Modelos de regresión

El análisis de regresión permite modelar, examinar y explorar relaciones espaciales y puede ayudar a explicar los factores detrás de los patrones espaciales observados.

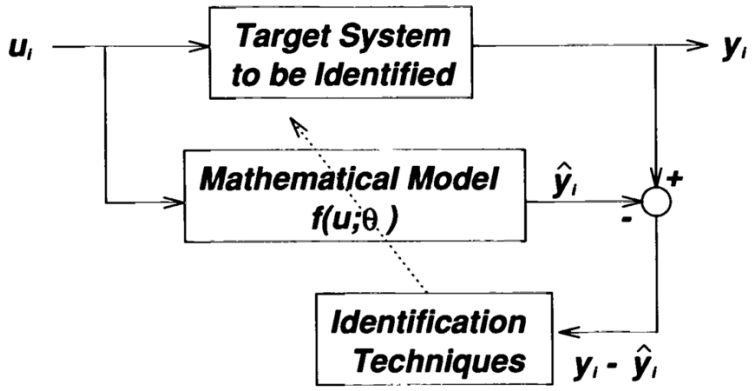

Figure 5.1. *Block diagram for parameter identification.*

El método tradicional considera un conjunto de datos de entrenamiento, $\{x_{p,i}; y_p\}_{p=1}^{q}$, con n variables exogena o explicativas (variables independientes: i=1,…,n) y una variable endogena dependiente del siguiente modelo:

$$\hat{y}_p = \theta_0 + \theta_1 x_{p,1} + \cdots + \theta_i x_{p,i} + \cdots + \theta_n x_{p,n}$$

$$y_p = \hat{y}_p + e_p$$

donde $\boldsymbol{\theta}$ es un vector de parámetros (desconocidos) y $\boldsymbol{e}$ es un error aleatorio normal de media cero y varianza σ^2, es decir $e_p \sim \mathcal{N}(0, \sigma^2)$ son *iid* (independientes e idénticamente distribuidos). Por tanto las hipótesis estructurales del modelo son:

- Linealidad
- Homocedasticidad (varianza constante del error)
- Normalidad (y homogeneidad: ausencia de valores atípicos y/o influyentes)
- Independencia de los errores

Hipótesis adicional en regresión:

- Ninguna de las variables explicativas es combinación lineal de las demás.

En el caso de regresión múltiple es de especial interés el fenómeno de la colinealidad (o multicolinealidad) relacionado con la última de estas. Además se da por hecho que el número de observaciones disponible (q) es como mínimo el número de parámetros ($\varrho = n + 1$), es decir $q > \varrho$.

El problema de la colinealidad

Si alguna de las variables explicativas no aporta información relevante sobre la respuesta puede aparecer el problema de la colinealidad.

En regresión múltiple se supone que ninguna de las variables explicativas es combinación lineal de las demás. Si una de las variables explicativas (variables independientes) es combinación lineal de las otras, no se pueden determinar los parámetros de forma única (sistema singular). Sin llegar a esta situación extrema, cuando algunas variables explicativas estén altamente correlacionadas entre sí, tendremos una situación de alta colinealidad. En este caso las estimaciones de los parámetros pueden verse seriamente afectadas:

- Tendrán varianzas muy altas (serán poco eficientes).
- Habrá mucha dependencia entre ellas (al modificar ligeramente el modelo, añadiendo o eliminando una variable o una observación, se producirán grandes cambios en las estimaciones de los efectos).

Selección de variables explicativas

Cuando se dispone de un conjunto grande de posibles variables explicativas suele ser especialmente importante determinar cuales de estas deberían ser incluidas en el modelo de regresión. Si alguna de las variables no contiene información relevante sobre la respuesta no se debería incluir (se simplificaría la interpretación del modelo, aumentaría la precisión de la estimación y se evitarían problemas como la colinealidad). Se trataría entonces de conseguir un buen ajuste con el menor número de variables explicativas posible.

Análisis e interpretación del modelo

Al margen de la colinealidad, si no se verifican las otras hipótesis estructurales del modelo, las conclusiones obtenidas pueden no ser fiables, o incluso totalmente erróneas:

- La falta de linealidad “invalida” las conclusiones obtenidas (cuidado con las extrapolaciones).
- La falta de normalidad tiene poca influencia si el número de datos es suficientemente grande. En caso contrario la estimación de la varianza, los intervalos de confianza y los contrastes podrían verse afectados.
- Si no hay igualdad de varianzas los estimadores de los parámetros no son eficientes pero sí insesgados. Las varianzas, los intervalos de confianza y contrastes podrían verse afectados.
- La dependencia entre observaciones puede tener un efecto mucho más grave.

Métricas de evaluación del modelo

Para evaluar el desempeño de un modelo de estimación, adoptamos el coeficiente de determinación R^2, coeficiente de correlación de Pearson (r) y el error cuadrático medio (MSE) que son las métricas de evaluación más utilizadas, y propusimos un índice para elegir y evaluar modelos computacionales. Dado un modelo, las observaciones experimentales y los valores predichos del modelo se define de la siguiente manera:

Coeficiente de determinación, R^2.

El coeficiente de determinación, se define como la proporción de variabilidad de la variable dependiente que es explicada por la regresión. El coeficiente de determinación se puede entender como una versión estandarizada del MSE, que proporciona una mejor interpretación del rendimiento del modelo. Técnicamente, el R^2 representa la varianza de las respuesta capturada por el modelo:

$$R^2 = 1 - \frac{\sum_{p=1}^{q}\left(y_p - \hat{y}_p\right)^2}{\sum_{p=1}^{q}\left(y_p - \bar{y}\right)^2}$$

Coeficiente de determinación ajustado, R^2_{ajus}.

Para evaluar la precisión de las predicciones podríamos utilizar el coeficiente de determinación ajustado, R^2_{ajus}, que estimaría la proporción de variabilidad explicada en una nueva muestra. Sin embargo, hay que tener en cuenta que su validez dependería de la de las hipótesis estructurales (especialmente de la linealidad, homocedasticidad e independencia), ya que se obtiene a partir de estimaciones de las varianzas residual y total:

$$R^2_{ajus} = 1 - \left(\frac{q-1}{q-\varrho-1}\right)(1-R^2)$$

donde q es el número de datos de entrenamiento y ϱ es el número de parámetros del modelo.

Coeficiente de correlación de Pearson, r.

El coeficiente de correlación de Pearson es la covarianza de las dos variables dividida por el producto de sus desviaciones estándar. La forma de la definición implica un "momento del producto", es decir, la media (el primer momento alrededor del origen) del producto de las variables aleatorias ajustadas a la media; de ahí el modificador momento-producto en el nombre.

El coeficiente de correlación de Pearson, cuando se aplica a una población, se representa comúnmente con la letra griega ρ (rho) y puede denominarse coeficiente de correlación poblacional o coeficiente de correlación poblacional de Pearson. Dado un par de variables aleatorias (X,Y) (por ejemplo, altura y peso), la fórmula para ρ es

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

donde $Cov(X,Y)$ es la covarianza de (X,Y), $Var(X)$ es la varianza de la variable X y $Var(Y)$ es la varianza de la variable Y.

El valor del índice de correlación $\rho(X,Y)$ varía en el intervalo [-1,1].

El coeficiente de correlación de Pearson, cuando se aplica a una muestra, se representa comúnmente por r_{xy} y puede denominarse coeficiente de correlación de la muestra o coeficiente de correlación de Pearson de la muestra. Podemos obtener una fórmula para r_{xy} sustituyendo en la fórmula anterior estimaciones de las covarianzas y varianzas basadas en una muestra. Dados datos emparejados $\{(x_1, y_1), \dots, (x_p, y_p), \dots, (x_q, y_q)\}$ que constan de n pares, r_{xy} se define como:

$$r_{xy} = \frac{\sum_{p=1}^{q}(x_p - \bar{x})(y_p - \bar{y})}{\sqrt{\sum_{p=1}^{q}(x_p - \bar{x})^2 \ \sum_{p=1}^{q}(y_p - \bar{y})^2}}$$

donde n es el tamaño de la muestra, x_p, y_p son los puntos muestrales individuales indexados con p, $\bar{x}$ es la media muestral y de manera análoga para $\bar{y}$.

Si este coeficiente es igual a 1 o -1 (o cercano a estos valores) significa que una variable es fruto de una transformación lineal de la otra. Teniendo una relación directa al tratarse de 1 (cuando una variable aumenta, la otra también), mientras que existirá una relación inversa al tratarse de -1 (cuando una variable aumenta la otra disminuye).

Mientras que, Si r = 0 (o cercano a este valor) no existe relación lineal, aunque puede existir algún otro tipo de relación no lineal.

El MSE del modelo se define de la siguiente manera:

$$MSE(Y, \hat{Y}) = \frac{1}{q}\sum_{p=1}^{q}(y_p - \hat{y}_p)^2$$

En general, cuanto más preciso es el modelo, mayor es la r y menor es el MSE. Al combinar las métricas r y MSE, el índice se define de la siguiente manera:

$$Index(Y, \hat{Y}) = \frac{r(Y, \hat{Y})}{MSE(Y, \hat{Y})}$$

Por lo tanto, cuanto mejor sea el modelo, mayor será el Índice.

