Medidas de rendimiento

Minería de Datos - Grado en Matemáticas

DSLAB

2025-10-26

Medidas de rendimiento

All models are wrong but some are useful

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?".

(Box, GEP, 1979, Robustness in the strategy of scientific model building, *Robustness in Statistics*, Academic Press, pp.201-236)

Medidas de rendimiento

$$DATOS = MODELO + ERROR$$

- **Datos**. La realidad que se quiere comprender, predecir o mejorar
- **Modelo**. Representación **simplificada** de la realidad que se propone para describirla e interpretarla más fácilmente
- **Error**. Diferencia entre la representación simplificada de la realidad (modelo) y los datos (describen la realidad de forma precisa)

Una vez construido un modelo –> evaluación del rendimiento –> ¿error aceptable?

Complejidad del modelo

- Más información en modelo (más variables) –> error suele reducirse
- Más variables –> más complejo es el modelo
- ¿Esto es bueno?
  - **Principio de parsimonia**. Priorizar modelos sencillos. Navaja de Occam
  - **Pérdida de generalidad**. Si añado demasiados parámetros de entrada a un modelo puedo representar exactamente la información de los datos que tengo, pero no funcionará bien con nuevos datos –> **sobreajuste** ("*overfitting*").

Complejidad del modelo

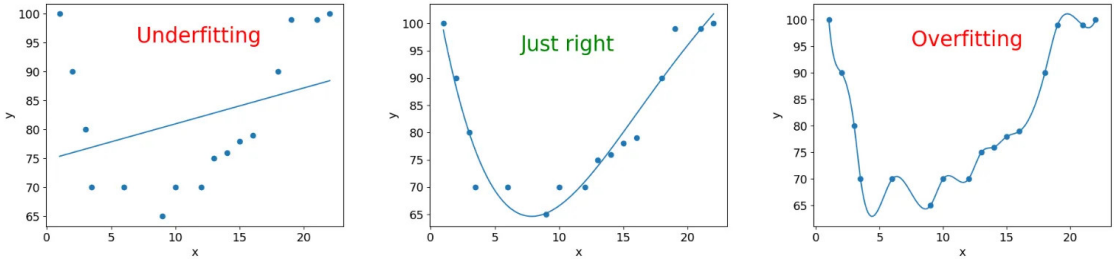

Figure 1: https://medium.com/@kiprono_65591/regularization-a-technique-used-to-prevent-over-fitting-886d5b361700

Balance sesgo-varianza

- **Sesgo**. Diferencia entre la predicción y lo real. Se refiere a la simplificación excesiva de un modelo, asumiendo que los datos de entrenamiento siguen una cierta estructura o patrón predefinido.
  - Sesgo alto –> subajusta los datos, no captura la complejidad de los datos ni representa la relación entre las variables
- **Varianza**. Sensibilidad de un modelo a las fluctuaciones en los datos de entrenamiento
  - Varianza alta –> demasiado ajuste en training –> mal rendimiento en nuevos datos

Balance sesgo-varianza

Equilibrio sesgo-varianza -> Modelos eficaces y con capacidad de generalización

- Modelo con **sesgo alto y varianza baja**: más simple y tiende a subajustar los datos
- Modelo con **sesgo bajo y varianza alta** se ajusta muy bien a los datos de entrenamiento pero generaliza mal

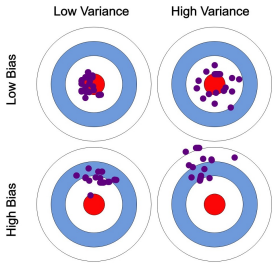

Figure 2: https://nvsyashwanth.github.io/machinelearningmaster/bias-variance/

Balance sesgo-varianza

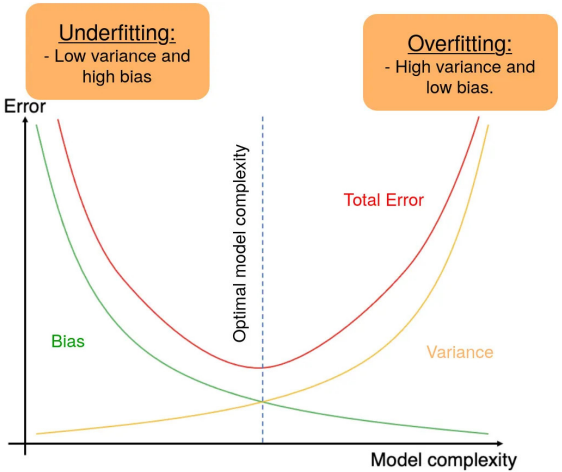

Figure 3: https://medium.com/@kiprono_65591/regularization-a-technique-used-to-prevent-over-fitting-886d5b361700

¿Cómo evitar el sobreajuste?

