UNIVERSIDAD EAFIT

Científico de Datos

Aliados:

Microsoft

idata
Advanced analytics for business

Vigilada Mineducación

UNIVERSIDAD EAFIT

SELECCIÓN DE MODELOS

Aliados:

Microsoft

idata
Advanced analytics for business

Vigilada Mineducación

OBJECTIVE

- ¿Qué modelo usar?
- ¿Cuáles son los mejores hiperparámetros?
- ¿Qué métrica usar?

MODEL SELECTION

Sabiendo que tenemos disponible tantos métodos, cómo saber cuál es el mejor?

Modelos

- Regresión logística
- SVM
- K-NN
- Árboles de decisión

Propiedades del modelo

Linealidad, costo computacional, interpretabilidad, etc.

Rendimiento del modelo

Qué tanto se equivoca?

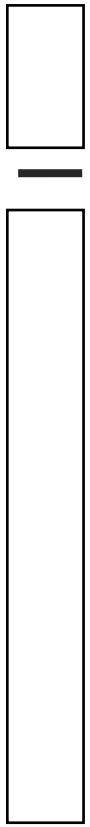
MODEL SELECTION

Sabiendo que tenemos disponible tantos métodos, con tantos hiperparámetros, cómo saber cuál es la mejor combinación de modelo e hiperparámetros?

Modelos	
Regresión logística	→ Coeficiente de regularización
SVM	→ Tipo de kernel, coeficientes del kernel
K-NN	→ Número de vecinos, métrica de distancia
Árboles de decisión	→ Criterio de impureza, profundidad

MODEL TESTING

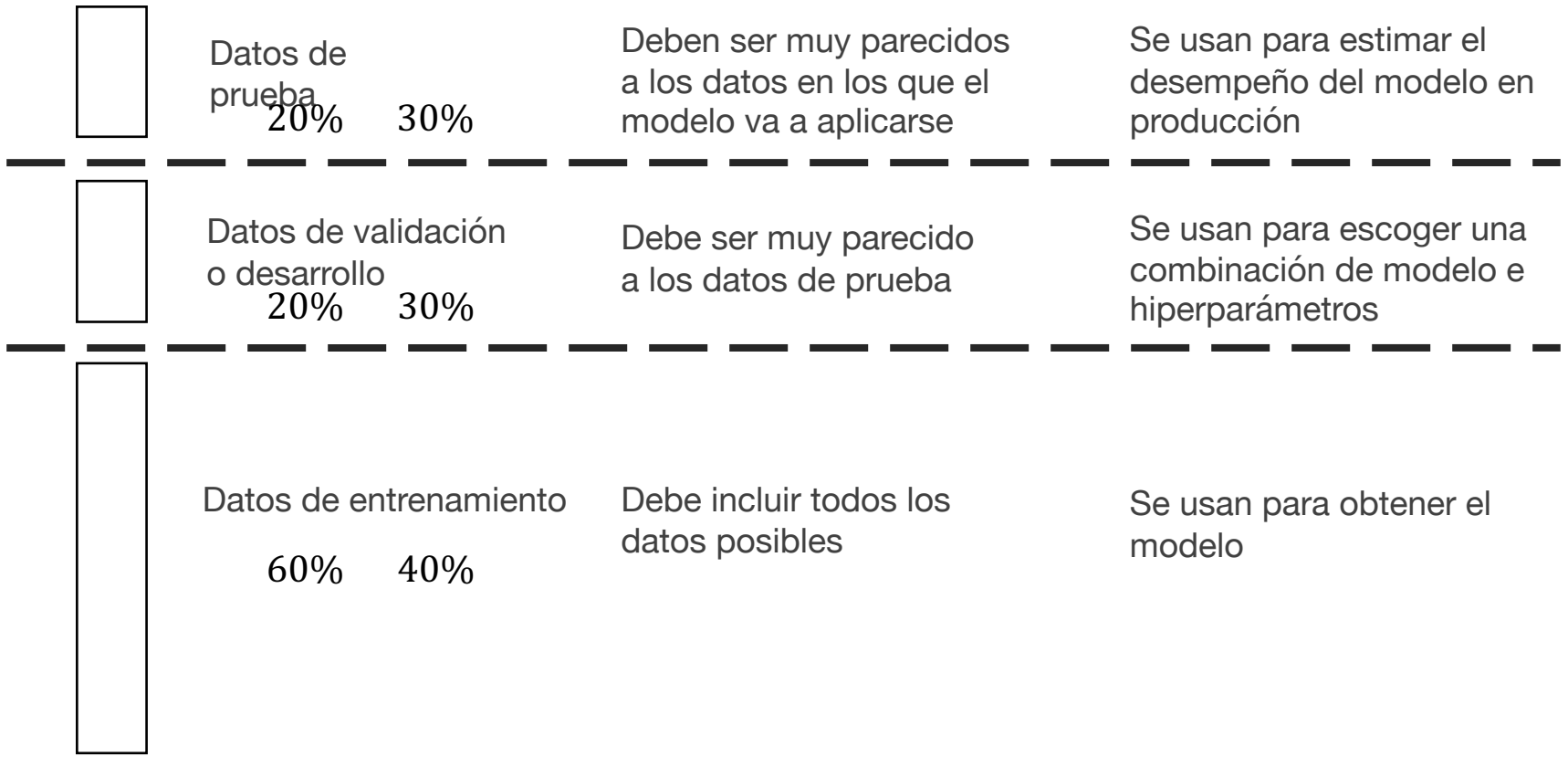
Evaluar diferentes modelos y ver cuál ofrece mejores resultados.



