Universidad del Rosario

Analisis Avanzado de Datos

W3. Evaluación de Modelos y Métodos de Regularización

FERNEY ALBERTO BELTRAN MOLINA
Escuela de Ingeniería, Ciencia y Tecnología
Matemáticas Aplicadas y Ciencias de la Computación

Profesor

FERNEY ALBERTO BELTRAN MOLINA
ferney.beltran@urosario.edu.co

Ingeniero Electrónico.
Magister en TIC
Candidato Doctor en TIC

Director del Centro de investigación e innovación CEINTECCI.
Miembro de la junta directiva Avanciencia

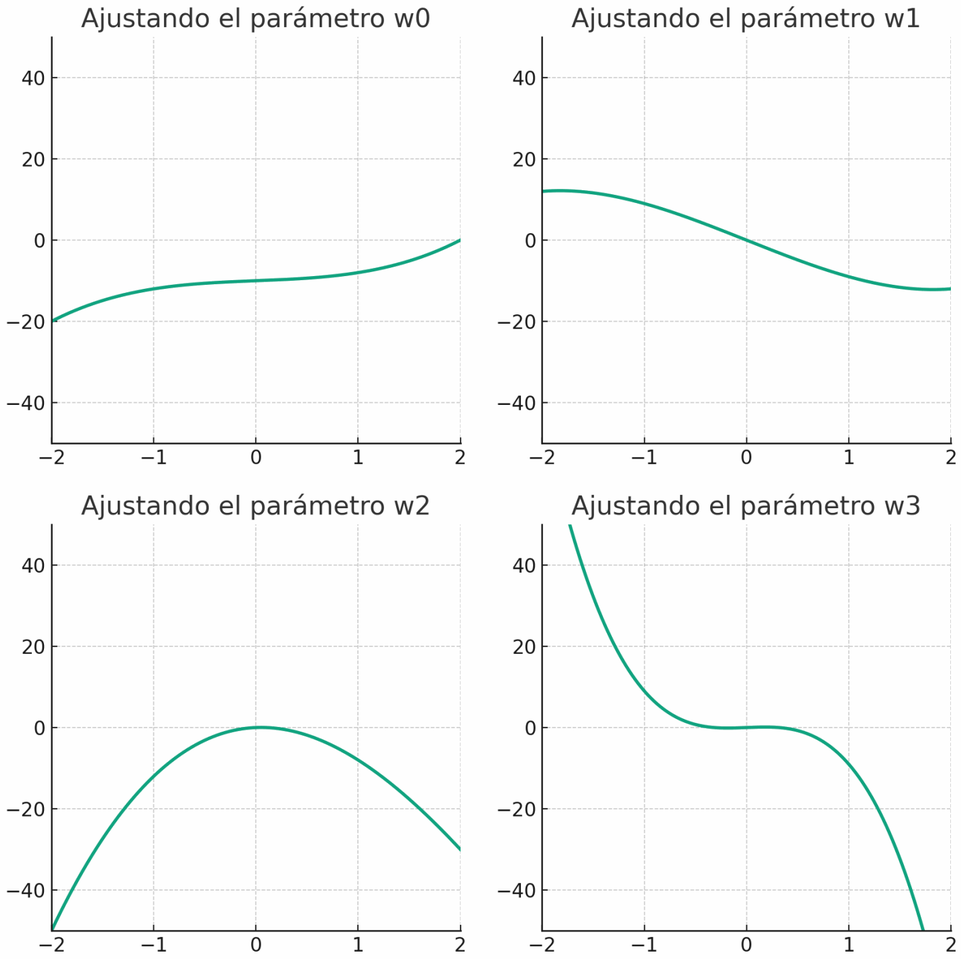
Procesamiento y análisis de datos basadas en IA.
Simulación y modelado por computación,
Optimizan Sistemas de procesamiento en hardware y software
Diseño de sistemas electrónicos reconfigurables

Flexibilidad

Los modelos nos permiten generar múltiples formas ajustando sus parámetros. Hablamos de los **grados de libertad** o la **complejidad** de un modelo para describir su capacidad para generar diferentes formas, es decir, su **flexibilidad**. Los grados de libertad de un modelo generalmente están relacionados con el número de parámetros del modelo:

- Un modelo lineal tiene dos parámetros y es inflexible, ya que sólo puede generar líneas rectas.
- Un modelo cúbico tiene 4 parámetros y es más flexible que uno lineal.

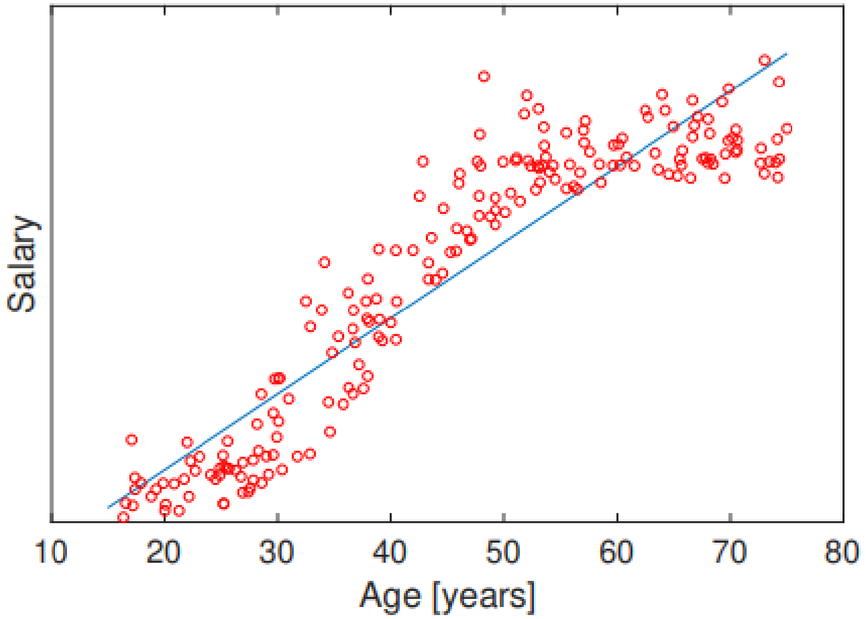
La flexibilidad de un modelo está relacionada con su **interpretabilidad** y **precisión**, y hay un equilibrio entre ambos.



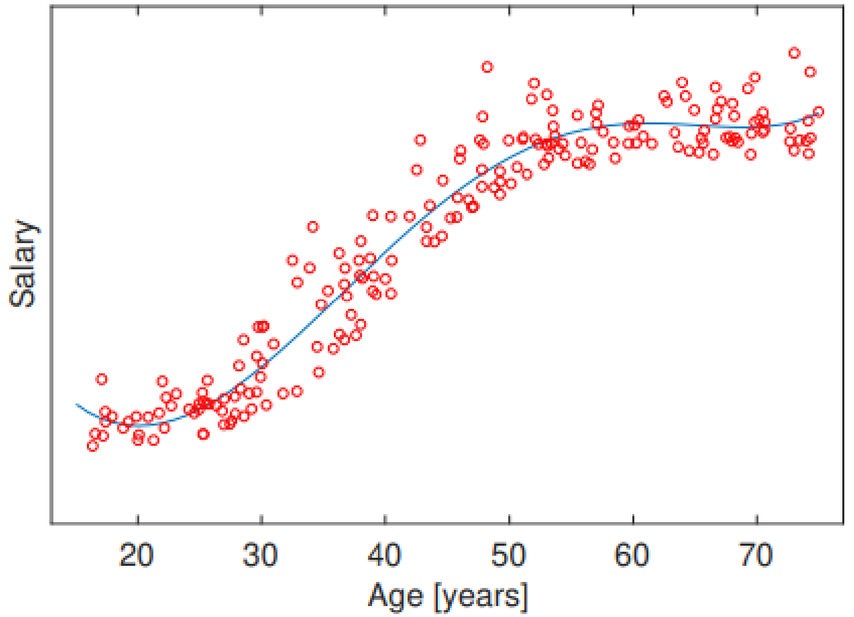
Modelos mas flexibles modelos mas complejos

Interpretabilidad

La **interpretabilidad** del modelo es crucial para nosotros, como humanos, para entender de una manera **cualitativa** el cómo un predictor se asigna a una etiqueta, Los modelos inflexibles producen soluciones que suelen ser más simples y fáciles de interpretar.