Criterio de información de Akaike (AIC) y criterio de información bayesiano (BIC)

Los criterios de información son útiles para comparar modelos alternativos para la misma variable endógena. El AIC tiene como objetivo seleccionar el modelo que mejor hace predicciones dentro de un conjunto de datos. El criterio AIC se define como:

$$AIC = -2\log(\mathcal{L}) + 2\varrho$$

donde $\mathcal{L}$ es la máxima verosimilitud del modelo y ϱ es el número de parámetros. El AIC es una medida de la capacidad predictiva del modelo y tiende a sobreparametrizarlo.

El BIC penaliza más la complejidad que AIC, busca el modelo más abstracto, más sencillo y hace predicciones en un contexto más amplio. La definición de BIC reemplaza la constante 2 por log(q):

$$BIC = -2\log(\mathcal{L}) + \varrho\log(q)$$

donde q es el número de muestras.

Para un modelo gaussiano lineal, la probabilidad logarítmica máxima se define como:

$$\log(\mathcal{L}) = -\frac{q}{2}\log(2\pi) - \frac{q}{2}(\sigma^2) - \frac{SSE}{2\sigma^2}$$

$$SSE = \sum_{p=1}^{q}\left(y_p - \hat{y}_p\right)^2$$

donde σ^2 es una estimación de la varianza del ruido, y_p y $\hat{y}_p$ son, respectivamente, los objetivos reales y previstos, y q es el número de muestras. Reemplazando la máxima verosimilitud logarítmica en la fórmula AIC se obtiene:

$$\text{AIC} = q\log(2\pi\sigma^2) + \frac{SSE}{\sigma^2} + 2\varrho$$

donde σ^2 es una estimación de la varianza del ruido, estimada a través del estimador definido como:

$$\sigma^2 = \frac{\sum_{p=1}^{q}\left(y_p - \hat{y}_p\right)^2}{q - \varrho}$$

Tenga en cuenta que esta fórmula solo es válida cuando $q > \varrho$.

Métodos de regularización

El procedimiento habitual para ajustar un modelo de regresión lineal es emplear mínimos cuadrados, es decir, utilizar como criterio de error la suma de cuadrados residual, SSE.

Si el modelo lineal es razonablemente adecuado, utilizar SSE va a dar lugar a estimaciones con poco sesgo, y si además $q \gg \varrho$, entonces el modelo también va a tener poca varianza (bajo las hipótesis estructurales, la estimación es insesgada y además de varianza mínima entre todas las técnicas insesgadas). Las dificultades surgen cuando ϱ es grande o cuando hay correlaciones altas entre las variables predictoras: tener muchas variables dificulta la interpretación del modelo, y si además hay problemas de colinealidad o se incumple $q \gg \varrho$, entonces la estimación del modelo va a tener muchas varianza y el modelo estará sobre ajustado. La solución pasa por forzar a que el modelo tenga menos complejidad para así reducir su varianza. Una forma de conseguirlo es mediante la regularización (regularization o shrinkage) de la estimación de los parámetros θ_1, θ_2, …, θ_n que consiste en considerar todas las variables predictoras pero forzando a que algunos de los parámetros se estimen mediante valores muy próximos a cero, o directamente con ceros. Esta técnica va a provocar un pequeño aumento en el sesgo pero a cambio una notable reducción en la varianza y una interpretación más sencilla del modelo resultante.

Hay dos formas básicas de lograr esta simplificación de los parámetros (con la consiguiente simplificación del modelo), utilizando una penalización cuadrática (norma L_2) o en valor absoluto (norma L_1).

1. Modelo de regresión lineal simple.

El modelo de regresión simple relaciona un predictor y una respuesta.

Sean q observaciones $\{x_p; y_p\}_{p=1}^{q}$ pares de predictores (**x**) y respuestas (**y**), tales que $e_p \sim \mathcal{N}(0, \sigma^2)$ son *iid* (independientes e idénticamente distribuidos). Para números reales fijos θ_0 y θ_1 (parámetros), el modelo es el siguiente:

$$y_p = \theta_0 + \theta_1 x_p + e_p$$

La hipótesis del modelo lineal, $h(x, \theta)$ es:

$$\hat{y}_p \equiv h(x, \theta) = \theta_0 + \theta_1 x_p = \theta_1 x_p + \theta_0$$

Término de error o residual, e. El término de error o residual, es la parte sin explicar de la variable dependiente, representada en la ecuación de regresión como el *término de error aleatorio,* **e**. Los valores conocidos de la variable dependiente se utilizan para crear y calibrar el modelo de regresión. La diferencia entre los valores **y** observados y los valores $\hat{\boldsymbol{y}}$ previstos se llama residual. La magnitud de los residuales de una ecuación de regresión es una medida del ajuste del modelo. Los grandes residuales indican un ajuste del modelo pobre.

$$e_p = y_p - \hat{y}_p$$

Función de perdida, $\mathcal{L}(y_p, \hat{y}_p)$. La función de pérdida captura la diferencia entre los valores reales y predichos para un solo registro.

$$\mathcal{L}(y_p, \hat{y}_p) = (y_p - \hat{y}_p)^2 = e_p^2$$

Funciones de costo: La función de costo agrega la perdida para todo el conjunto de datos.

$$E \equiv SSE = \sum_{p=1}^{q} \mathcal{L}(y_p, \hat{y}_p) = \sum_{p=1}^{q} e_p^2$$

$$MSE = \frac{SSE}{q}$$

$$RMSE = \sqrt{MSE}$$

funciones gradiente, ∇*:*

$$\frac{\partial E}{\partial \theta_0} = 2\sum_{p=1}^{q} e_p \frac{\partial e_p}{\partial \hat{y}_p} \frac{\partial \hat{y}_p}{\partial \theta_0} = -2\sum_{p=1}^{q} e_p \frac{\partial \hat{y}_p}{\partial \theta_0} = -2\sum_{p=1}^{q} e_p$$

$$\frac{\partial E}{\partial \theta_1} = 2\sum_{p=1}^{q} e_p \frac{\partial e_p}{\partial \hat{y}_p} \frac{\partial \hat{y}_p}{\partial \theta_1} = -2\sum_{p=1}^{q} e_p \frac{\partial \hat{y}_p}{\partial \theta_1} = -2\sum_{p=1}^{q} e_p\, x_p$$

$$\nabla_E(\theta) = \begin{bmatrix} \frac{\partial E}{\partial \theta_1} \\ \frac{\partial E}{\partial \theta_0} \end{bmatrix}$$

$$\nabla_{MSE} = \frac{\nabla_E(\theta)}{q}$$

$$\nabla_{RMSE} = \frac{\nabla_E(\theta)}{2qRMSE}$$

Optimización de la función de costo, $E(\theta)$:

$$\hat{\theta} = arg\min_{\theta} E$$

$$\frac{\partial E}{\partial \theta_0} = -2\sum_{p=1}^{q}(y_p - \hat{y}_p) = -2\sum_{p=1}^{q}(y_p - [\hat{\theta}_0 + \hat{\theta}_1 x_p]) = 0$$

$$\frac{\partial E}{\partial \theta_1} = -2\sum_{p=1}^{q}(y_p - \hat{y}_p)\, x_p = -2\sum_{p=1}^{q}(y_p - [\hat{\theta}_0 + \hat{\theta}_1 x_p])\, x_p = 0$$

simplificando, obtenemos el sistema de ecuaciones lineales normales

$$q\,\hat{\theta}_0 + \left(\sum_{p=1}^{q} x_p\right)\hat{\theta}_1 = \sum_{p=1}^{q} y_p$$

$$\left(\sum_{p=1}^{q} x_p\right)\hat{\theta}_0 + \left(\sum_{p=1}^{q} x_p^2\right)\hat{\theta}_1 = \sum_{p=1}^{q} y_p x_p$$

$$\begin{bmatrix} q & \sum_{p=1}^{q} x_p \\ \sum_{p=1}^{q} x_p & \sum_{p=1}^{q} x_p^2 \end{bmatrix} \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix} = \begin{bmatrix} \sum_{p=1}^{q} y_p \\ \sum_{p=1}^{q} y_p x_p \end{bmatrix}$$

$$\boldsymbol{A\hat{\theta} = b}$$

$$\boldsymbol{\hat{\theta} = A^{-1}b}$$

donde

$$\boldsymbol{\hat{\theta}} = \begin{bmatrix} \hat{\theta}_0 \\ \hat{\theta}_1 \end{bmatrix}$$

$$\boldsymbol{A} = \begin{bmatrix} q & \sum_{p=1}^{q} x_p \\ \sum_{p=1}^{q} x_p & \sum_{p=1}^{q} x_p^2 \end{bmatrix}$$

$$\boldsymbol{b} = \begin{bmatrix} \sum_{p=1}^{q} y_p \\ \sum_{p=1}^{q} y_p x_p \end{bmatrix}$$

re-inscribiendo el modelo lineal simple en forma matricial, obtenemos:

Modelo lineal simple, $\hat{y}(\theta)$:

$$y_p = \theta_1 x_p + \theta_0 + e_p$$

donde

$$\hat{y}_p = \theta_1 x_p + \theta_0$$

En notación detallada tenemos:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \\ \vdots \\ y_q \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_p & 1 \\ \vdots & \vdots \\ x_q & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_0 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_p \\ \vdots \\ e_q \end{bmatrix}$$

$$\boldsymbol{y} = A\,\boldsymbol{\theta} + \boldsymbol{e}$$

donde

$$\hat{\boldsymbol{y}} = A\,\boldsymbol{\theta}$$

donde $\boldsymbol{A} = [\boldsymbol{x} \quad | \quad 1_{q\times 1}]$ es la matriz de diseño de tamaño $q \times 2$ y $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_0 \end{bmatrix}$ de tamaño 2×1