Datos de prueba 20% 30%	Deben ser muy parecidos a los datos en los que el modelo va a funcionar	Se usan para estimar el desempeño del modelo en producción
Datos de entrenamiento 80% 70%	Debe incluir todos los datos posibles	Se usan para obtener el modelo

MODEL TESTING

Evaluar diferentes combinaciones de modelos he hiperparámetros y ver cuál ofrece mejores resultados.



MODEL TESTING

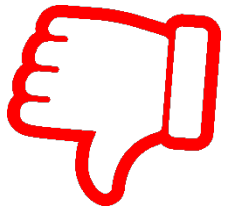
Ejercicio mental: suponga que queremos identificar automáticamente los daños en todos los vehículos de todas las marcas en Colombia.

Bogotá
Medellín
Cali
Bucaramanga

Entrenamiento de los modelos

Leticia
Pasto
Manizales
Pereira

Prueba de los modelos



MODEL TESTING

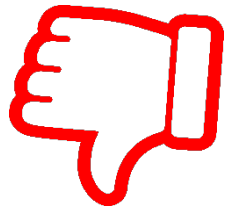
Ejercicio mental: suponga que queremos identificar automáticamente los daños en todos los vehículos de todas las marcas en Colombia.

Bogotá
Medellín
Cali
Bucaramanga

Entrenamiento y validación de modelos

Leticia
Pasto
Manizales
Pereira

Prueba de los modelos



MODEL TESTING

Ejercicio mental: suponga que queremos identificar automáticamente los daños en todos los vehículos de todas las marcas en Colombia.

Bogotá
Medellín
Cali
Bucaramanga
Leticia
Pasto
Manizales
Pereira

Entrenamiento de modelos

Validación y prueba de los modelos

MODEL TESTING

Ejercicio mental: suponga que queremos identificar automáticamente los daños en todos los vehículos de todas las marcas en Colombia.

Bogotá
Medellín
Cali
Bucaramanga
Leticia
Pasto
Manizales
Pereira

Entrenamiento, validación y prueba modelos

MODEL TESTING

Ejercicio mental: suponga que queremos identificar automáticamente los daños en todos los vehículos de todas las marcas en Colombia.

Bogotá	Entrenamiento
Medellín	
Cali	
Bucaramanga	Validación y prueba modelos
Leticia	
Pasto	
Manizales	
Pereira	

HYPERPARAMETERS SEARCH

Cómo debemos escoger (buscar) los hiperparámetros?

Curvas de validación

Define una métrica de interés

Define un hiperparámetro de interés

Define un rango de valores para el hiperparámetro → No uses valores muy cercanos, de esta forma tu búsqueda será más efectiva

Entrena un modelo diferente para cada valor del hiperparámetro

Grafica el resultado de aplicar la métrica a los modelos anteriores en los conjuntos de datos de entrenamiento y validación → Obviamente el siguiente paso es analizar esta gráfica

TRAINING SIZE SEARCH

Cómo saber si me hacen falta datos de entrenamiento?