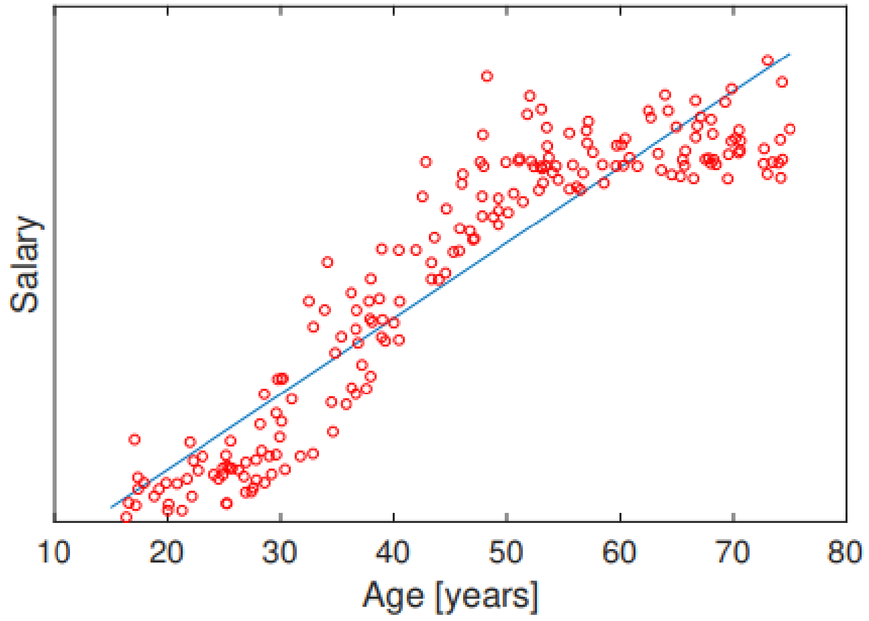
Según este modelo lineal, cuanto mayor te haces, más dinero ganas



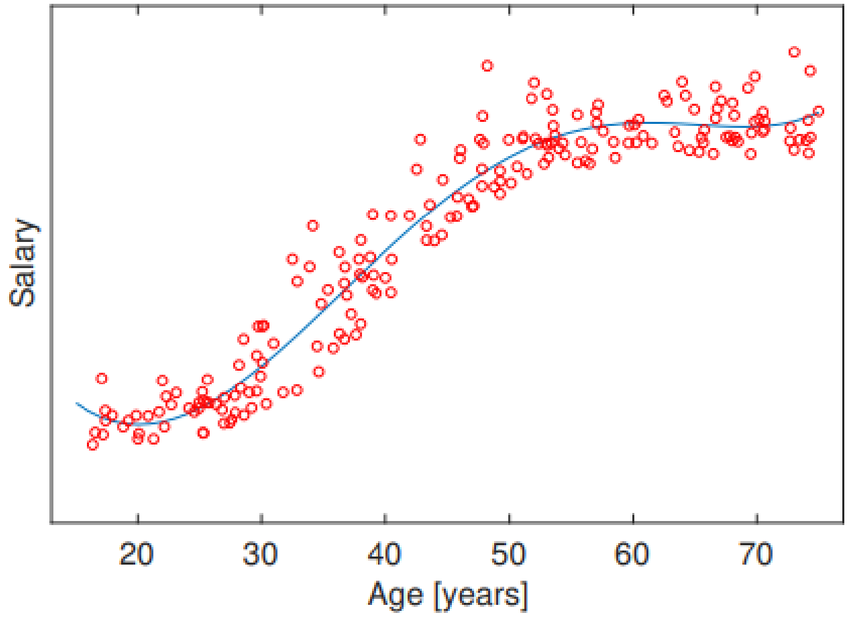
En este modelo polinomial, nuestro salario sigue siendo el mismo que el de los adolescentes, luego aumenta entre los 20 y los 50 años, luego …

Precisión

La **precisión** de un modelo también está relacionada con su flexibilidad. Durante el entrenamiento, los modelos flexibles generalmente tienen errores más bajos que los modelos inflexibles



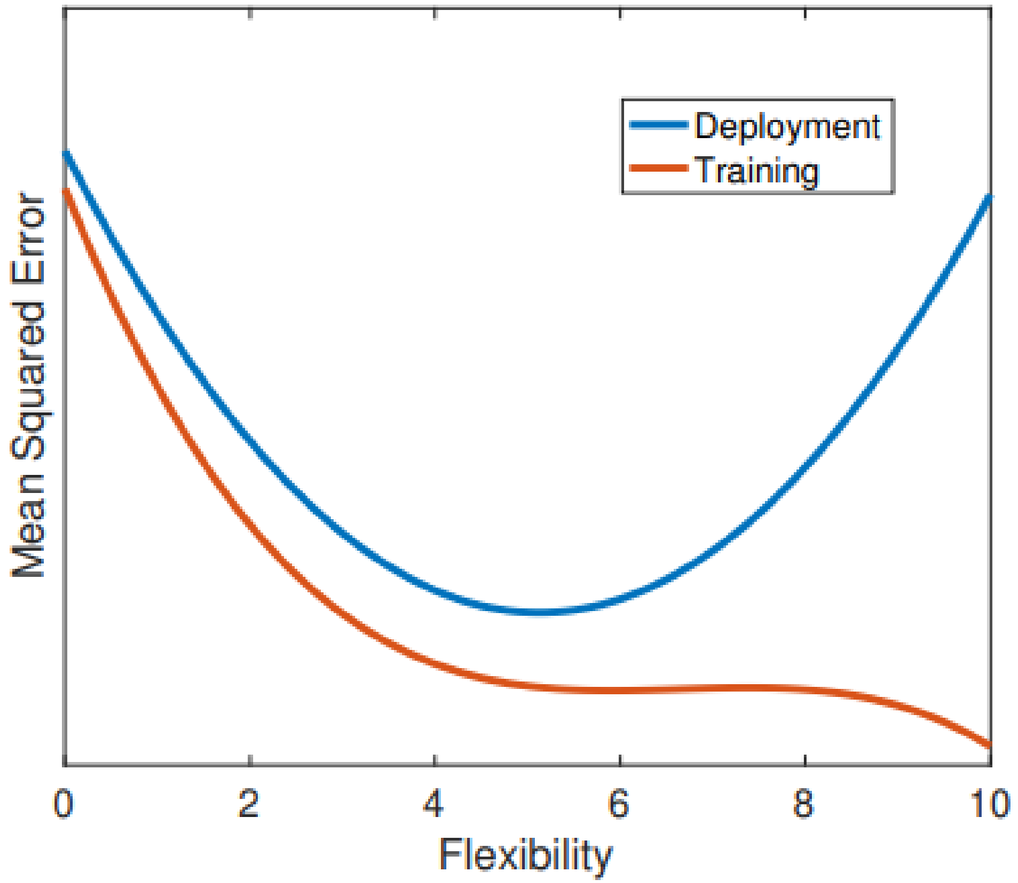
El error de entrenamiento del mejor modelo lineal es EMSE = 0.0983



El error de entrenamiento del mejor modelo polinómico es EMSE = 0.0379

cuando mas complejo el modelo, menor el error en los datos de entrenamiento

Generalización (principio fundamental)



- En esta figura, la curva roja representa el MSE de entrenamiento de diferentes modelos de creciente complejidad, mientras que la curva azul representa el MSE de despliegue para los mismos modelos.

¿Qué está pasando?