Término de error, $\boldsymbol{e}$:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

Funciones de costo:

En el caso que la función de perdida total cuadrática (suma de los errores al cuadrado, SSE) sin regularización, entonces la función de costo se define como:

$$SSE = \|\boldsymbol{e}\|_F^2 = tr(\boldsymbol{e}^{\mathsf{T}}\boldsymbol{e})$$

donde $\|\cdot\|_F$ se llama la norma de Frobenius o norma de Hilbert-Schmidt y $tr(\cdot)$ es la traza de una matriz cuadrada.

$$E \equiv SSE = \boldsymbol{e}'\boldsymbol{e}$$

$$MSE = \frac{SSE}{q}$$

$$RMSE = \sqrt{MSE}$$

funciones gradiente, ∇:

$$\nabla_E(\theta) = -2A'e$$

$$\nabla_{MSE} = \frac{\nabla_E(\theta)}{q}$$

$$\nabla_{RMSE} = \frac{\nabla_E(\theta)}{2qRMSE}$$

Optimización de la función de costo, $E(\theta)$:

$$\hat{\theta} = arg\min_{\theta} E$$

$$\nabla_E(\hat{\theta}) = -2A'e = 0$$

$$A'e = 0$$

$$A'(y - A\hat{\theta}) = 0$$

$$A'y - A'A\hat{\theta} = 0$$

$$A'y = A'A\hat{\theta}$$

Sistema de ecuaciones normales de regresión

Las ecuaciones normales son un conjunto de ecuaciones lineales que se utilizan para resolver los coeficientes de un modelo de regresión lineal. Estas ecuaciones se derivan estableciendo el gradiente de la suma de errores cuadrados en cero, lo que da como resultado un sistema de ecuaciones que se pueden resolver para los coeficientes. Las ecuaciones normales son particularmente útiles cuando la cantidad de características en los datos es relativamente pequeña y los datos no son demasiado grandes.

$$A'A\hat{\theta} = A'y$$

$$\hat{\theta} = (A'A)^{-1}A'y$$

Predicción

$$\hat{\boldsymbol{y}} = A\,\hat{\boldsymbol{\theta}}$$

Para estimar la varianza del término de error puede usarse la expresión:

$$\hat{e} = y - \hat{y} = y - A\,\hat{\theta}$$

$$\hat{\sigma}_e^2 = \frac{\hat{e}'\hat{e}}{q - \varrho}$$

Consecuentemente, la matriz de covarianzas del estimador OLS puede estimarse usando la expresión:

$$cov\left(\hat{\theta}\right) = \hat{\sigma}_e^2 (A'A)^{-1}$$

2. Modelo de regresión lineal múltiple

El modelo de Regresión Múltiple relaciona más de un predictor y una respuesta.

Sean q observaciones $\{x_{p,i}; y_p\}_{p=1}^{q}$ con n predictores ($\mathbf{X}$) y una respuesta ($\mathbf{y}$). Sea $\boldsymbol{y}$ el vector de respuesta q×1, $\boldsymbol{A}$ es la matriz de diseño $q \times \varrho$ tal que las primeras n columnas son los predictores ($\mathbf{X}$) y la última columna sean 1′s, es decir $\boldsymbol{A} = [\boldsymbol{X}|\boldsymbol{1_{q\times 1}}]$. Sea $\boldsymbol{e}$ un vector q×1 tal que $e_p \sim \mathcal{N}(0, \sigma^2)$ sean *iid* (independientes e idénticamente distribuidos), $\boldsymbol{\theta}$ sea un vector $\varrho \times 1$ de parámetros fijos y $\varrho = n + 1$. El modelo es el siguiente:

$$y_p = \theta_0 + \theta_1 x_{p,1} + \cdots + \theta_i x_{p,i} + \cdots + \theta_n x_{p,n} + e_p$$

$$y_p = \theta_1 x_{p,1} + \cdots + \theta_i x_{p,i} + \cdots + \theta_n x_{p,n} + \theta_0 + e_p$$

$$y_p = \sum_{i=1}^{n} \theta_i x_{p,i} + \theta_0 + e_p$$

donde

$$\hat{y}_p = \sum_{i=1}^{n} \theta_i x_{p,i} + \theta_0$$

Término de error, ***e****:*

$$e_p = y_p - \hat{y}_p$$

Función de perdida, $\mathcal{L}(y_p, \hat{y}_p)$

$$\mathcal{L}(y_p, \hat{y}_p) = (y_p - \hat{y}_p)^2 = e_p^2$$

Funciones de costo.

$$E \equiv SSE = \sum_{p=1}^{q} \mathcal{L}(y_p, \hat{y}_p) = \sum_{p=1}^{q} (y_p - \hat{y}_p)^2 = \sum_{p=1}^{q} e_p^2$$

$$MSE = \frac{SSE}{q}$$

$$RMSE = \sqrt{MSE}$$

funciones gradiente, ∇*:*

$$\frac{\partial E}{\partial \theta_1} = 2 \sum_{p=1}^{q} e_p \frac{\partial e_p}{\partial \hat{y}_p} \frac{\partial \hat{y}_p}{\partial \theta_1} = -2 \sum_{p=1}^{q} e_p \frac{\partial \hat{y}_p}{\partial \theta_1} = -2 \sum_{p=1}^{q} e_p \, x_{p,1}$$

$$\vdots$$

$$\frac{\partial E}{\partial \theta_i} = 2 \sum_{p=1}^{q} e_p \frac{\partial e_p}{\partial \hat{y}_p} \frac{\partial \hat{y}_p}{\partial \theta_i} = -2 \sum_{p=1}^{q} e_p \frac{\partial \hat{y}_p}{\partial \theta_i} = -2 \sum_{p=1}^{q} e_p \, x_{p,i}$$

$$\vdots$$

$$\frac{\partial E}{\partial \theta_n} = 2 \sum_{p=1}^{q} e_p \frac{\partial e_p}{\partial \hat{y}_p} \frac{\partial \hat{y}_p}{\partial \theta_n} = -2 \sum_{p=1}^{q} e_p \frac{\partial \hat{y}_p}{\partial \theta_n} = -2 \sum_{p=1}^{q} e_p \, x_{p,n}$$

$$\frac{\partial E}{\partial \theta_0} = 2 \sum_{p=1}^{q} e_p \frac{\partial e_p}{\partial \hat{y}_p} \frac{\partial \hat{y}_p}{\partial \theta_0} = -2 \sum_{p=1}^{q} e_p \frac{\partial \hat{y}_p}{\partial \theta_0} = -2 \sum_{p=1}^{q} e_p$$

$$\nabla_E(\theta) = \begin{bmatrix} \frac{\partial E}{\partial \theta_1} \\ \vdots \\ \frac{\partial E}{\partial \theta_i} \\ \vdots \\ \frac{\partial E}{\partial \theta_n} \\ \frac{\partial E}{\partial \theta_0} \end{bmatrix}$$

$$\nabla_{MSE} = \frac{\nabla_E(\theta)}{q}$$

$$\nabla_{RMSE} = \frac{\nabla_E(\theta)}{2qRMSE}$$

re-inscribiendo el modelo lineal simple en forma matricial, obtenemos:

$$y_p = \theta_1 x_{p,1} + \cdots + \theta_i x_{p,i} + \cdots + \theta_n x_{p,n} + \theta_0 + e_p$$

En notación detallada tenemos:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_p \\ \vdots \\ y_q \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,i} & \cdots & x_{1,n} & 1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ x_{p,1} & \cdots & x_{p,i} & \cdots & x_{p,n} & 1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ x_{q,1} & \cdots & x_{q,2} & \cdots & x_{q,n} & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_i \\ \vdots \\ \theta_n \\ \theta_0 \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_p \\ \vdots \\ e_q \end{bmatrix}$$

$$\boldsymbol{y} = \boldsymbol{A\theta} + \boldsymbol{e}$$

donde $\hat{\boldsymbol{y}} = \boldsymbol{A\theta}$, $\boldsymbol{A}$ es la matriz de diseño de tamaño $q \times \varrho$, $\boldsymbol{\theta}$ el vector de parámetros de tamaño $\varrho \times 1$ y $\varrho = n + 1$.

Término de error, $\boldsymbol{e}$:

$$\boldsymbol{e} = \boldsymbol{y} - \hat{\boldsymbol{y}}$$

Funciones de costo.