- Mayor tamaño muestral
- Validación cruzada
- Selección de variables (evitar variables redundantes)
- Modelos simples
- Partición de los datos

Medidas de evaluación

- Estudio del error del modelo
- Evaluación si todas las variables son útiles
- Comparación de modelos
- Comparación del error en distintos conjuntos de datos (train-test-validation) –> capacidad de generalización
- Distintas técnicas para regresión y clasificación
- **Error**: diferencia (en base a alguna medida) entre el valor observado y el predicho

$$error_i \propto target\_observado_i - target\_predicho_i$$

Medidas para regresión

- **Modelo de regresión**: técnica estadística que analiza la relación entre una variable dependiente (o de respuesta) continua y una o más variables independientes (o predictoras)
- **Error en regresión**: diferencia entre los valores predichos y los observados

Medidas para regresión

- **Error Cuadrático Medio (Mean Squared Error, MSE)**: Promedio de las diferencias al cuadrado entre las predicciones del modelo y los valores reales. Sensible a los errores grandes debido al término de cuadrado. Cuanto menor sea el MSE, mejor será el ajuste del modelo:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

donde y_i es el valor de la variable objetivo en la observación $\mathbf{x}_i$ con $i = 1, \ldots, n$, y $f(x_i)$ es la predicción del modelo de ML

- **Error Absoluto Medio (Mean Absolute Error, MAE)**: Similar al MSE, pero con el valor absoluto. Menos sensible a los errores extremos.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - f(x_i)|$$

Medidas para regresión

- **Raíz del Error Cuadrático Medio (Root Mean Squared Error, RMSE)**: Raíz cuadrada del MSE. Ofrece una medida del error en la misma unidad que la variable objetivo, lo que facilita su interpretación.
- **R-cuadrado (R-squared, R^2)**: Proporciona una medida de la proporción de la variabilidad en la variable dependiente que es explicada por el modelo. Un R^2 más alto indica un mejor ajuste del modelo a los datos, con un valor máximo de 1.

$$R^2 = \frac{\sum_{i=1}^{n}(f(x_i) - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

donde $\bar{y}$ es el valor medio de la variable objetivo

- **Error Porcentual Absoluto Medio (Mean Absolute Percentage Error, MAPE)**: Calcula el porcentaje promedio de error absoluto en relación con los valores reales.

$$MAPE = \frac{100}{n} \sum_{i=1}^{n} |\frac{y_i - f(x_i)}{y_i}|$$

Medidas para clasificación

- **Clasificación binaria**: dividir las observaciones dadas en dos clases mutuamente excluyentes $\{-1,+1\}$
- Medidas se obtienen a partir de la **matriz de confusión**:

		Valor observado	
		-1	1
Valor predicho	-1	TN	FN
	1	FP	TP

  - **TP**: "True positive", 1 clasificados como 1
  - **TN**: "True negative" -1 clasificados como -1
  - **FP**: "False positive" -1 erróneamente clasificados como 1
  - **FN**: "False negative" 1 erróneamente clasificados como -1
- Nótese que: $n = TP + FP + TN + FN$
- Importancia relativa de FP y FN. Ejemplo: control de accesos

Medidas para clasificación

- **Exactitud (Accuracy)**: Medida más común. Representa la proporción de observaciones correctamente predichas, es decir:

$$Accuracy = \frac{TP + TN}{n}$$

- **Error**: recíproco de la exactitud:

$$Error = \frac{FP + FN}{n}$$

- **Sensibilidad (recall)**: también conocida como Recuperación o Tasa de Verdaderos Positivos (TPR). Puede verse como la probabilidad de que un 1 observado sea clasificado efectivamente como 1:

$$Recall = \frac{TP}{TP + FN}$$

Medidas para clasificación

- **Especificidad (specificity)**: también conocida Tasa de Verdaderos Negativos puede verse como la probabilidad de que un -1 observado sea clasificado efectivamente como -1:

$$Specificity = \frac{TN}{TN + FP}$$

- **Precisión**: también conocida Valor Predictivo Positivo puede verse como la probabilidad de que acierto cuando se predice un valor 1:

$$Precision = \frac{TP}{TP + FP}$$

- **Valor Predictivo Negativo (NPV, "Negative Predictive Value")**: tasa de acierto cuando se predice un valor -1:

$$NPV = \frac{TN}{TN + FN}$$

Medidas para clasificación

- **F1-score**: media armónica de Precisión y Recuperación:

$$F_1 - score = 2\frac{Precision * Recall}{Precision + Recall}$$

- **F-score generalizado**: media armónica ponderada de Precisión y Recall:

$$F_\beta = (1 + \beta^2)\frac{Precision * Recall}{\beta^2 Precision + Recall}$$

Cuando $\beta = 1$, se tiene la medida anterior: $F_1 - score$. Si $\beta > 1$, se da mayor peso a la Recall que a la Precisión. Si $\beta < 1$ se da mayor peso a la Precisión que a la Recall.

Rendimiento de las particiones

- Partición Train-Test-Validation –> 3 medidas de rendimiento
- Validación cruzada en t trozos
  - t medidas de rendimiento (referentes a Train-Test) –> media y desviación típica
  - La medida de rendimiento de Validation

Ajuste hiperparámetros

- Modelo knn
- Elección del número de vecinos k: validación cruzada (t trozos) probando distintos valores de k
- Ejemplo de resultados:

Número de vecinos k	Media (Desv. Std.)
1	1,54 (0,33)
2	1,75 (0,41)
3	1,68 (0,28)
4	1,68 (0,35)
5	1,70 (0,34)
6	1,89 (0,38)
7	1,91 (0,36)
8	2,12 (0,45)
9	2,25 (0,46)
10	2,34 (0,34)
11	2,33 (0,38)

Comparación de modelos

- Modelos devuelven como salida, para cada observación, la **probabilidad de pertenencia** a las diferentes clases de la variable respuesta
- Ejemplo: knn devuelve % de tus vecinos que pertenecen a cada clase
- Esta probabilidad se **binariza** para determinar a qué clase pertenece cada observación
- ¿Con qué **umbral**? $->$ En general 0.5 (**¡Ojo! otro parámetro!!**)
- ¿Datos desbalanceados?

Comparación de modelos: Curvas ROC

- **Curva ROC ("*Receiver Operating Characteristic curve*", o curva característica de operación)**: método gráfico para ilustrar la capacidad predictiva de un modelo de ML binario

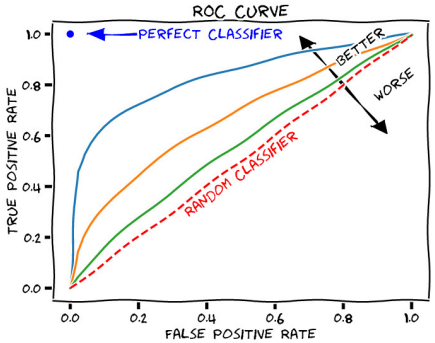

Figure 4: https://sefiks.com/2020/12/10/a-gentle-introduction-to-roc-curve-and-auc/

Comparación de modelos: Curvas ROC

Cálculo de los valores *False Positive Rate* y *True Positive Rate* de la curva ROC*:*

- Se fija un umbral para binarizar
- Se obtiene la matriz de confusión y, con ella, el valor de FPR y TPR
- Se grafica el punto
- Se repite el proceso con otro valor del umbral (en orden creciente, ej: 0.1, 0.2,..., 0.9, 1)
- Finalmente se unen todos los puntos

Comparación de modelos: Curvas ROC

- **¿Mejor solución?** (*FPR=0, TPR=1)* –> no hay errores en la clasificación
- Curva ROC sirve para elegir el mejor umbral (el más cercano al punto ideal (0,1))
- También sirve para elegir en qué modo de funcionamiento ajustar nuestro modelo
  - Modelo con muy alta recall (*TPR*), sacrificando la *FPR* (baja especificidad)
  - Modelo con baja recall y alta especificidad
  - Deseable: recall y especificidad altos

Comparación de modelos: AUC

- **Área bajo la curva (AUC, "Area Under Curve"):** medida resumen de la curva ROC
  - $AUC \approx 1$ –> mejor modelo
  - $AUC \approx 0.5$ –> predicción cercana al azar

Referencias

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1-758). New York: springer.