La generalización es la capacidad de nuestro modelo para traducir exitosamente lo que aprendió durante la etapa de aprendizaje al despliegue

Otras medidas de Calidad

Además del MSE, promedio de los errores cuadrados entre las predicciones y los valores reales., podemos considerar otras métricas de calidad:

- **Error cuadrático medio raíz**. Mide la desviación estándar de muestra del error de predicción.

$$E_{RMSE} = \sqrt{\frac{1}{N}\sum e_i^2}$$

- **Error absoluto medio**. Mide el promedio del error absoluto de predicción.

$$E_{MAE} = \frac{1}{N}\sum |e_i|$$

- **R-cuadrado**. Mide la proporción de la varianza en la respuest que es predecible a partir de los predictores.

$$E_R = 1 - \frac{\sum e_i^2}{\sum (y_i - \bar{y})^2},$$

where $\bar{y} = \frac{1}{N}\sum y_i$

Otras medidas de Calidad

Además del MSE, promedio de los errores cuadrados entre las predicciones y los valores reales., podemos considerar otras métricas de calidad:

- **Residual Estándar (RSE).** Desviación Estándar Residual, es una métrica que se utiliza para medir la cantidad de variación que no es explicada por el modelo de regresión. Es especialmente relevante en el contexto de la regresión lineal.

$$\mathrm{RSE} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-p-1}}$$

n es el número total de observaciones

p es el número de predictores en el modelo (sin contar el término constante)

Cuánto error hay en las predicciones del modelo en comparación con la variabilidad de la variable dependiente. Es una medida de la dispersión de los residuos. Un RSE pequeño indica que el modelo ajusta bien los datos, mientras que un RSE grande sugiere que el modelo no captura toda la variabilidad en los datos.

Y tenemos mas de una variable dependiente

$$MSE_{\text{promedio}} = \frac{1}{m}\sum_{j=1}^{m} MSE_j$$

Donde MSE_j para cada variable dependiente j se define como:

$$MSE_j = \frac{1}{n}\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2$$

$$RSE_{\text{promedio}} = \frac{1}{m}\sum_{j=1}^{m} RSE_j$$

Donde RSE_j para cada variable dependiente (j) se define como:

$$RSE_j = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_{ij} - \hat{y}_{ij})^2}$$

Te has preguntado de donde viene el termino de regresión ?

En el siglo XIX, Galton observó que los hijos de personas altas tienden a ser más altos que el promedio, pero no tanto como sus padres los hijos de padres extremadamente altos o bajos tendían a "regresar" hacia la media poblacional en términos de altura. Es decir, no eran tan extremos como sus padres en términos de altura.

Galton inicialmente llamó a este fenómeno "reversión hacia la mediocridad" (reversion to mediocrity), pero más tarde el término cambió a "regresión hacia la media" (regression to the mean).

Superficie de Error

Teoría de la optimización

Supongamos que tenemos:

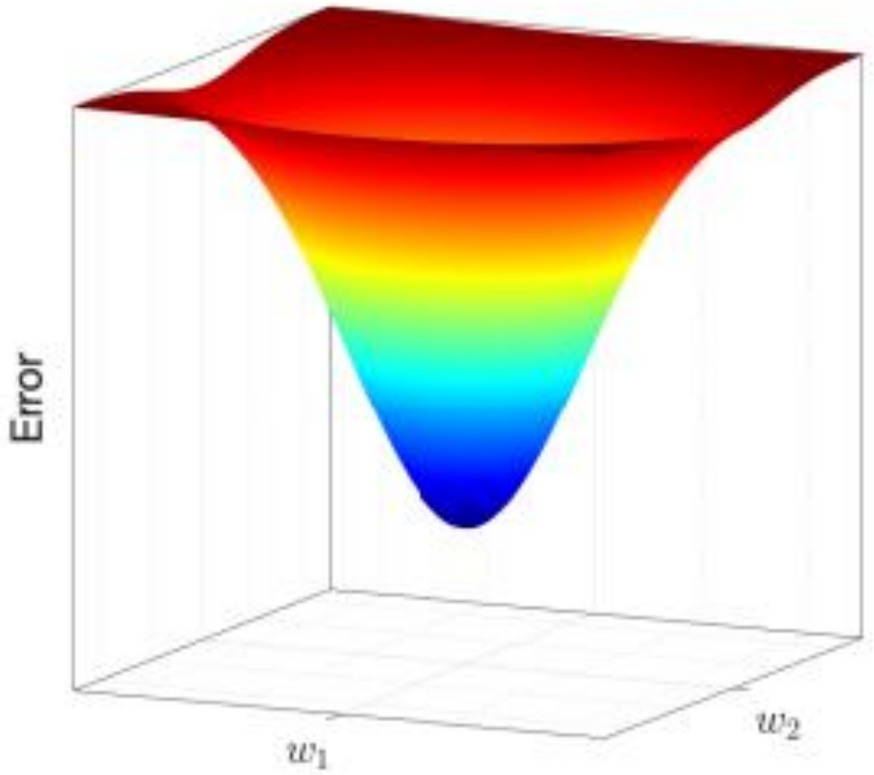
- Una familia de **modelos candidatos**, por ejemplo, modelos lineales.
- Una **métrica de calidad** (la misma que se usa para las pruebas), por ejemplo, MSE
- Una **descripción ideal** de la población objetivo.

La optimización nos permite identificar, entre todos los modelos candidatos, aquel que logra la mayor calidad en la población objetivo, es decir, el modelo **óptimo**.

El concepto de **superficie de error** (también conocido como función de error, objetivo, pérdida o coste) es el concepto principal en la teoría de optimización. Asumiremos que podemos obtenerlo utilizando la descripción ideal de nuestra población objetivo.

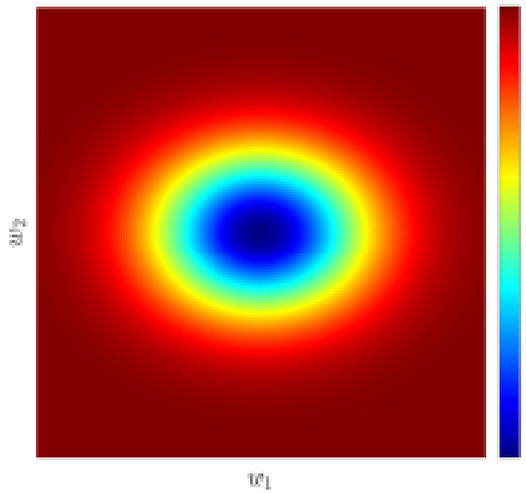
Superficie de Error

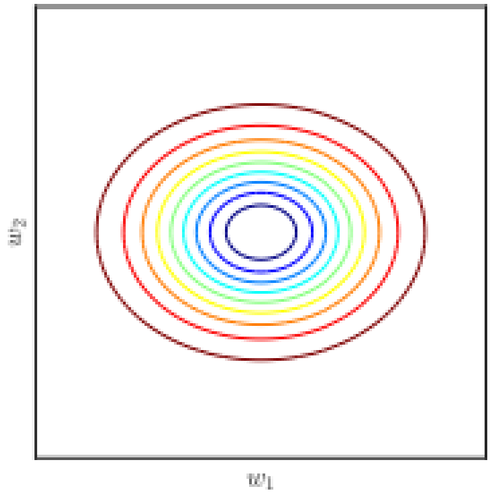
Una superficie de error asigna a cada modelo candidato su error. La denotaremos por $E(w)$, donde w representa un modelo. En esta figura, $w=[w1,w2]$



