

Machine Learning

Training vs testing
Interpretación de la cota de la generalización

Angel Vázquez-Patiño
angel.vazquezp@ucuenca.edu.ec

Departamento de Ciencias de la Computación
Universidad de Cuenca

7 de noviembre de 2017

Objetivo

- Entender las implicaciones prácticas de la dvc

Contenido

Complejidad de la muestra
Penalización para la complejidad del modelo
El test set

Cota de la generalización

- Muy útil en la práctica
- Regla (práctica)
- N al menos $10d_{vc}$ para una generalización decente

Complejidad de la muestra

Complejidad de la muestra

- Cuántos ejemplos de entrenamiento N se necesitan para obtener un cierto rendimiento de generalización
- ϵ , tolerancia del error, error de generalización permitida
- δ , parámetro de confianza, qué tan a menudo el error de tolerancia ϵ es violado

$$N \geq \frac{8}{\epsilon^2} \ln \left(\frac{4m_{\mathcal{H}}(2N)}{\delta} \right)$$

Complejidad de la muestra

- Error de generalización de a lo mucho ε con probabilidad de a lo mucho $1-\delta$
- Basado en la VC dimension

$$N \geq \frac{8}{\epsilon^2} \ln \left(\frac{4((2N)^{d_{vc}} + 1)}{\delta} \right)$$

- 10 veces la d_{vc}

Penalización por complejidad del modelo

Penalización por complejidad de \mathcal{H}

- En la complejidad de la muestra se fijan ε y δ y se estima el N
- En la mayoría de situaciones prácticas, se tiene un data set \mathcal{D} fijo por ende N también es fijo
- La pregunta importante en ese caso es ¿qué rendimiento se espera para ese N particular?
- Con probabilidad de al menos $1-\delta$

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \left(\frac{4m_{\mathcal{H}}(2N)}{\delta} \right)}$$

Penalización por complejidad de \mathcal{H}

- Basado en d_{vc}

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \left(\frac{4((2N)^{d_{vc}} + 1)}{\delta} \right)}$$

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \Omega(N, \mathcal{H}, \delta),$$

$$\begin{aligned} \Omega(N, \mathcal{H}, \delta) &= \sqrt{\frac{8}{N} \ln \left(\frac{4m_{\mathcal{H}}(2N)}{\delta} \right)} \\ &\leq \sqrt{\frac{8}{N} \ln \left(\frac{4((2N)^{d_{vc}} + 1)}{\delta} \right)} \end{aligned}$$

Although $\Omega(N, \mathcal{H}, \delta)$ goes up when \mathcal{H} has a higher VC dimension, E_{in} is likely to go down with a higher VC dimension as we have more choices within \mathcal{H} to fit the data. Therefore, we have a tradeoff: more complex models help E_{in} and hurt $\Omega(N, \mathcal{H}, \delta)$. The optimal model is a compromise that minimizes a combination of the two terms, as illustrated informally in Figure 2.3.

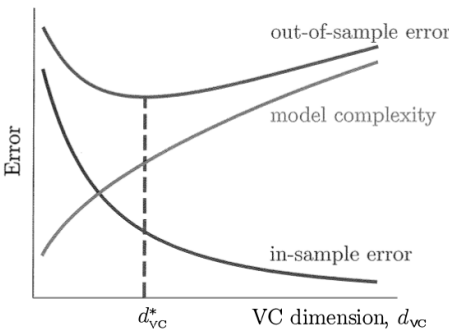


Figure 2.3: When we use a more complex learning model, one that has higher VC dimension d_{vc} , we are likely to fit the training data better resulting in a lower in sample error, but we pay a higher penalty for model complexity. A combination of the two, which estimates the out of sample error, thus attains a minimum at some intermediate d_{vc}^* .

El test set

El test set

As we have seen, the generalization bound gives us a loose estimate of the out-of-sample error E_{out} based on E_{in} . While the estimate can be useful as a guideline for the training process, it is next to useless if the goal is to get an accurate forecast of E_{out} . If you are developing a system for a customer, you need a more accurate estimate so that your customer knows how well the system is expected to perform.

El test set

- Etest aproxima Eout
- Mientras más grande el test set, más preciso Etest será como estimación de Eout
- Test set no tiene sesgo (bias)
- Train y test sets tiene muestras finitas que tendrán cierta varianza de acuerdo al tamaño de la muestra
- Pero el test set no tiene un sesgo ni optimista ni pesimista (no fue utilizado en el entrenamiento)
- Tradeoff para separar ejemplos de test

Conceptos y términos importantes


Conceptos y términos importantes

- Utilidad práctica de estudiar
 - Complejidad de la muestra
 - Penalización para la complejidad del modelo
 - El test set

Referencia

- Abu-Mostafa, Y.S., Magdon-Ismail, M., Lin, H.-T., 2012. Learning from data: a short course. AMLbook.com, USA.

Preguntas



Cota de la generalización

Angel Vázquez-Patiño

19/19