En el caso que la función de perdida total sea cuadrática (suma de los errores al cuadrado, SSE) sin regularización, entonces la función de costo se define como:

$$SSE = \|\boldsymbol{e}\|_F^2 = tr(\boldsymbol{e}^\intercal \boldsymbol{e})$$

donde $\|\cdot\|_F$ se llama la norma de Frobenius o norma de Hilbert-Schmidt y $tr(\cdot)$ es la traza de una matriz cuadrada.

$$E \equiv SSE = \boldsymbol{e}'\boldsymbol{e}$$

$$MSE = \frac{SSE}{q}$$

$$RMSE = \sqrt{MSE}$$

funciones gradiente, ∇:

$$\nabla_E(\theta) = -2\boldsymbol{A}'\boldsymbol{e}$$

$$\nabla_{MSE} = \frac{\nabla_E(\theta)}{q}$$

$$\nabla_{RMSE} = \frac{\nabla_E(\theta)}{2q\sqrt{MSE}}$$

Optimización de la función de costo, $E(\theta)$:

$$\hat{\theta} = arg\min_{\theta} E$$

$$\nabla_E(\hat{\theta}) = -2A'e = 0$$

$$A'e = 0$$

$$A'(y - A\hat{\theta}) = 0$$

$$A'y - A'A\hat{\theta} = 0$$

$$A'y = A'A\hat{\theta}$$

Sistema de ecuaciones normales de regresión

$$A'A\hat{\theta} = A'y$$

$$\hat{\theta} = (A'A)^{-1}A'y$$

Predicción

$$\hat{\boldsymbol{y}} = A\,\hat{\boldsymbol{\theta}}$$

3. Regresión multivariada

El modelo de regresión multivariada relaciona más de un predictor y más de una respuesta.

La regresión multivariada es un método utilizado para medir el grado en que más de una variable independiente (predictores) y más de una variable dependiente (respuestas) están relacionadas linealmente. El método se usa ampliamente para predecir el comportamiento de las variables de respuesta asociadas a los cambios en las variables predictoras, una vez que se ha establecido un grado de relación deseada.

Sean q observaciones $\{x_{p,i}; y_{p,j}\}_{p=1}^{q}$ con *n* predictores (**X**) y *m* respuestas (**Y**). Sea $\boldsymbol{Y}$ la matriz de respuestas q×m, $\boldsymbol{A}$ es la matriz de diseño de tamaño $q \times \varrho$ tal que las *n* primeras columnas son los predictores (**X**) y la última columna sean 1's, es decir $\boldsymbol{A} = [\boldsymbol{X}|\boldsymbol{1}_{\boldsymbol{q}\times\boldsymbol{1}}]$. Sea $\boldsymbol{\Theta}$ una matriz $\varrho \times m$ de parámetros fijos, $\boldsymbol{\Xi}$ una matriz q×m tal que $\boldsymbol{\Xi} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ (multivariada normalmente distribuida con matriz de covarianza $\boldsymbol{\Sigma}$) y $\varrho = n + 1$. El modelo es el siguiente:

$$y_{p,j} = \theta_{0,j} + \theta_{1,j}x_{p,1} + \cdots + \theta_{i,j}x_{p,i} + \cdots + \theta_{n,j}x_{p,n} + e_{p,j}$$

$$y_{p,j} = \theta_{1,j}x_{p,1} + \cdots + \theta_{i,j}x_{p,i} + \cdots + \theta_{n,j}x_{p,n} + \theta_{0,j} + e_{p,j}$$

donde

$$\hat{y}_{p,j} = \sum_{i=1}^{n} \theta_{i,j}x_{p,i} + \theta_{0,j}$$

Término de error, $\boldsymbol{\Xi}$*:*

$$e_{p,j} = y_{p,j} - \hat{y}_{p,j}$$

Función de perdida, $\mathcal{L}(y, \hat{y})$

$$\mathcal{L}(y_p, \hat{y}_p) = \sum_{j=1}^{m} (y_{p,j} - \hat{y}_{p,j})^2 = \sum_{j=1}^{m} e_{p,j}^2$$

Funciones de costo.

$$E \equiv SSE = \sum_{p=1}^{q} \mathcal{L}(y_p, \hat{y}_p) = \sum_{p=1}^{q}\sum_{j=1}^{m} \left(y_{p,j} - \hat{y}_{p,j}\right)^2 = \sum_{p=1}^{q}\sum_{j=1}^{m} e_{p,j}^2$$

$$MSE = \frac{SSE}{qm}$$

$$RMSE = \sqrt{MSE}$$

funciones gradiente, ∇:

$$\frac{\partial E}{\partial \theta_{0,j'}} = 2\sum_{p=1}^{q}\sum_{j=1}^{m} e_{p,j}\,\frac{\partial e_{p,j}}{\partial \hat{y}_{p,j}}\,\frac{\partial \hat{y}_{p,j}}{\partial \theta_{0,j'}} = -2\sum_{p=1}^{q}\sum_{j=1}^{m} e_{p,j}\,\frac{\partial \hat{y}_{p,j}}{\partial \theta_{0,j'}} = -2\sum_{p=1}^{q} e_{p,j}$$

$$\frac{\partial E}{\partial \theta_{i,j'}} = 2\sum_{p=1}^{q}\sum_{j=1}^{m} e_{p,j}\,\frac{\partial e_{p,j}}{\partial \hat{y}_{p,j}}\,\frac{\partial \hat{y}_{p,j}}{\partial \theta_{i,j'}} = -2\sum_{p=1}^{q}\sum_{j=1}^{m} e_{p,j}\,\frac{\partial \hat{y}_{p,j}}{\partial \theta_{i,j'}} = -2\sum_{p=1}^{q} e_{p,j}\,x_{p,i}$$

donde

$$\frac{\partial \hat{y}_{p,j}}{\partial \theta_{0,j'}} = \begin{cases} 0 & j \neq j' \\ x_{p,i} & otherwise \end{cases}$$

$$\nabla_E(\Theta) = \begin{bmatrix} \frac{\partial E}{\partial \theta_{1,1}} & \cdots & \frac{\partial E}{\partial \theta_{1,j}} & \cdots & \frac{\partial E}{\partial \theta_{1,m}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial \theta_{i,1}} & \cdots & \frac{\partial E}{\partial \theta_{i,j}} & \cdots & \frac{\partial E}{\partial \theta_{i,m}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial \theta_{n,1}} & \cdots & \frac{\partial E}{\partial \theta_{n,j}} & \cdots & \frac{\partial E}{\partial \theta_{n,m}} \\ \frac{\partial E}{\partial \theta_{0,1}} & \cdots & \frac{\partial E}{\partial \theta_{0,j}} & \cdots & \frac{\partial E}{\partial \theta_{0,m}} \end{bmatrix}$$, de tamaño (n+1) x m

$$\nabla_{MSE} = \frac{\nabla_E(\Theta)}{qm}$$

$$\nabla_{RMSE} = \frac{\nabla_E(\Theta)}{2qm\,RMSE}$$

re-inscribiendo el modelo lineal simple en forma matricial, obtenemos:

$$y_{p,j} = \theta_{1,j}x_{p,1} + \cdots + \theta_{i,j}x_{p,i} + \cdots + \theta_{n,j}x_{p,n} + \theta_{0,j} + e_{p,j}$$

En notación matricial tenemos:

$$\begin{bmatrix} y_{1,1} & \cdots & y_{1,j} & \cdots & y_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{p,1} & \cdots & y_{p,j} & \cdots & y_{p,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ y_{q,1} & \cdots & y_{q,j} & \cdots & y_{q,m} \end{bmatrix} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,i} & \cdots & x_{1,n} & 1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ x_{p,1} & \cdots & x_{p,i} & \cdots & x_{p,n} & 1 \\ \vdots & \ddots & \vdots & \ddots & \vdots & \vdots \\ x_{q,1} & \cdots & x_{q,2} & \cdots & x_{q,n} & 1 \end{bmatrix} \begin{bmatrix} \theta_{1,1} & \cdots & \theta_{1,j} & \cdots & \theta_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_{i,1} & \cdots & \theta_{i,j} & \cdots & \theta_{i,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \theta_{n,1} & \cdots & \theta_{n,j} & \cdots & \theta_{n,m} \\ \theta_{0,1} & \cdots & \theta_{0,j} & \cdots & \theta_{0,m} \end{bmatrix} + \begin{bmatrix} e_{1,1} & \cdots & e_{1,j} & \cdots & e_{1,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ e_{p,1} & \cdots & e_{p,j} & \cdots & e_{p,m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ e_{q,1} & \cdots & e_{q,j} & \cdots & e_{q,m} \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{A\Theta} + \boldsymbol{\Xi}$$

donde $\widehat{\boldsymbol{Y}} = \boldsymbol{A\Theta}$, donde $\boldsymbol{A}$ es la matriz de diseño de tamaño $q \times \varrho$, $\boldsymbol{\Theta}$ el vector de parámetros de tamaño $\varrho \times m$ y $\varrho = n + 1$.

Término de error, $\boldsymbol{\Xi}$*:*

$$\boldsymbol{\Xi} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

donde

$\boldsymbol{\Xi} = [e_{p,1} \quad \cdots \quad e_{p,j} \quad \cdots \quad e_{p,m}]$, de tamaño q x m

$\boldsymbol{Y} = [y_{p,1} \quad \cdots \quad y_{p,j} \quad \cdots \quad y_{p,m}]$, de tamaño q x m

$\widehat{\boldsymbol{Y}} = [\hat{y}_{p,1} \quad \cdots \quad \hat{y}_{p,j} \quad \cdots \quad \hat{y}_{p,m}]$, de tamaño q x m

Funciones de costo.