Superficie de Error

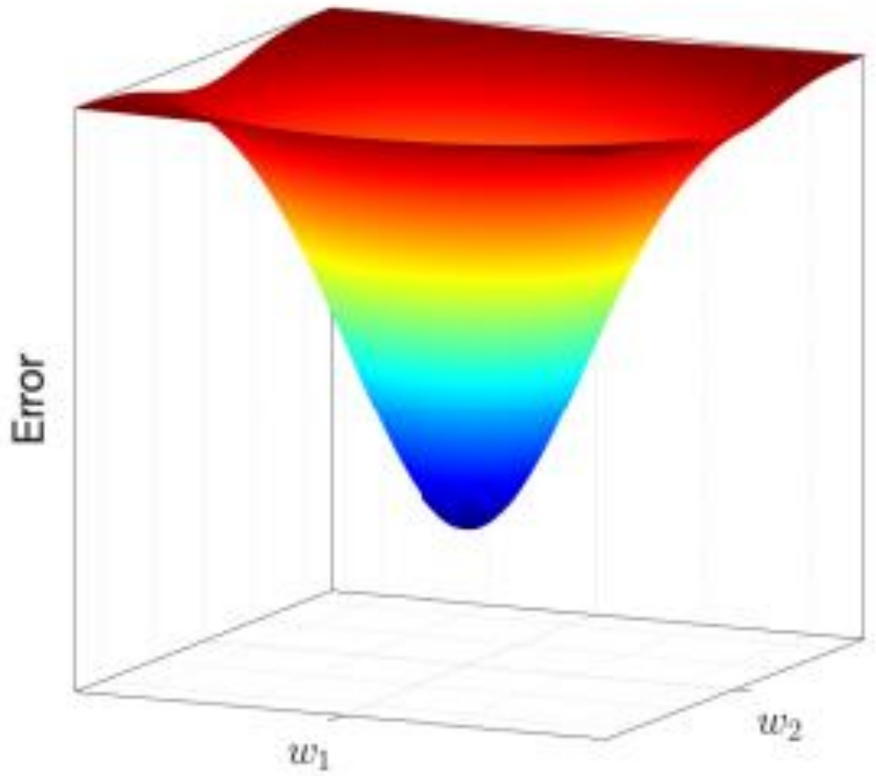
Las superficies de error también pueden representarse como mapas codificados por colores y mapas de contorno, donde la paleta de colores codifica los valores del error





Superficie de Error y modelo óptimo

El modelo **óptimo** se puede identificar como aquel con el error más bajo

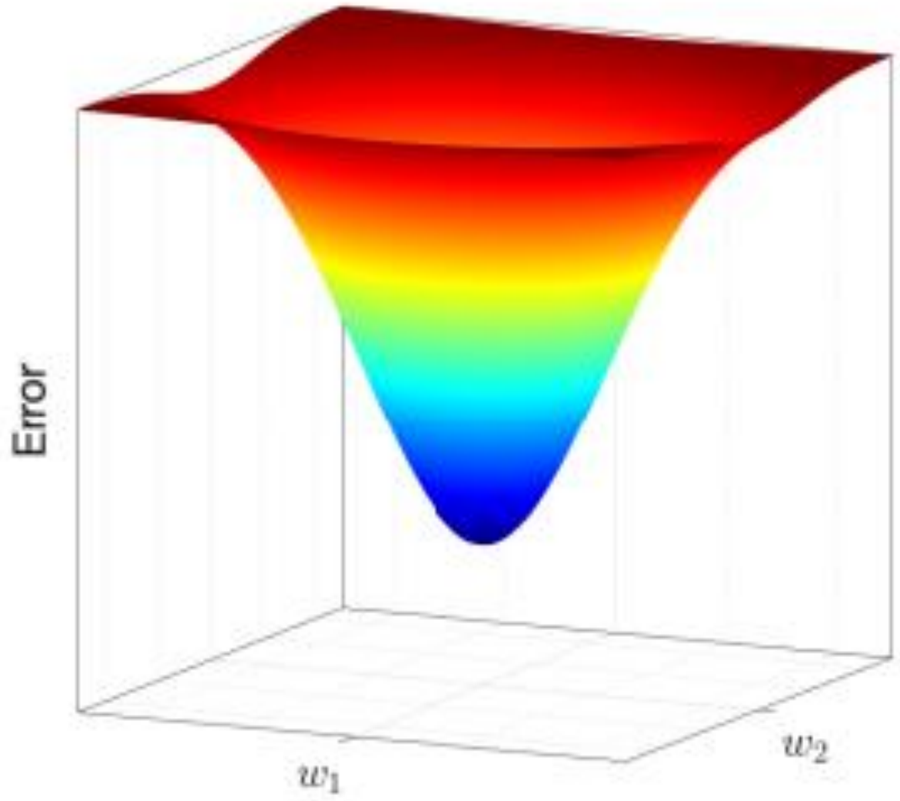


Superficie de Error y modelo óptimo

El **gradiente** (pendiente) de la superficie de error, $\nabla E(w)$, es cero en el modelo óptimo, por lo tanto, podemos centrarnos en modelos donde

$$\nabla E(\boldsymbol{w}) = 0.$$

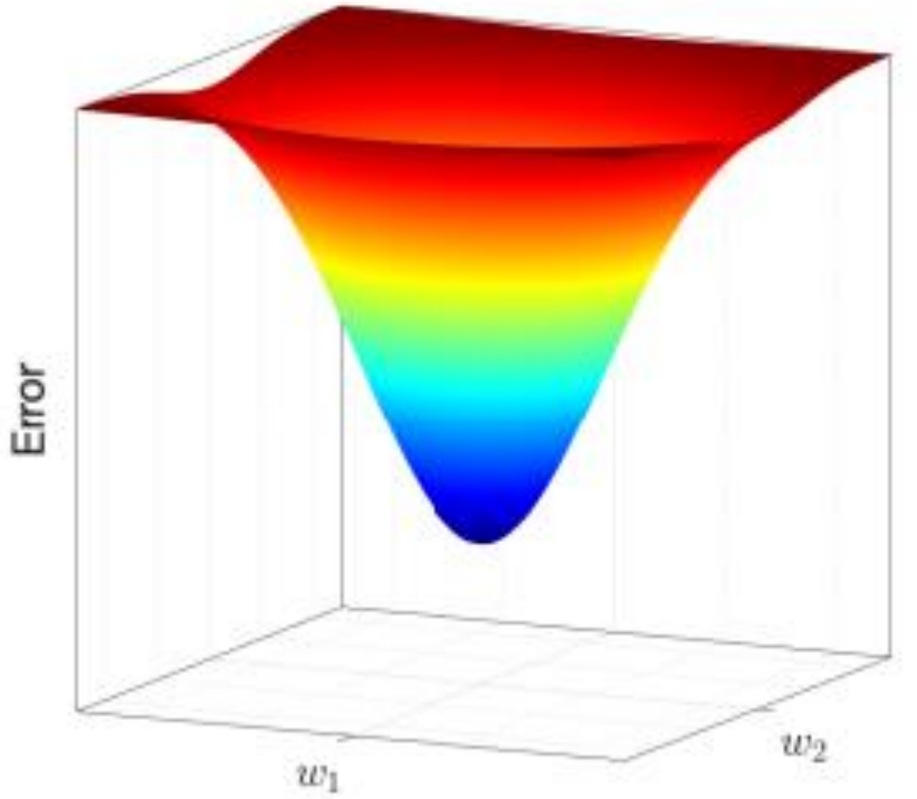
La pendiente de la superficie es cero



Superficie de Error y modelo óptimo

¿Qué sucede si no se dispone de suficiente capacidad de cálculo para obtener el error o el **gradiente de cada modelo candidato**?

¿Cómo podemos encontrar el modelo óptimo en ese caso?



Hazte esta pregunta: ¿puedes subir/bajar una montaña en completa oscuridad?

Descenso por Gradiente

El descenso de gradiente es un método de optimización numérica en el cual **actualizamos iterativamente** nuestro modelo utilizando el gradiente de la superficie de error.

