal{B}_\varepsilon(\theta^*)}|R(\theta)-R_n(\theta)|\geq\frac{4M\varepsilon}{\sqrt{\kappa_{n,\gamma}}}\right)\leq c_d(1+\varepsilon\Lambda)^d\cdot\kappa_{n,\gamma^2}^{\frac{d_{n,\gamma,\varepsilon}}{2}}\,\exp\left(-\frac{16\pi M^2\varepsilon^2\log n}{B^2\gamma}\right) \tag{3}$$

donde $M = M_1M_2$, $d_{n,\gamma,\varepsilon}$ es la dimensión local efectiva $d_{n,\gamma}\left(\mathcal{M}_{\mathcal{B}_\varepsilon(\theta^*)}\right)$.

Contenido

Implementaciones

Se usaron dos dataset de imagenes NMIST y CIFAR10, con ambos usaron dos tipos de experimentos:

- Variar la cantidad de parametros al aumentar la cantidad de neuronas por capa.
- Variar el porcentaje de aleatoriedad de los datos de etiquetados.

En cada variación se calculaba el error generalizado y la dimensión local efectiva.

Contenido

Conclusiones y trabajo futuro

- Al tener una cantidad de datos de entrenamiento finita, la dimensión local efectiva permite medir la capacidad de generalización del modelo, esto es importante para tener certeza que al ingresar nuevos datos el modelos entregará resultados confiables, bajo los parámetros con los cuales se entrenó.
- El tratamiento que se aplica en este trabajo es en general utilizado para tratar con modelos estadísticos, entonces abre la posibilidad de adentrarse en la generalización del error incluso en modelos de aprendizaje automático cuánticos.

Contenido

Referencias

A. Abbas, D. Sutter, C. Zoufal, A. Lucchi, A. Figalli, and S. Woerner. The power of quantum neural networks. Nature Computational, 2021.

A. Abbas, D. Sutter, A. Figalli and S. Woerner. Effective dimension of machine learning models. arXiv, 2021.