En el caso que la función de perdida total sea cuadrática (suma de los errores al cuadrado, SSE) sin regularización, entonces la función de costo se define como:

$$SSE = \|\boldsymbol{\Xi}\|_F^2 = tr(\boldsymbol{\Xi}^\intercal\boldsymbol{\Xi})$$

donde $\|\cdot\|_F$ se llama la norma de Frobenius o norma de Hilbert-Schmidt y $tr(\cdot)$ es la traza de una matriz cuadrada.

$$E \equiv SSE = \vec{\boldsymbol{\Xi}}'\vec{\boldsymbol{\Xi}}$$

donde la matriz $\boldsymbol{\Xi}$ vectorizada es $\vec{\boldsymbol{\Xi}}$ de tamaño mq x 1

$$MSE = \frac{SSE}{qm}$$

$$RMSE = \sqrt{MSE}$$

funciones gradiente, ∇*:*

$\nabla_E(\Theta) = -2\boldsymbol{A'\Xi}$, es una matriz de tamaño $\varrho \times m$

$$\nabla_{MSE} = \frac{\nabla_E(\Theta)}{qm}$$

$$\nabla_{RMSE} = \frac{\nabla_E(\Theta)}{2qm\sqrt{MSE}}$$

Optimización de la función de costo, $E(\Theta)$*:*

$\widehat{\Theta} = arg\min_{\Theta} E$

$\nabla E\left(\widehat{\Theta}\right) = -2A'e = 0$

$A'e = 0$

$A'\left(Y - A\widehat{\Theta}\right) = 0$

$A'Y - A'A\widehat{\Theta} = 0$

$A'Y = A'A\widehat{\Theta}$

Sistema de ecuaciones normales de regresión

$A'A\widehat{\Theta} = A'Y$

$\widehat{\Theta} = (A'A)^{-1}A'Y$

$\widehat{\Theta} = (A'A)^{+}A'Y$, donde la pseudoinversa de Moore-Penrose, $(A'A)^{+}$ de una matriz $A'A$ es una generalización de la matriz inversa.

Predicción

$\widehat{\boldsymbol{Y}} = A\,\widehat{\Theta}$

función gradiente de E vectorizada, $\nabla_E(\vec{\Theta})$:

$\nabla_E(\vec{\Theta}) = 2\, J_{\vec{\Xi}}' \, \vec{\Xi}$

donde la matriz Θ vectorizada es el vector $\vec{\Theta}$ de tamaño $\varrho m \times 1$. La matriz $\nabla_E(\Theta)$ vectorizada es el vector $\nabla_E(\vec{\Theta})$ de tamaño $\varrho m \times 1$. La matriz jacobiana de $\vec{\Xi}$ denotada, $J_{\vec{\Xi}}$, se define como: $J_{\vec{\Xi}} \equiv \frac{\partial \vec{\Xi}}{\partial \vec{\Theta}}$ de tamaño de $mq \times \varrho m$.

Entonces la jacobiana $J_{\vec{\Xi}}$ para m variables dependientes y ϱ parámetros es:

$$J_{\vec{\Xi}} \equiv \frac{\partial \vec{\Xi}}{\partial \vec{\Theta}} = -\begin{bmatrix} A_{q\times\varrho} & 0_{q\times\varrho} & \cdots & 0_{q\times\varrho} \\ 0_{q\times\varrho} & A_{q\times\varrho} & \cdots & 0_{q\times\varrho} \\ \vdots & \vdots & \ddots & \vdots \\ 0_{q\times\varrho} & 0_{q\times\varrho} & \cdots & A_{q\times\varrho} \end{bmatrix}$$

El algoritmo RLS (Recursive-Least-Squares algorithm) se usa en filtros adaptativos para encontrar los coeficientes del filtro que permiten obtener el mínimo cuadrado de la señal de error (definida como la diferencia entre la señal deseada y la señal producida a la salida del filtro) en forma recursiva.

Algoritmo RLS

(B) $\mathrm{P}_0 = \gamma I$, where $\gamma \to \infty$, $I \in \mathbb{R}^{\rho\times\rho}$
$\Theta = 0$

(R) for k=0,1,…,q-1

$$\mathrm{P}_{k+1} = \mathrm{P}_k - \frac{\mathrm{P}_k a_{k+1}' a_{k+1} \mathrm{P}_k}{1 + a_{k+1} \mathrm{P}_k a_{k+1}'}$$

$$\Theta_{k+1} = \Theta_k + \mathrm{P}_{k+1} a_{k+1}' (y_{k+1} - a_{k+1}\Theta_k)$$

end

where $(a_{k,:}; y_{k,:})$ is the kth row data pair, $A = [a_{k,j}]$, $a_{k,:} \in \mathbb{R}^{1\times\rho}$ and $y_{k,:} \in \mathbb{R}^{1\times m}$

4. Regresión polinomial multivariada

La regresión polinomial multivariada de orden superior se utiliza para modelar relaciones complejas con múltiples variables. Estas relaciones complejas suelen ser no lineales y de grandes dimensiones. Una vez que se crea o encuentra el modelo, esta ecuación se puede usar para futuras predicciones.

La regresión polinomial multivariada es una extensión de la regresión multivariada que permite múltiples variables de entrada y relaciones no lineales entre las variables de entrada y la variable

objetivo. En un modelo de regresión polinomial multivariada, las variables de entrada se elevan a diferentes potencias, creando una ecuación polinómica. Los coeficientes de la ecuación polinómica se determinan utilizando un proceso de optimización de mínimos cuadrados, al igual que en la regresión lineal.

El modelo de regresión polinomial multivariable de orden superior es una función polinómica multivariable de grado τ, $P_\tau(\boldsymbol{x})$ definida como:

$$P_\tau(\boldsymbol{x}) = \theta_0 + \sum_{l_1=1}^{n} \theta_{l_1} x_{l_1} + \sum_{l_1=1}^{n} \sum_{l_2=l_1}^{n} \theta_{l_1,l_2} x_{l_1} x_{l_2} + \sum_{l_1=1}^{n} \sum_{l_2=l_1}^{n} \cdots \sum_{l_\tau=l_{\tau-1}}^{n} \theta_{l_1,l_2,\ldots,l_\tau} x_{l_1} x_{l_2} \cdots x_{l_\tau}$$

con ϱ parámetros en relación al número de variables n y el grado del polinomio τ;

$$\varrho = \sum_{\ell=0}^{\tau} \frac{(\ell + n - 1)!}{(n-1)!\,\ell!}$$

donde $x_1, \ldots, x_n$ son las variables de entrada del sistema, el grado del polinomio es τ, es un número entero no negativo y θ son los coeficientes del polinomio, $P_\tau(\boldsymbol{x})$. Teniendo en cuenta que la función multivariable $P_\tau(x_1, \ldots, x_n)$ es lineal con respecto a sus coeficientes polinómicos, podemos adoptar la estimación estándar de mínimos cuadrados (OLS) para encontrar estos coeficientes polinómicos a partir de observaciones experimentales.

Al usar un modelo de regresión polinomial multivariada para modelar un sistema, se debe tener en cuenta que existen dos problemas técnicos principales: requerir grandes conjuntos de datos y que la matriz $\boldsymbol{A'A}$ sea singular.

Para obtener un modelo de regresión polinomial multivariada, a menudo se elige un polinomio de orden relativamente alto que tiene muchos coeficientes polinómicos para estimar. Esto significa que se necesitan muchos puntos de datos de entrenamiento. Esta solución obtendría el rendimiento estable relativo, porque los conjuntos de prueba son los conjuntos de datos experimentales completos. Dado que los términos no lineales $x_{l_1} x_{l_2} \cdots x_{l_\tau}$ para τ =1,2,… pueden dar lugar a la aparición de una matriz singular numérica, se utilizan los métodos de descomposición en valores singulares (SVD) y de estimación por mínimos cuadrados ordinario (OLS). Las solución esta dada por la siguiente ecuación.

$$\widehat{\boldsymbol{\Theta}} = (\boldsymbol{A'A})^{+}\boldsymbol{A'Y}$$

donde $\boldsymbol{A}$ es la matriz de diseño de tamaño $q \times \varrho$ e $\boldsymbol{Y}$ es la salida experimental de tamaño $q \times m$.

Predicción

$$\widehat{\boldsymbol{Y}} = A\,\widehat{\Theta}$$

Por ejemplo, para un modelo de regresión polinómica de grado τ de una variable, su hipótesis se define como:

$$\hat{y}_p = P_\tau(x) = \theta_0 + \theta_1 x_p + \theta_2 x_p^2 + \cdots + \theta_k x_p^k + \cdots + \theta_\tau x_p^\tau = \theta_0 + \sum_{k=1}^{\tau} \theta_k x_p^k$$

$$\hat{y} = \begin{bmatrix} 1 & x_p & x_p^2 & \cdots & x_p^k & \cdots & x_p^\tau \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \\ \vdots \\ \theta_\tau \end{bmatrix} = A\theta$$

El número de parámetros es $\varrho = \tau + 1$. El vector $\boldsymbol{\theta}$ es de tamaño $\varrho \times 1$. Entonces la matriz de diseño $\boldsymbol{A}$ de tamaño $q \times \varrho$ se define como;

$$A = \begin{bmatrix} 1 & x_p & x_p^2 & \cdots & x_p^k & \cdots & x_p^\tau \end{bmatrix}$$

Asimismo supongamos un polinomio de grado 2 y dos variables independientes, $P_2(x_1, x_2)$, descrito como:

$$\hat{y}_p = P_2(x_1, x_2) = \theta_0 + \theta_1 x_{1,p} + \theta_2 x_{2,p} + \theta_3 x_{1,p} x_{2,p} + \theta_4 x_{1,p}^2 + \theta_5 x_{2,p}^2$$

$$\hat{y} = \begin{bmatrix} 1 & x_{1,p} & x_{2,p} & x_{1,p}x_{2,p} & x_{1,p}^2 & x_{2,p}^2 \end{bmatrix} \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \theta_5 \end{bmatrix} = \boldsymbol{A\theta}$$

El número de parámetros es $\varrho = 6$. El vector $\boldsymbol{\theta}$ es de tamaño $\varrho \times 1$. Entonces la matriz de diseño $\boldsymbol{A}$ de tamaño $q \times 6$ se define como;

$$A = \begin{bmatrix} 1 & x_{1,p} & x_{2,p} & x_{1,p}x_{2,p} & x_{1,p}^2 & x_{2,p}^2 \end{bmatrix}$$

El siguiente algoritmo se usa para evaluar la matriz de diseño de regresión polinomial multivariada de orden superior.