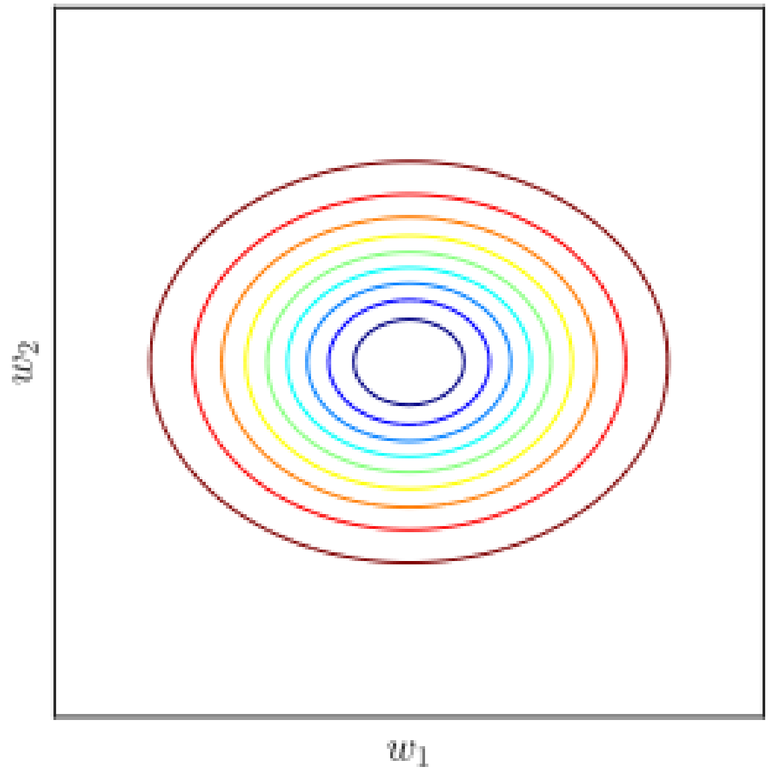
El gradiente proporciona la dirección en la cual el error aumenta más. Utilizando el gradiente, podemos crear la siguiente regla de actualización

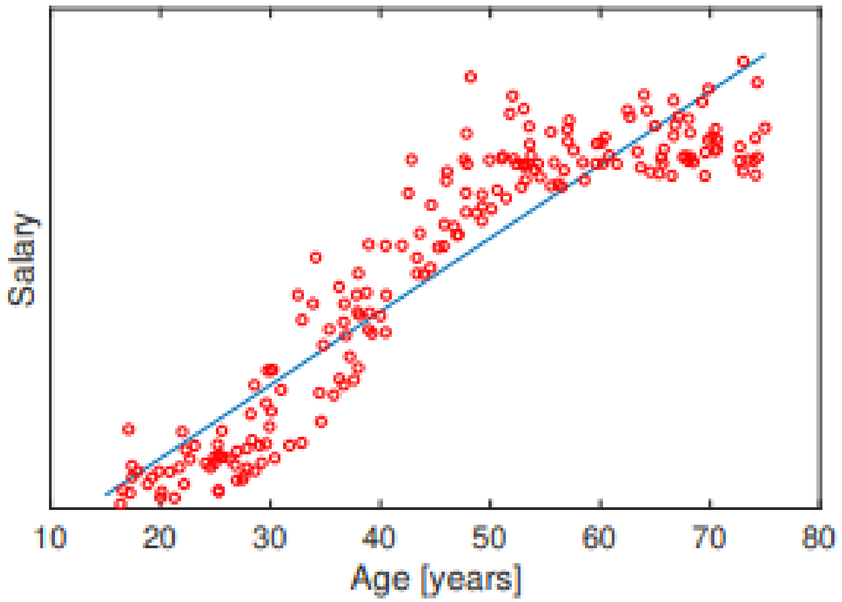
$$\boldsymbol{w}_{\text{new}} = \boldsymbol{w}_{\text{old}} - \epsilon \nabla E(\boldsymbol{w}_{old})$$

donde **ε** es conocida como la **tasa de aprendizaje (**learning rate**)** o el tamaño del paso.

Con cada iteración ajustamos los parámetros w de nuestro modelo. Por esta razón, este proceso también es conocido como **ajuste de parámetros**.

Descenso por Gradiente





Ten en cuenta que estamos representando gráficamente un conjunto de datos con fines ilustrativos. Sin embargo, en esta sección se supone que la superficie de error ha sido derivada de una descripción ideal de la población.

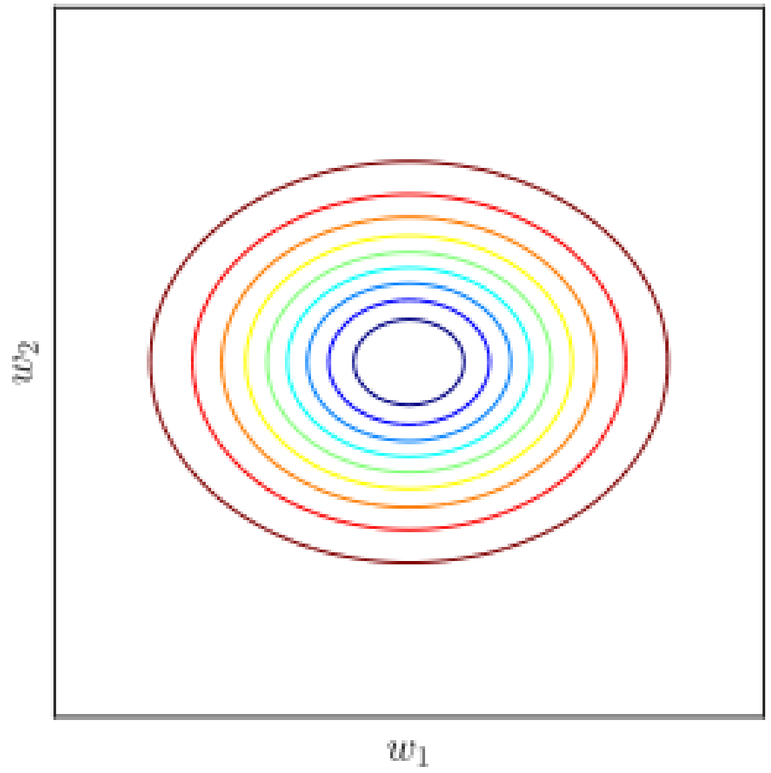
Tasa de aprendizaje (learning rate)

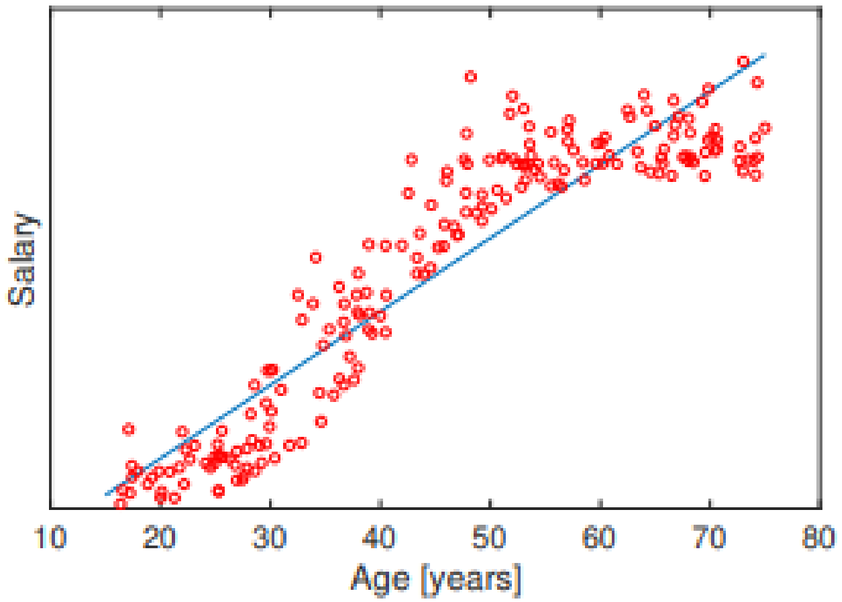
La tasa de aprendizaje **ε** controla cuánto cambiamos los parámetros w de nuestro modelo en cada iteración del descenso de gradiente:

- Valores pequeños de **ε** resultan en una convergencia lenta hacia el modelo óptimo.
- Valores grandes de **ε** corren el riesgo de sobrepasar el modelo óptimo.

Enfoques adaptativos pueden ser implementados, donde el valor de la tasa de aprendizaje disminuye progresivamente.

