

Parcial TAM

Juan Esteban Villada Sierra

DD MM AA
20 05 25

Preguntas:

II) Sea el modelo de regresión $t_n = \phi(x_n)w^T + r_n$, con $t_n \in \mathbb{R}$, $x_n \in \mathbb{R}^P$, $\{x_n\}_{n=1}^N$, $w \in \mathbb{R}^Q$, $\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$, $Q \geq P$, y $r_n \sim N(r_n | 0, \sigma_r^2)$.

Entonces PROBLEMA DE OPTIMIZACIÓN:

Modelo de mínimos cuadrados

Sea el modelo $t_n = \phi(x_n)w^T + r_n$ donde:

$t_n \in \mathbb{R}$ — Es la variable objetivo (salida) para la n -ésima observación.

$x_n \in \mathbb{R}^P$ — Vector de características (entrada) para la n -ésima observación.

$w \in \mathbb{R}^Q$ — Vector de pesos (parámetros) que queremos "APRENDER".

$\phi: \mathbb{R}^P \rightarrow \mathbb{R}^Q$ — Es una función de base que mapea el espacio de entrada a un espacio de características de mayor o igual dimensión ($Q \geq P$)

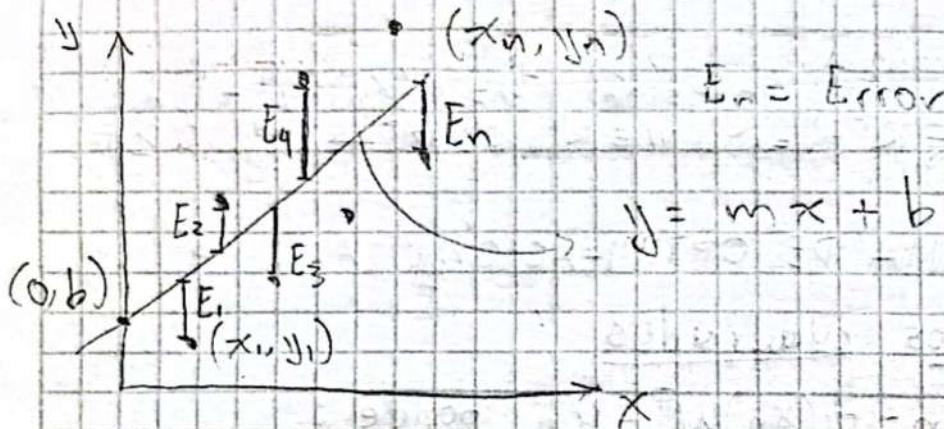
$r_n \sim N(r_n | 0, \sigma_r^2)$ — Ruido aditivo gaussiano con media cero y varianza σ_r^2

N datos — Independientes e identicamente distribuidos (i.i.d) $\{x_n, t_n\}_{n=1}^N$

Nuestro objetivo es estimar el vector de pesos w basándonos en los datos observados.

El objetivo en mínimos cuadrados es encontrar el vector de pesos w que minimice la suma de los errores cuadráticos entre las predicciones del modelo y los valores objetivo observados.

Teniendo en cuenta y recordando:



$$E_n = \text{Error}$$

$$y = mx + b$$

De este modo para $E_1 = y_1 - (mx_1 + b)$

$$E_2 = y_2 - (mx_2 + b)$$

$$E_n = y_n - (mx_n + b)$$

Por lo tanto el Error cuadrático de esta linea:

$$E_{\text{linea}} = (y_1 - (mx_1 + b))^2 + (y_2 - (mx_2 + b))^2 + \dots + (y_n - (mx_n + b))^2$$