Algorithm: ***designMatrix(*****X**, τ***)***
// Matriz de diseño para regresión polinomial multivariable de orden superior
Requerimientos:

- $\boldsymbol{X}_{q\times n}$: Datos de entrada
- τ : Grado del polinomio
- ϱ: Número de parámetros del modelo

Resultado: Matriz de diseño, $\boldsymbol{A}_{q\times\varrho}$

1. $\boldsymbol{A} \leftarrow \emptyset$
2. $q, n \leftarrow |\boldsymbol{X}|$
3. ***for*** $p \leftarrow 1$ ***to*** q
4. $M \leftarrow \boldsymbol{powerVector}\left(X_{p,:}, \tau\right)$
5. $\boldsymbol{A} \leftarrow \begin{bmatrix} \boldsymbol{A} \\ - \\ M \end{bmatrix}$
6. ***end***

est matriz de disenia es la que estaremos programando

Algorithm: ***powerVector(****V*, τ***)***
// Vector de potencias para regresión polinomial multivariable de orden superior
Requerimientos:

- $\boldsymbol{V}$: Vector renglón $X_{p,:}$ de tamaño $1 \times n$
- τ : Grado del polinomio

Resultado: Vector de potencias **M** de tamaño $1 \times \varrho$

1. ***if*** $|V| = 0$ ***or*** $\tau = 0$
2. $M \leftarrow 1$
3. ***else***
4. $M \leftarrow \emptyset$
5. $Z \leftarrow V_{1:n-1}$
6. $W \leftarrow V_n$
7. ***for*** $k \leftarrow 0$ ***to*** τ
8. $M \leftarrow [M \quad | \quad powerVector(Z, \tau - k) \odot W^k]$
9. ***end***
10. ***end***

5. Modelo de regresión no-lineal múltiple

El modelo de regresión no-lineal múltiple relaciona más de un predictor y una respuesta.

Sean q observaciones $\{x_{p,i}; y_p\}_{p=1}^{q}$ con *n*-variables predictoras ($X = [x_{p,i}]$ para i=1,…,n) y una variable de respuesta (**y**). Sea $\boldsymbol{y}$ el vector de respuesta q×1, $\boldsymbol{e}$ un vector q×1 tal que $e_p \sim \mathcal{N}(0, \sigma^2)$ sean *iid* (independientes e idénticamente distribuidos), $\boldsymbol{\theta}$ sea un vector $\varrho \times 1$ de parámetros. El modelo es el siguiente:

Modelo no-lineal múltiple:

$$y_p = f\left(x_{p,i}, \boldsymbol{\theta}\right) + e_p$$

donde $\hat{\boldsymbol{y}}_p = f(x_{p,i}, \boldsymbol{\theta})$, $\boldsymbol{X}$ es una matriz de tamaño $q \times n$, $\boldsymbol{y}$ es un vector de tamaño qx1 y $\boldsymbol{\theta} = \left(\theta_1, \dots, \theta_k \dots, \theta_\varrho\right)'$ de tamaño $\varrho \times 1$

Término de error, $\boldsymbol{e}$:

$e_p = y_p - \hat{y}_p$

Función de perdida, $\mathcal{L}(y_p, \hat{y}_p)$

$$\mathcal{L}(y_p, \hat{y}_p) = \left(y_p - \hat{y}_p\right)^2 = \left(y_p - f(x_{p,i}, \boldsymbol{\theta})\right)^2 = e_p^2$$

Funciones de costo.

$$E \equiv SSE = \sum_{p=1}^{q} \mathcal{L}(y_p, \hat{y}_p) = \sum_{p=1}^{q} \left(y_p - f(x_{p,i}, \boldsymbol{\theta})\right)^2 = \sum_{p=1}^{q} e_p^2$$

En el caso que la función de perdida total sea cuadrática (suma de los errores al cuadrado, SSE) sin regularización, entonces la función de costo se define como:

$$SSE = \|\boldsymbol{e}\|_F^2 = tr(\boldsymbol{e}^\intercal \boldsymbol{e})$$

donde $\|\cdot\|_F$ se llama la norma de Frobenius o norma de Hilbert-Schmidt y $tr(\cdot)$ es la traza de una matriz cuadrada.

$E \equiv SSE = \boldsymbol{e}'\boldsymbol{e}$

$$MSE = \frac{SSE}{q}$$

$$RMSE = \sqrt{MSE}$$

funciones gradiente, ∇:

$$\frac{\partial E}{\partial \theta_k} = 2\sum_{p=1}^{q} e_p \frac{\partial e_p}{\partial \theta_k} = 2\sum_{p=1}^{q} e_p \frac{\partial e_p}{\partial \hat{y}_p}\frac{\partial \hat{y}_p}{\partial \theta_k} = -2\sum_{p=1}^{q} e_p \frac{\partial \hat{y}_p}{\partial \theta_k}$$

$$\nabla_E(\theta) = \begin{bmatrix} \frac{\partial E}{\partial \theta_1} \\ \vdots \\ \frac{\partial E}{\partial \theta_k} \\ \vdots \\ \frac{\partial E}{\partial \theta_\varrho} \end{bmatrix}$$, de tamaño $\varrho \times 1$

$$\mathrm{J}_e(\theta) = \left[\frac{\partial e_p}{\partial \theta_1}, \ldots, \frac{\partial e_p}{\partial \theta_k}, \ldots, \frac{\partial e_p}{\partial \theta_\varrho}\right] = -\left[\frac{\partial \hat{y}_p}{\partial \theta_1}, \ldots, \frac{\partial \hat{y}_p}{\partial \theta_k}, \ldots, \frac{\partial \hat{y}_p}{\partial \theta_\varrho}\right]$$, de tamaño $qx\varrho$

donde $\frac{\partial e_p}{\partial \theta_k} = \frac{\partial e_p}{\partial \hat{y}_p}\frac{\partial \hat{y}_p}{\partial \theta_k} = -\frac{\partial \hat{y}_p}{\partial \theta_k}$

$$\nabla_E(\theta) = 2\,\mathrm{J}_e'\,e$$

$$\nabla_{MSE} = \frac{\nabla_E(\theta)}{q}$$

$$\nabla_{RMSE} = \frac{\nabla_E(\theta)}{2qRMSE}$$

Optimización de la función de costo, $E(\theta)$:

$$\hat{\theta} = arg\min_{\theta} E$$

Utilizar métodos de optimización numérica (SGD, Variable Learning Rate Gradient Descent (gdx), Adam, Levenberg-Marquardt, …etc) para minimizar la función de costo, $E(\theta)$

predicción

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$$

6. Regresión no-lineal multivariada

El modelo de regresión no-lineal multivariada relaciona más de un predictor y más de una respuesta.

La regresión no-lineal multivariada es un método utilizado para medir el grado en que más de una variable independiente (predictores) y más de una variable dependiente (respuestas) están relacionadas no-lineal. El método se usa ampliamente para predecir el comportamiento de las variables de respuesta asociadas a los cambios en las variables predictoras, una vez que se ha establecido un grado de relación.

Sean q observaciones $\{x_{p,i}; y_{p,j}\}_{p=1}^{q}$ con *n* variables predictoras ($\boldsymbol{X} = [x_{p,i}]$ para i=1,…,n) y *m* variables de respuesta ($\boldsymbol{Y} = [y_{p,j}]$ para j=1,…,m). Sea $\boldsymbol{Y}$ la matriz de respuestas q×m. Sea $\boldsymbol{\Theta}$ una matriz $\varrho \times m$ de parámetros, $\boldsymbol{\Xi}$ una matriz q×m tal que $\boldsymbol{\Xi} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ (multivariada normalmente distribuida con matriz de covarianza $\boldsymbol{\Sigma}$). El modelo es el siguiente:

$$
\begin{aligned}
y_{p,1} &= f_1(x_{p,i}, \boldsymbol{\theta}_{:,1}) + e_{p,1} \\
\dots \quad & \dots \quad \dots \\
y_{p,j} &= f_j(x_{p,i}, \boldsymbol{\theta}_{:,j}) + e_{p,j} \\
\dots \quad & \dots \quad \dots \\
y_{p,m} &= f_m(x_{p,i}, \boldsymbol{\theta}_{:,m}) + e_{p,j}
\end{aligned}
$$