Tasa de aprendizaje (learning rate)





Ten en cuenta que estamos representando gráficamente un conjunto de datos con fines ilustrativos. Sin embargo, en esta sección se supone que la superficie de error ha sido derivada de una descripción ideal de la población.

Starting and stopping

Para que el descenso de gradiente comience, necesitamos un modelo inicial. La elección del modelo inicial puede ser crucial. Los parámetros iniciales w generalmente se eligen al **azar (random)** dentro de un rango de valores que depende del tipo de modelo.

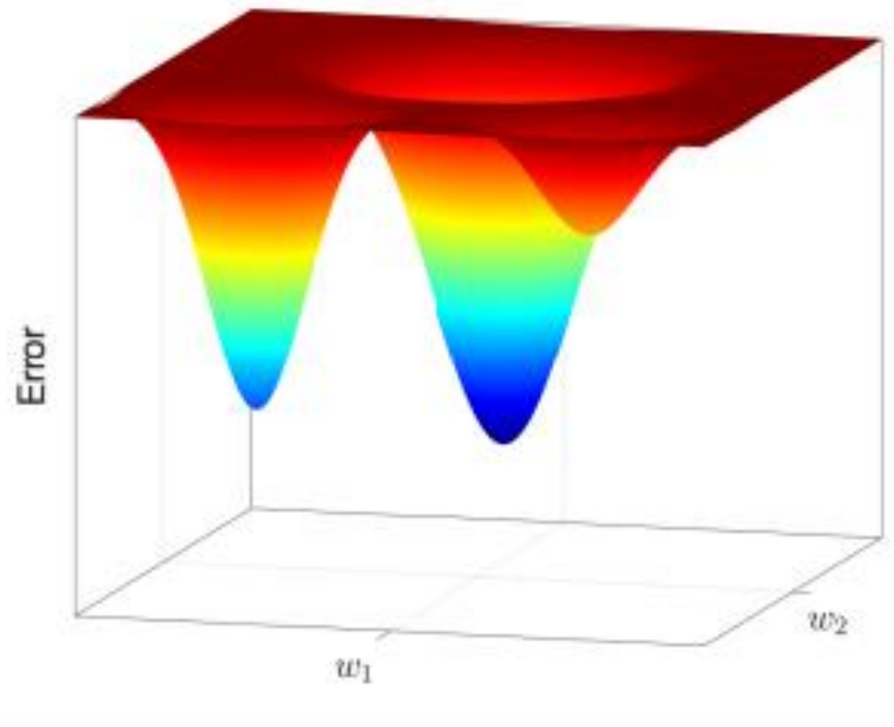
En general, el descenso de gradiente no alcanzará el modelo óptimo, por lo tanto, es necesario diseñar una estrategia de detención. Las opciones comunes incluyen:

- Número de iteraciones.
- Tiempo de procesamiento.
- Valor de error.
- Cambio relativo del valor de error.

Soluciones locales y globales

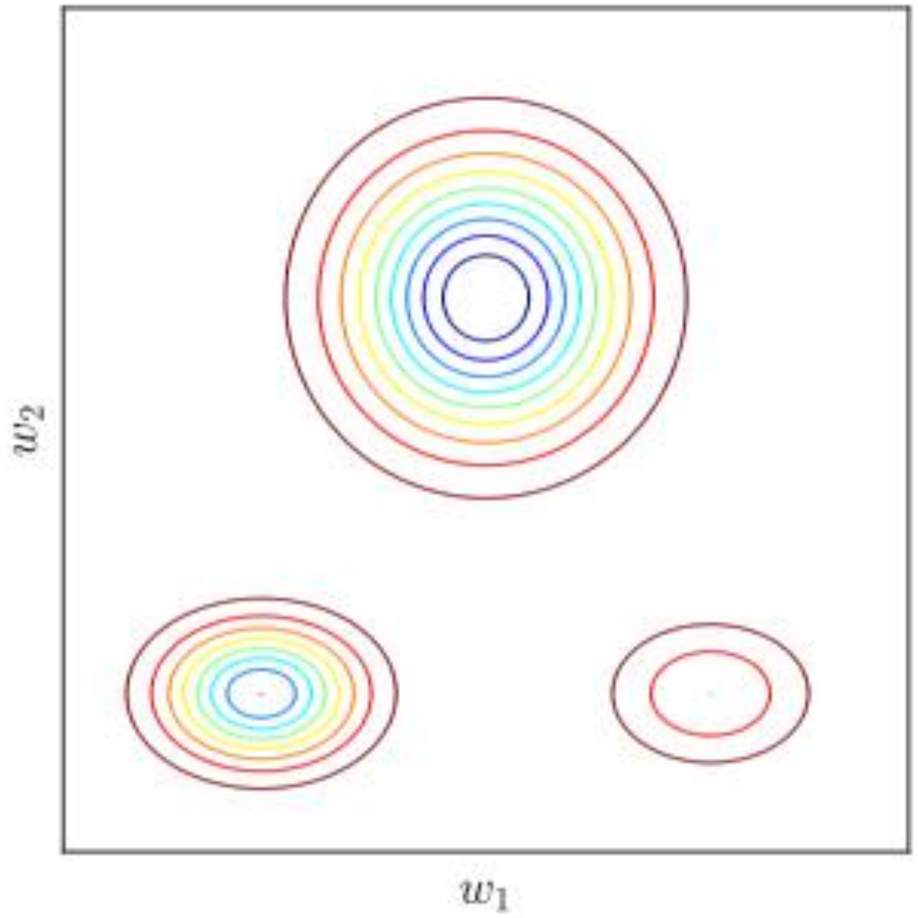
Hasta ahora, hemos considerado superficies de error convexas. Sin embargo, las superficies de error pueden ser complejas y tener:

- Óptimos **locales** (modelo con el error más bajo dentro de una región).
- Óptimos **globales** (modelo con el error más bajo entre todos los modelos).



Soluciones locales y globales

El descenso de gradiente puede quedar **atrapado** en óptimos locales. Para evitar esto, podemos repetir el procedimiento desde varios **modelos iniciales** y seleccionar el mejor.



¿Dónde está mi superficie de error?

En el Análisis avanzado de datos tenemos:

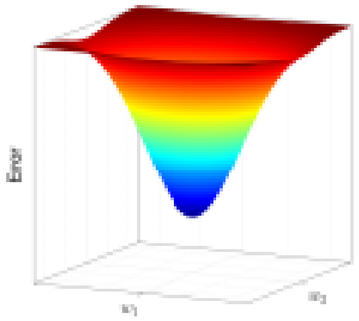
- Una familia de modelos candidatos (por ejemplo, modelos lineales).
- Una métrica de calidad (por ejemplo, el error).
- Datos extraídos de la población (es decir, no una descripción ideal).

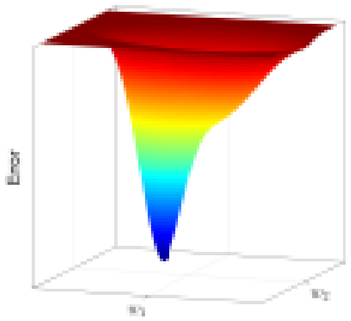
La noción de superficie de error nos permite identificar el modelo con el error más bajo entre los modelos candidatos. Sin embargo, ¿dónde está mi superficie de error?