Encontrando los m y b que minimizan E_{linea}

$$E_{\text{cl}} = [y_1^2 - 2y_1(mx_1 + b) + (mx_1 + b)^2] + \dots +$$

$$[y_n^2 - 2y_n(mx_n + b) + (mx_n + b)^2]$$

$$= \dots + y_1^2 - 2y_1mx_1 - 2y_1b + m^2x_1^2 + 2mx_1b + b^2 +$$

$$y_2^2 - 2y_2mx_2 - 2y_2b + m^2x_2^2 + 2mx_2b + b^2 +$$

$$\vdots$$
$$y_n^2 - 2y_nmx_n - 2y_nb + m^2x_n^2 + 2mx_nb + b^2$$

Sumando las columnas:

$$\begin{aligned} c &= (y_1^2 + y_2^2 + \dots + y_n^2) + 2m(x_1y_1 + x_2y_2 + \dots + x_ny_n) \\ &\quad - 2b(y_1 + y_2 + \dots + y_n) + m^2(x_1^2 + x_2^2 + \dots + x_n^2) \\ &\quad + 2mb(x_1 + x_2 + \dots + x_n) + nb^2 \end{aligned}$$

Simplificando

$$\bar{y}^2 = \bar{y}_1^2 + \bar{y}_2^2 + \dots + \bar{y}_n^2 \Rightarrow n\bar{y}^2 = \sum_{i=1}^n y_i^2$$

$$\bar{xy} = \frac{x_1y_1 + x_2y_2 + \dots + x_ny_n}{n} \Rightarrow n\bar{xy} = \sum_{i=1}^n x_iy_i$$

$$\bar{y} = \bar{y}_1 + \bar{y}_2 + \dots + \bar{y}_n \Rightarrow n\bar{y} = \sum_{i=1}^n y_i$$

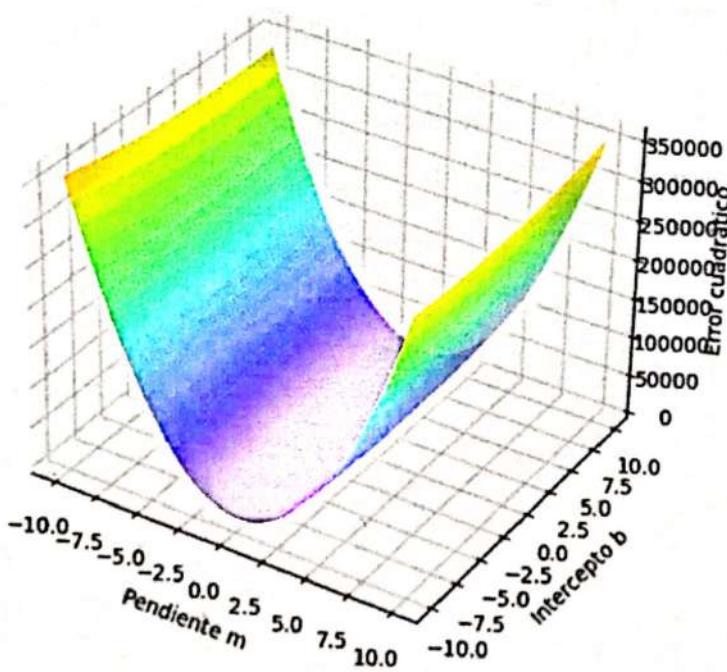
Así mismo para \bar{x}^2 y \bar{x}

Por lo tanto

$$c = n\bar{y}^2 - 2m(n\bar{xy}) - 2b(n\bar{y}) + m^2(n\bar{x}^2) + 2mb(n\bar{x}) + nb^2$$

Optimizando

Superficie del error cuadrático



Por lo tanto
para hallar el
punto crítico

$$\frac{\partial S E}{\partial m} = 0$$

$$\frac{\partial S E}{\partial b} = 0$$

donde $SE = \text{error cuadra-} + \text{tico}$

Entonces derivando con respecto a m :

$$\frac{\partial Ecl}{\partial m} = \frac{\partial n\bar{y}^2}{\partial m} - \frac{\partial 2mn\bar{xy}}{\partial m} - \frac{\partial 2bn\bar{x}}{\partial m} + \frac{\partial mn^2\bar{x}^2}{\partial m} \\ + \frac{\partial 2mnb\bar{x}}{\partial m} + \frac{\partial nb^2}{\partial m}$$

$$\frac{\partial Ecl}{\partial m} = 0 - 2n\bar{y}\bar{y} - 0 + 2mn\bar{x}^2 + 2bn\bar{x} + 0$$