Es decir

$\hat{y}_{p,j} = f_j(x_{p,i}, \theta_{k,j})$ para i=1,…,n; j=1,…,m y k=1,…, ϱ

Término de error, $\boldsymbol{\Xi}$*:*

$$e_{p,j} = y_{p,j} - \hat{y}_{p,j}$$

Función de perdida, $\mathcal{L}(y_p, \hat{y}_p)$

$$\mathcal{L}(y_p, \hat{y}_p) = \sum_{j=1}^{m} (y_{p,j} - \hat{y}_{p,j})^2 = \sum_{j=1}^{m} e_{p,j}^2$$

Funciones de costo.

$$E \equiv SSE = \sum_{p=1}^{q} \mathcal{L}(y_p, \hat{y}_p) = \sum_{p=1}^{q} \sum_{j=1}^{m} (y_{p,j} - \hat{y}_{p,j})^2 = \sum_{p=1}^{q} \sum_{j=1}^{m} e_{p,j}^2$$

$$MSE = \frac{SSE}{qm}$$

$$RMSE = \sqrt{MSE}$$

El modelo no-lineal multivariado general se describe como:

$$\boldsymbol{Y} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\Theta}) + \boldsymbol{\Xi}$$

donde $\widehat{\boldsymbol{Y}} = \boldsymbol{f}(\boldsymbol{X}, \boldsymbol{\Theta})$, $\boldsymbol{f}$ es un vector de funciones no-lineales multivariable

Término de error, $\boldsymbol{\Xi}$*:*

$$\boldsymbol{\Xi} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$$

donde

$\boldsymbol{\Xi} = [e_{p,1} \quad \cdots \quad e_{p,j} \quad \cdots \quad e_{p,m}]$, de tamaño q x m

$\boldsymbol{Y} = [y_{p,1} \quad \cdots \quad y_{p,j} \quad \cdots \quad y_{p,m}]$, de tamaño q x m

$\widehat{\boldsymbol{Y}} = [\hat{y}_{p,1} \quad \cdots \quad \hat{y}_{p,j} \quad \cdots \quad \hat{y}_{p,m}]$, de tamaño q x m

Funciones de costo.

En el caso que la función de perdida total sea cuadrática (suma de los errores al cuadrado, SSE) sin regularización, entonces la función de costo se define como:

$$SSE = \|\boldsymbol{\Xi}\|_F^2 = tr(\boldsymbol{\Xi}^\intercal \boldsymbol{\Xi})$$

donde $\|\cdot\|_F$ se llama la norma de Frobenius o norma de Hilbert-Schmidt y $tr(\cdot)$ es la traza de una matriz cuadrada.

$$E \equiv SSE = \vec{\boldsymbol{\Xi}}' \vec{\boldsymbol{\Xi}}$$

donde la matriz $\boldsymbol{\Xi}$ vectorizada es $\vec{\boldsymbol{\Xi}}$ de tamaño mq x 1

$$MSE = \frac{SSE}{qm}$$

$$RMSE = \sqrt{MSE}$$

funciones gradiente, ∇*:*

$$\nabla_E(\Theta) = \begin{bmatrix} \frac{\partial E}{\partial \theta_{1,1}} & \cdots & \frac{\partial E}{\partial \theta_{1,j}} & \cdots & \frac{\partial E}{\partial \theta_{1,m}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial \theta_{k,1}} & \cdots & \frac{\partial E}{\partial \theta_{k,j}} & \cdots & \frac{\partial E}{\partial \theta_{k,m}} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial E}{\partial \theta_{\varrho,1}} & \cdots & \frac{\partial E}{\partial \theta_{\varrho,j}} & \cdots & \frac{\partial E}{\partial \theta_{\varrho,m}} \end{bmatrix}$$, de tamaño ϱ x m

$$\nabla_{MSE} = \frac{\nabla_E(\Theta)}{qm}$$

$$\nabla_{RMSE} = \frac{\nabla_E(\Theta)}{2qm\,RMSE}$$

Optimización de la función de costo, $E(\theta)$:

$$\hat{\theta} = arg\,\min_{\theta} E$$

Utilizar métodos de optimización numérica (SGD, Variable Learning Rate Gradient Descent (gdx), Adam, Levenberg-Marquardt, …etc) para minimizar la función de costo, $E(\theta)$

predicción

$$\hat{\boldsymbol{y}} = f(\boldsymbol{x}, \hat{\boldsymbol{\theta}})$$

7. Regularización

En muchas técnicas de aprendizaje automático, el aprendizaje consiste en encontrar los coeficientes que minimizan una función de coste. La regularización consiste en añadir una penalización a la función de coste. Esta penalización produce modelos más simples que generalizan mejor. Las regularizaciones más usadas en machine learning: Lasso (conocida como L1), Ridge (conocida también como L2) y ElasticNet que combina tanto Lasso como Ridge.

Una formulación genérica de las técnicas de regularización en el contexto de modelos lineales puede realizarse de la siguiente manera:

$$J(\Theta) = \sum_{p=1}^{q} \mathcal{L}\left(y_p, \hat{y}_p\right) + \lambda\Omega(\Theta)$$

donde $\Omega(\cdot)$ es la función de regularización, para $\lambda \geq 0$.

La ventaja de utilizar lasso es que va a forzar a que algunos parámetros sean cero, con lo cual también se realiza una selección de las variables más influyentes. Por el contrario, ridge regression va a incluir todas las variables predictoras en el modelo final, si bien es cierto que algunas con parámetros muy próximos a cero: de este modo va a reducir el riesgo del sobreajuste, pero no resuelve el problema de la interpretabilidad. Otra posible ventaja de utilizar lasso es que cuando hay variables predictoras correlacionadas tiene tendencia a seleccionar una y anular las demás (esto también se puede ver como un inconveniente, ya que pequeños cambios en los datos pueden dar lugar a distintos modelos), mientras que ridge tiende a darles igual peso.

Resumen de técnicas regularización.

1. Ordinary Least Squares (OLS) Regression

$$\hat{\theta} = \min_{\theta} \|y - A\theta\|_2^2 = \min_{\theta} SSE$$

2. Ridge regression

esta es la tecnica que se usa

$$\hat{\theta} = \min_{\theta} \|y - A\theta\|_2^2 + \lambda\|\theta\|_2^2 = \min_{\theta} SSE + \lambda\|\theta\|_2^2$$

3. Lasso regression

$$\hat{\theta} = \min_{\theta} \frac{1}{2q}\|y - A\theta\|_2^2 + \lambda\|\theta\|_1 = \min_{\theta} \frac{1}{2} MSE + \lambda\|\theta\|_1$$

4. Multi-Task-Lasso Regression

$$\hat{\theta} = \min_{\theta} \frac{1}{2q}\|y - A\theta\|_{Fro}^2 + \lambda\|\theta\|_{21}$$

donde $\|X\|_{Fro} = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} x_{i,j}^2} = \sqrt{trace(X'X)}$, $\|X\|_{21} = \sum_{i=1}^{m}\sqrt{\sum_{j=1}^{n} x_{i,j}^2}$

5. Elastic-Net Regression

$$\hat{\theta} = \min_{\theta} \frac{1}{2q}\|y - A\theta\|_2^2 + \lambda\rho\|\theta\|_1 + \lambda\left(\frac{1-\rho}{2}\right)\|\theta\|_2^2$$

6. Multi-Task Elastic-Net

$$\hat{\theta} = \min_{\theta} \frac{1}{2q}\|y - A\theta\|_{Fro}^2 + \lambda\rho\|\theta\|_{21} + \lambda\left(\frac{1-\rho}{2}\right)\|\theta\|_{Fro}^2$$

7. Huber Regression

$$\hat{\theta} = \min_{\theta} \sum_{p=1}^{q}\left(\sigma + H_{\epsilon}\left(\frac{y_p - A_{p,:}\theta}{\sigma}\right)\sigma\right) + \lambda\|\theta\|_2^2$$

$$H_{\epsilon}(z) = \begin{cases} z^2 & if\ |z| < \epsilon \\ 2\epsilon|z| - \epsilon^2 & otherwise \end{cases}$$

Se recomienda establecer el parámetro épsilon (ϵ) en 1.35 para lograr una eficiencia estadística del 95 %.

8. Quantile Regression

La regresión de cuantiles estima la mediana u otros cuantiles de **y** condicional en x, mientras que los mínimos cuadrados ordinarios (OLS) estiman la media condicional.

Como modelo lineal, Quantile Regressor proporciona predicciones lineales $\hat{y} = A\theta$ para el q-ésimo cuantil, $q \in (0,1)$. Luego, los pesos o coeficientes se encuentran mediante el siguiente problema de minimización:

$$\hat{\theta} = \min_{\theta} \frac{1}{N} \sum_{p=1}^{N} PB_q(y_p - A_{p,:}\theta) + \lambda \|\theta\|_1$$

Esto consiste en la pérdida de pinball (también conocida como pérdida lineal), $PB_q(t)$:

$$PB_q(t) = q\ max(t,0) + (1-q)\max(-t,0) = \begin{cases} qt & t > 0 \\ 0 & t = 0 \\ (q-1)t & t < 0 \end{cases}$$

Como la pérdida pinball es solo lineal en los residuos, la regresión por cuantiles es mucho más sólida para los valores atípicos que la estimación de la media basada en el error cuadrático.

La regresión de cuantiles puede ser útil si uno está interesado en predecir un intervalo en lugar de una predicción puntual. A veces, los intervalos de predicción se calculan basándose en la suposición de que el error de predicción se distribuye normalmente con media cero y varianza constante. La regresión cuantil proporciona intervalos de predicción sensibles incluso para errores con varianza no constante (pero predecible) o distribución no normal.