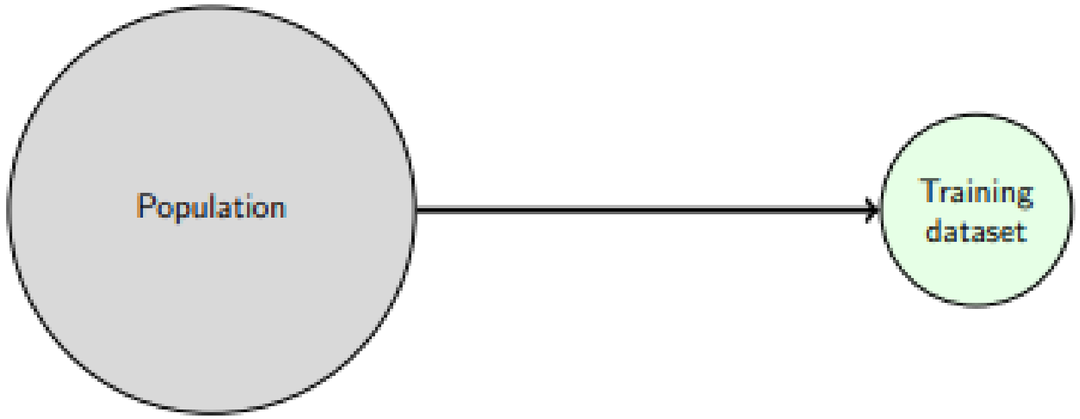
Si tuviéramos una descripción ideal de la población objetivo, podríamos calcular la superficie de error y su gradiente. En el aprendizaje automático, nuestro punto de partida es que no la tenemos y solo tenemos datos.

La superficie de error en el dataset

Las superficies de error empírica y verdadera son, en general, diferentes. Por lo tanto, sus modelos óptimos pueden ser diferentes, es decir, el mejor modelo para el conjunto de datos de entrenamiento podría no ser el mejor para la población.

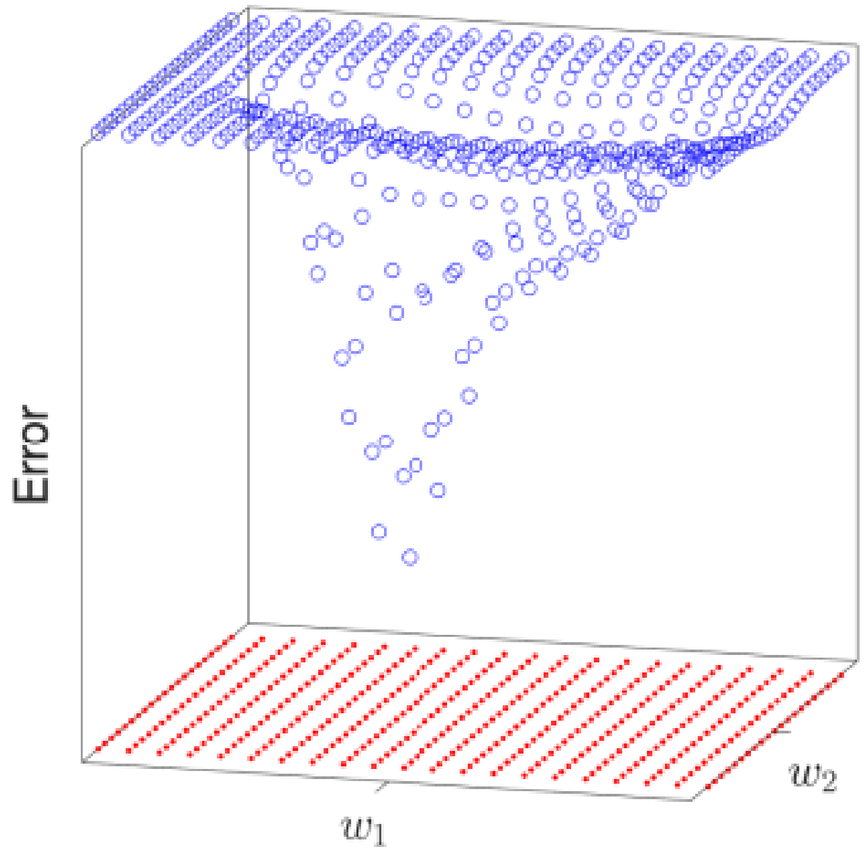






Problemática

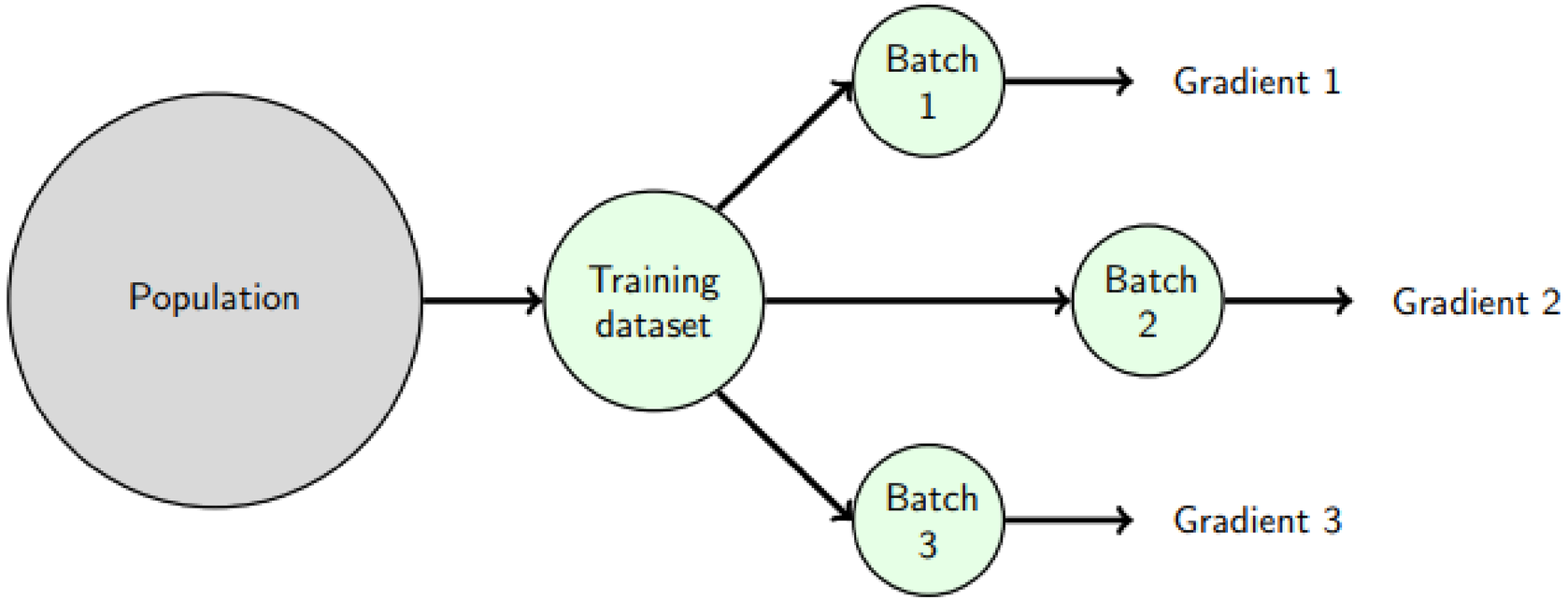
Evaluar muchísimos modelos, búsqueda exhaustivo



Descenso por gradiente de dataset

El descenso de gradiente se puede implementar estimando el gradiente utilizando nuestro conjunto de datos de entrenamiento.

En cada iteración, se utiliza un subconjunto (batch) del conjunto de datos de entrenamiento para calcular el gradiente de la superficie de error.