Como $Ecl = SE$ (error cuadrático) entonces:

$$\frac{\partial Ecl}{\partial m} = \frac{\partial SE}{\partial m} \quad \text{Por lo tanto, necesitamos que: } \frac{\partial SE}{\partial m} = 0$$

$$\text{Donde } -2n\bar{y}\bar{y} + 2mn\bar{x}^2 + 2bn\bar{x} = 0 \quad [1]$$

Y también, derivando con respecto a b :

$$\frac{\partial SE}{\partial b} = 0 \quad \text{donde:}$$

$$\frac{\partial SE}{\partial b} = 0 - 0 - 2n\bar{y} + 0 + 2mn\bar{x} + 2bn$$

$$\text{Así: } -2n\bar{y} + 2mn\bar{x} + 2bn = 0 \quad [2]$$

De este modo nos queda un sistema lineal con 2 ecuaciones donde dividiremos entre $2n$ para: [1] y [2]:

$$- \bar{y} + m\bar{x}^2 + b\bar{x} = 0 \quad [3]$$

$$- \bar{y} + m\bar{x} + b = 0 \quad [4]$$

Para poder decir algo en términos de la recta ideal quiero pasar ambas ecuaciones a la forma:

$$y = mx + b \quad \text{por lo tanto.}$$

$$\bar{y} = m\bar{x}^2 + b\bar{x} \quad [5] \quad \bar{y} = m\bar{x} + b \quad [6]$$

Por que se hace esto? Supongamos que en nuestro modo óptimo ya está hecha por:

$$y = mx + b \quad \text{donde } m \text{ y } b \text{ ya son óptimas}$$

Por lo tanto m y b son las mismas que satisfacen el sistema de ecuaciones.

Justo por eso, podemos concluir que la recta óptima b sea la linea que mejor se ajusta contiene al punto (\bar{x}, \bar{y}) que está sobre la recta.

Es decir que la recta óptima va a contener el punto donde $x = \bar{x}$ y $y = \bar{y}$

seg que la recta tiene al punto cuyas entradas son los promedios de los datos que nos dan.

Por lo tanto, para [5]:

$$\bar{y} = m\bar{x}^2 + b\bar{x} \quad \text{donde dividendo entre } \bar{x}:$$

$$\frac{m\bar{x}^2}{\bar{x}} + b = \frac{\bar{y}}{\bar{x}} \quad [7] \quad \text{por lo tanto el punto: } \left(\frac{\bar{x}^2}{\bar{x}}, \frac{\bar{y}}{\bar{x}} \right) \text{ está}$$

Sobre la linea óptima

De este modo queremos hallar m

$$m\bar{x}^2 + b\bar{x} = \bar{y} \quad [5] \Rightarrow [7] \quad m\frac{\bar{x}^2}{\bar{x}} + b = \frac{\bar{y}}{\bar{x}}$$

$$m\bar{x} + b = \bar{y} \quad [6]$$

$$[6] - [7] = m\left(\bar{x} - \frac{\bar{x}^2}{\bar{x}}\right) = \bar{y} - \frac{\bar{y}}{\bar{x}}$$

$$m = \frac{\bar{y} - \frac{\bar{y}}{\bar{x}}}{\bar{x} - \frac{\bar{x}^2}{\bar{x}}}$$

Observar que es lo mismo que obtener la pendiente de la recta

Entonces multiplicando en $m = \frac{\bar{x}\bar{y} - \bar{xy}}{(\bar{x})^2 - \bar{x}^2}$

y así se puede llegar

$$b = \bar{y} - m\bar{x}$$

Entonces definimos la función de error de suma de cuadrados:

$$Efs(w) = \frac{1}{2} \sum_{n=1}^N (t_n - \phi(x_n)w^\top)^2$$

El factor $\frac{1}{2}$ se incluye por conveniencia matemática al tomar la derivada.