8. Processing Functions

Processes matrices by normalizing the minimum and maximum values of each row to *apply* [y_{min}, y_{max}]

$$y = (y_{max} - y_{min}) \frac{x - x_{min}}{x_{max} - x_{min}} + y_{min}$$

Processes matrices by reverse

$$x = (x_{max} - x_{min}) \frac{y - y_{min}}{y_{max} - y_{min}} + x_{min}$$

Processes matrices by transforming the mean and standard deviation of each row to *apply* $\bar{y}$ and $\boldsymbol{\sigma_y}$

$$y = \frac{\boldsymbol{\sigma_y}}{\boldsymbol{\sigma_x}}(x - \bar{x}) + \bar{y}$$

Processes matrices by reverse

$$x = \frac{\boldsymbol{\sigma_x}}{\boldsymbol{\sigma_y}}(y - \bar{y}) + \bar{x}$$

9. Aprendizaje supervisado

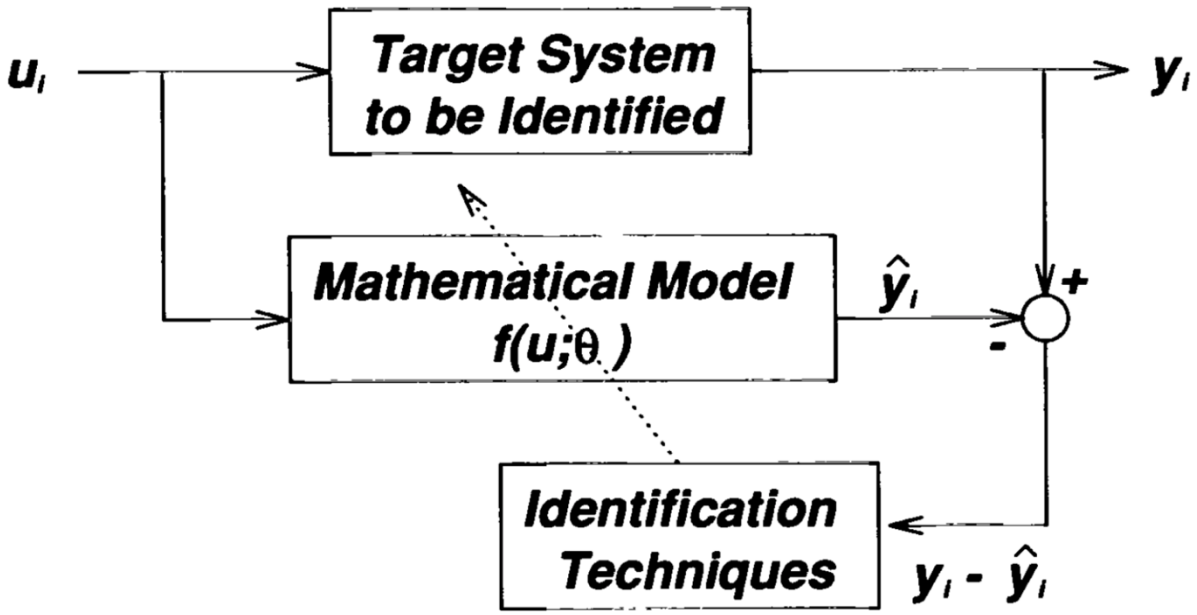

Figure 5.1. *Block diagram for parameter identification.*

En el aprendizaje supervisado, se presenta una entrada ($\boldsymbol{u}$) al modelo junto con una respuesta deseada ($\boldsymbol{y}$), el proceso de aprendizaje (optimización) se basa en la comparación entre la salida del modelo ($\hat{\boldsymbol{y}}$) y la respuesta deseada ($\boldsymbol{y}$), generando un error ($\boldsymbol{e = y - \hat{y}}$), el error se utiliza para cambiar los parámetros ($\boldsymbol{\theta}$) del modelo de modo que resulte un mejor rendimiento.

9.1 Gradiente descendente estocástico, lote y mini lote

Gradiente descendente estocástico

El gradiente descendente estocástico (SGD), calcula el error para cada dato de entrenamiento y ajusta los parámetros o pesos inmediatamente. Si tenemos 100 puntos de datos de entrenamiento, el SGD ajusta los parámetros 100 veces. La Figura 2-15 muestra cómo la actualización de los parámetros del SGD se relaciona con todos los datos de entrenamiento.

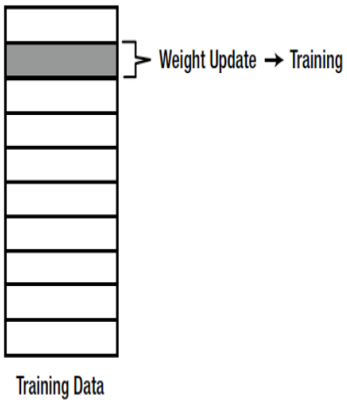

Figure 2-15. How the weight update of the SGD is related to the entire training data

A medida que el SGD ajusta los parámetros para cada punto de datos, el rendimiento del modelo se actualiza mientras se realiza el proceso de entrenamiento. El nombre "estocástico" implica el comportamiento aleatorio del proceso de entrenamiento. El SGD calcula las actualizaciones de los parámetros del modelo, $\boldsymbol{\theta}$, como:

$$\Delta\boldsymbol{\theta} = -\eta\nabla_{\boldsymbol{\theta}}E$$

Esta ecuación implica que la regla delta se basa en el enfoque SGD.

Gradiente descendente por lotes

En el gradiente descendente por lotes, cada actualización de los parámetros ($\Delta\boldsymbol{\theta}$) se calcula para todos los errores de los datos de entrenamiento y el promedio de las actualizaciones de los parámetros se usan para ajustar los parámetros. Este método utiliza todos los datos de entrenamiento y se actualiza solo una vez por cada iteración (época). La Figura 2-16 explica el cálculo de actualización de los parámetros y el proceso de entrenamiento del método por lotes.

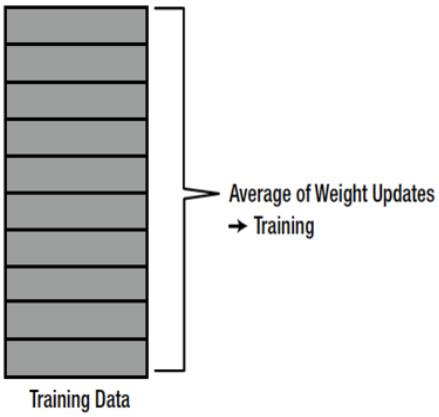

Figure 2-16. The batch method's weight update calculation and training process

El método por lotes calcula la actualización de peso como:

$$^{avg}\Delta\boldsymbol{\theta} = \frac{1}{q}\sum_{p=1}^{q}\Delta_p\boldsymbol{\theta}$$

donde $\Delta_p\boldsymbol{\theta}$ es la actualización de peso para el p-ésimo dato de entrenamiento y q es el número total de datos de entrenamiento del lote. Debido al cálculo de actualización de peso promedio, el método por lotes consume una cantidad significativa de tiempo para el entrenamiento.

Gradiente descendente por mini lotes

El método de mini lotes es una combinación del SGD y el método de lotes. Selecciona una parte del conjunto de datos de entrenamiento y los usa para entrenar en el método por lotes. Por lo tanto, calcula las actualizaciones $\Delta_p\boldsymbol{\theta}$ de los parámetros de los datos seleccionados y optimiza el modelo con la actualización de peso promedio, $^{avg}\Delta\boldsymbol{\theta}$. Por ejemplo, si se seleccionan 20 puntos de datos arbitrarios de 100 puntos de datos de entrenamiento, el método por lotes se aplica a los 20 puntos de datos. En este caso, se realizan un total de cinco ajustes de parámetros para completar el proceso de entrenamiento para todos los puntos de datos (5 = 100/20). La Figura 2-17 muestra cómo el esquema de mini lotes selecciona los datos de entrenamiento y calcula la actualización de peso.

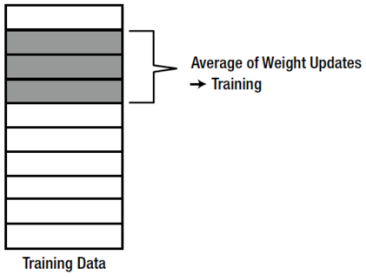

Figure 2-17. *How the mini batch scheme selects training data and calculates the weight update*

El método de mini lotes, cuando selecciona un número apropiado de puntos de datos, obtiene los beneficios de ambos métodos: velocidad del SGD y estabilidad del lote. Por esta razón, a menudo se utiliza en aprendizaje profundo, que manipula una cantidad significativa de datos.

Ahora, profundicemos un poco en el SGD, el lote y el mini lote en términos de la época. La época es el número de ciclos de entrenamiento completados para todos los datos de entrenamiento. En el método por lotes, el número de ciclos de entrenamiento del modelo es igual a una época. Esto tiene mucho sentido, porque el método por lotes utiliza todos los datos para un proceso de entrenamiento.

Por el contrario, en el mini lote, el número de procesos de entrenamiento para una época varía según el número de puntos de datos en cada lote. Cuando tenemos q puntos de datos de entrenamiento en total, el número de procesos de entrenamiento por época es mayor que uno, que corresponde al método por lotes, y menor que q, que corresponde al SGD.