Descenso por gradiente de dataset

Dependiendo de la cantidad de datos utilizada en cada iteración, es común (aunque no realmente útil) distinguir entre:

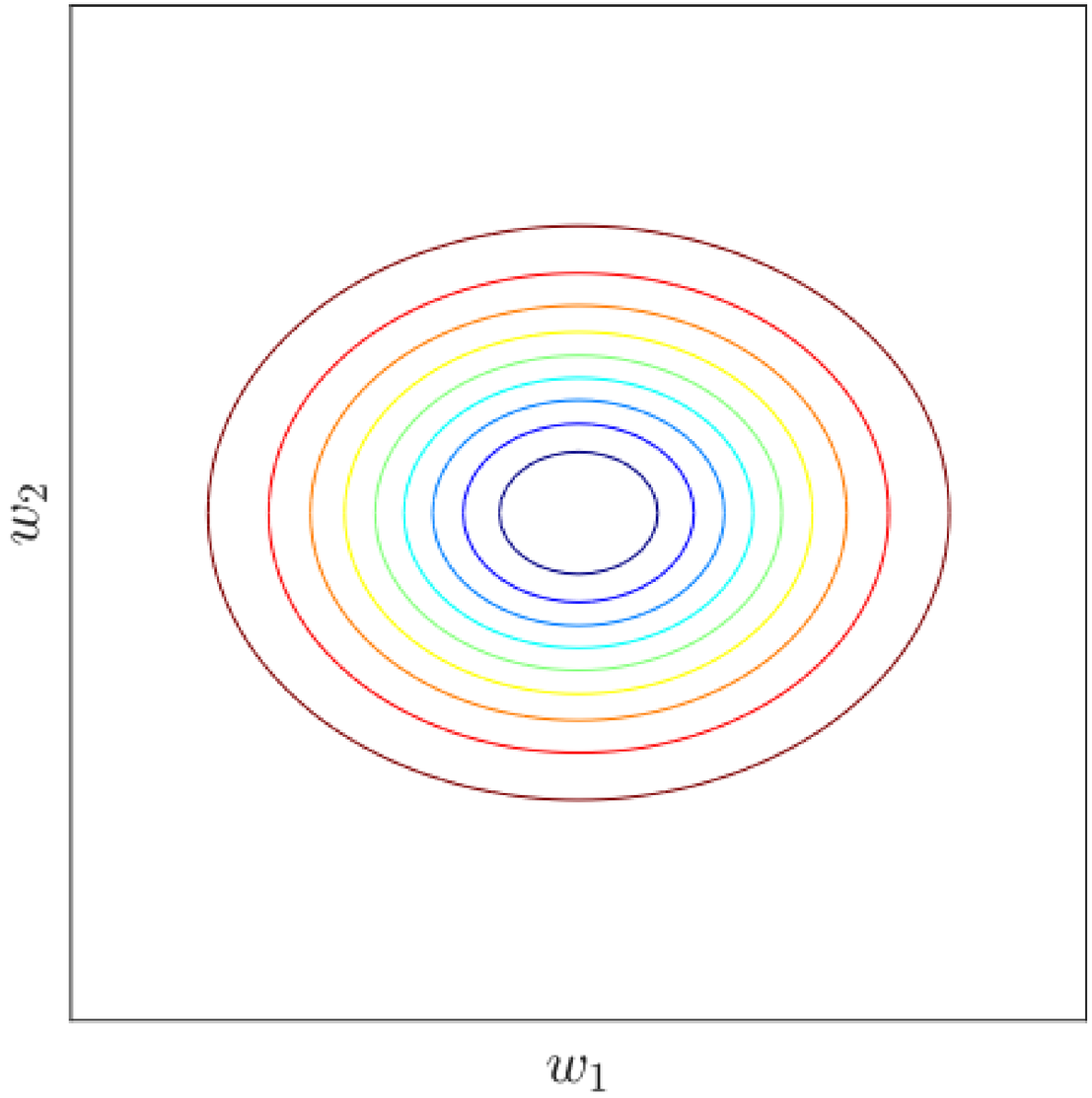
- **Batch**(se utiliza todo el conjunto de datos de entrenamiento).
- **Estocástico** (u online) (se utiliza una muestra).
- **Mini-batch** (se utiliza un pequeño subconjunto del conjunto de datos de entrenamiento).

Es más útil hablar sobre el **tamaño del batch** (un número entre 1 y el tamaño del conjunto de datos de entrenamiento). Independientemente del valor del tamaño del lote, usaremos el término descenso de **gradiente estocástico para este enfoque**.

Los lotes pequeños generan versiones ruidosas del gradiente de la superficie de error empírica, lo que puede ayudar a escapar de los mínimos locales.

Batch- size en descenso por gradiente de dataset

El descenso de gradiente estocástico es el algoritmo de optimización más utilizado, aunque a veces puede ser lento. Otros algoritmos populares de optimización basados en el gradiente

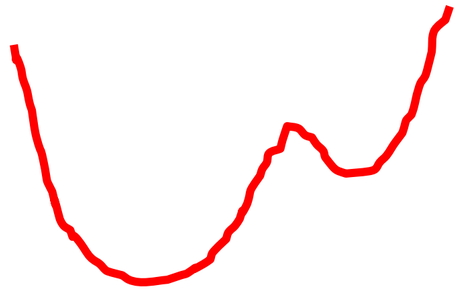


Momentum (direccion y magnitud), **RMSPop** adapta la tasa de aprendizaje , **Adam** combina ambas

Superficie de Error y sobreajuste

Las superficies de error empírica y verdadera son, en general, diferentes. Cuando se utilizan **conjuntos de datos pequeños** y **modelos complejos**, las diferencias entre ambas pueden ser muy grandes, lo que da como resultado modelos entrenados que funcionan muy bien para la superficie de error empírica pero muy mal para la superficie de error verdadera.

Esto, por supuesto, es otra manera de ver el **sobreajuste**. Al aumentar el tamaño del conjunto de datos de entrenamiento, las superficies de error empíricas se acercan más a las superficies de error verdaderas y el riesgo de sobreajuste disminuye.



Regularización

Las regularizaciones modifican la superficie de error empírica al agregar un término que limita los valores que los parámetros del modelo pueden tener. Una opción común es la superficie de error regularizada ER(w), definida como:

$$E_R(\boldsymbol{w}) = E(\boldsymbol{w}) + \lambda \boldsymbol{w}^T \boldsymbol{w}$$

Por ejemplo, el error cuadrático medio (MSE) en regresión puede regularizarse de la siguiente manera

$$E_{MSE+R} = \frac{1}{N} \sum_{i=1}^{N} e_i^2 + \lambda \sum_{i=1}^{K} w_k^2$$

Y la solución para MMSE

$$\boldsymbol{w} = \left(\mathbf{X}^T \mathbf{X} + N\lambda \mathbf{I}\right)^{-1} \mathbf{X}^T \boldsymbol{y}$$

A medida que λ aumenta, la complejidad de la solución resultante disminuye y también lo hace el riesgo de sobreajuste

Ridge Regression (Regularización Ridge):

$$ER(w) = MSE(w) + \lambda * \Sigma w^2$$

En Ridge Regression, se agrega un término de regularización que **penaliza los coeficientes del modelo** en función de su magnitud al cuadrado. El hiperparámetro λ controla cuánta penalización se aplica. A medida que λ aumenta, los coeficientes tienden a disminuir, lo que reduce la complejidad del modelo y previene el sobreajuste.

Ridge es especialmente útil cuando se tiene un conjunto de datos con correlación entre variables predictoras

Lasso Regression (Regularización Lasso):

$$ER(w) = MSE(w) + \lambda * \Sigma|w|$$

En Lasso Regression, se agrega un término de regularización al error cuadrático medio. Sin embargo, en Lasso, en lugar de penalizar los coeficientes al cuadrado, se penaliza sus valores absolutos.

Esto tiene la propiedad de forzar algunos coeficientes a volverse exactamente cero, lo que puede conducir a la selección automática de características.

Lasso es útil cuando se sospecha que solo un subconjunto de características es relevante para el modelo, ya que tiende a reducir la importancia de coeficientes irrelevantes.

Ejercicio de Regresión con Múltiples Predictores

Objetivo: Comprender y aplicar técnicas de regresión en un conjunto de datos con múltiples predictores y evaluar diferentes modelos y métricas.

- Elige un conjunto de datos que te interese y que esté disponible públicamente, que tenga varias características.
- Identifica una problemática que pueda ser abordada mediante regresión. Asegúrate de que el conjunto de datos que seleccionaste tenga relevancia para esta problemática.
- Implementa y entrena diferentes modelos de regresión.
- Evalúa cada modelo utilizando métricas adecuadas para regresión
- Analiza y compara el desempeño de los diferentes modelos.
- Utiliza bibliotecas como scikit-learn o statsmodels para implementar la regresión lineal y la regularización.
- Reflexiona sobre la importancia de la regularización y cómo puede ayudar a prevenir el sobreajuste, especialmente cuando se tienen muchas características.