Curvas de aprendizaje

Define una métrica de interés

Define un rango de valores para el tamaño de los datos de entrenamiento

Entrena un modelo diferente para cada tamaño del conjunto de datos de entrenamiento

Grafica el resultado de aplicar la métrica a los modelos anteriores en los conjuntos de datos de entrenamiento y validación

No uses valores muy cercanos, de esta forma tu búsqueda será ~~más efectiva~~

→ Obviamente el siguiente paso es analizar esta gráfica

NUMBERS OF ITERATIONS SEARCH

Curvas de aprendizaje

Define una métrica de interés

Define una frecuencia con la cuál quieras conocer el rendimiento de tu modelo. → No uses una frecuencia muy alta, de esta forma tu no le quitarás mucha velocidad a tu algoritmo de entrenamiento

Grafica el resultado de aplicar la métrica al estado del modelo en las iteraciones definidas usando los datos de entrenamiento y validación → Obviamente el siguiente paso es analizar esta gráfica

DEMO

Demo - Curvas de validación-1

CROSS VALIDATION

Validación cruzada (K-fold cross validation)

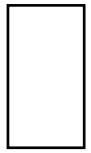
Algunas veces…

No se tienen suficientes datos para separar los datos en entrenamiento validación y prueba.

Queremos tener una medida más robusta del error en los datos de validación, ya que la anterior estrategia puede estar sesgada. .

CROSS VALIDATION

Validación cruzada (K-fold cross validation)

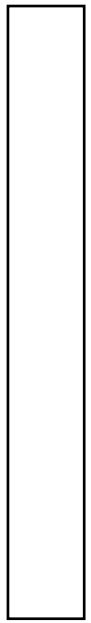


Datos de prueba

20% 30%

Deben ser muy parecidos a los datos en los que el modelo va a funcionar

Se usan para estimar el desempeño del modelo en producción



Debe incluir todos los datos posibles

Se usan para obtener el modelo

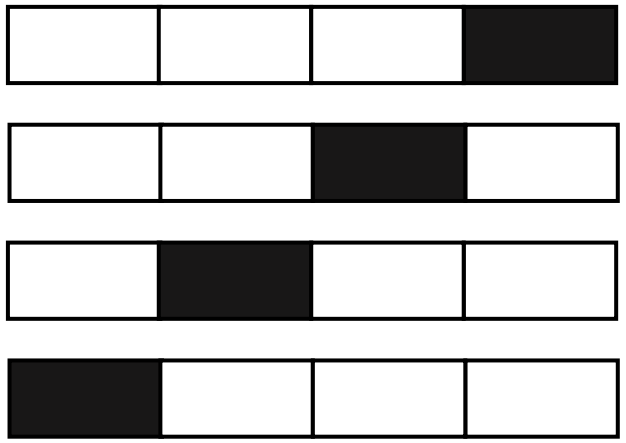
Datos de entrenamiento

80% 70%

Es posible hacer una estimación no sesgada del desempeño de este modelo, con ciertos hiperparámetros, usando solo este conjunto de datos.

CROSS VALIDATION

Cómo funciona?



- Divida los datos de entrenamiento en K partes iguales.
- Cada parte tomará el papel de conjunto de validación en una ocasión.
- El resto de partes se usa para entrenar el modelo cada vez.
- El error de validación se estima promediando los el de todos las K partes.

DEMO

Demo - Curvas de validación-2

HYPERPARAMETERS SEARCH

Búsqueda automática de hiperparámetros

Buscar los hiperparámetros es un problema de optimización: **encontrar la combinación de hiperparámetros que minimiza el error de validación**

HYPERPARAMETERS SEARCH

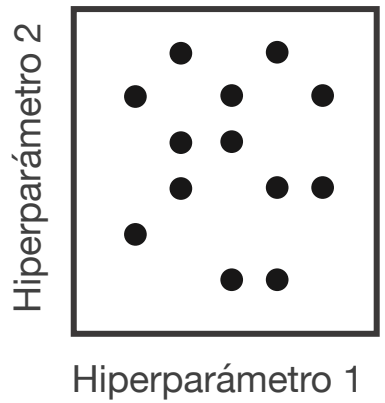
One by one

Curvas de validación

Grid search

Evaluar **todas** las posibles combinaciones de los distintos hiperparámetros

Randomized search



Evaluar **algunas** de las posibles combinaciones de los distintos hiperparámetros

DEMO

Búsqueda hiperparámetros

UNIVERSIDAD EAFIT

¡Gracias!

Aliados:

Microsoft

idata
Advanced analytics for business

Vigilada Mineducación