Solución

Para encontrar el w que minimiza $Efs(w)$, tomamos el gradiente de Efs con respecto a w y lo igualamos a cero. Por lo visto anteriormente

$$\nabla_w Efs(w) = \sum_{n=1}^N (t_n - \phi(x_n)w^\top)(-\phi(x_n)) = 0$$

Reorganizando la ecuación:

$$\sum_{n=1}^N \phi(x_n)w^\top \phi(x_n)^\top = \sum_{n=1}^N t_n \phi(x_n)$$

Entonces sea Φ la matriz de diseño de $N \times Q$, donde la n -ésima fila es $\phi(x_n)$:

$$\Phi = \begin{pmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_n) \end{pmatrix} \quad \text{y} \quad \text{sea } t = \begin{pmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{pmatrix} \quad \text{el vector de valores objetivo}$$

Entonces la ecuación anterior se puede escribir en forma matricial como:

$$\sum \Phi^T \Phi w = \sum \Phi^T t$$

Si la matriz $\Phi^T \Phi$ es invertible, la solución para w_{ls} es:

$$w_{ls} = (\Phi^T \Phi)^{-1} \Phi^T t$$

La matriz $(\Phi^T \Phi)^{-1} \Phi^T$ se conoce como la pseudo-inversa de Moore-Penrose de Φ cuando $N \geq Q$ o Φ no tiene rango completo.

Si $N \geq Q$ y Φ tiene rango completo, entonces $(\Phi^T \Phi)^{-1} \Phi^T = A$

$$A = (\Phi^{-1})^T (\Phi^T)^{-1} \Phi^T = \Phi^{-1}$$

Modelo de mínimos cuadrados regularizados (RLS)

* PROBLEMA:

Para evitar el sobreajuste, especialmente cuando $Q > P$ o tenemos pocos datos, se añade un término de regularización a la función de error de mínimos cuadrados. Una forma común de regularización es la regularización L2 (también conocida como regularización de Ridge):

$$E_{RLS}(w) = \frac{1}{2} \sum_{n=1}^N (t_n - \phi(x_n)w^T)^2 + \frac{\lambda}{2} w^T w$$

donde $\lambda > 0$ es el parámetro de regularización que controla la fuerza de la penalización sobre la magnitud de los pesos

Solución: Similar al caso de mínimos cuadrados, tomamos el gradiente de $E_{RLS}(w)$ con respecto a w y lo igualamos a cero

$$\nabla_w E_{RLS}(w) = \sum_{n=1}^N (t_n - \phi(x_n)^T w)^T (-\phi(x_n)) + \lambda w = 0$$

Reorganizando la ecuación en forma matricial:

$$\Phi^T \Phi w + \lambda I w = \Phi^T t$$

$$(\Phi^T \Phi + \lambda I) w = \Phi^T t$$

La solución para w_{RLS} es:

$$w_{RLS} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T t$$

La adición del término λI (donde I es la matriz identidad de $Q \times Q$) asegura que la matriz $(\Phi^T \Phi + \lambda I)$ siempre sea invertible si $\lambda > 0$, incluso si $\Phi^T \Phi$ no lo es. Esto estabiliza la solución y reduce la varianza de los pesos.

Haciendo una analogía con el modelo de regresión simple, podemos notar que:

$$\text{con } y = mx + b$$

y : Analoga a t_n . Representa la salida que queremos predecir.

x : Analoga a un componente de x_n . Representa la entrada o información que usamos para hacer la predicción.

m y b : Analoga a un componente de w . Son parámetros que aprendemos del modelo (el b podría estar implícito en $\phi(x_n)$).

Por lo tanto, para nuestro modelo de regresión general:

$\phi(x_n)$: Es la función de base que transforma el vector de entrada x_n a un nuevo espacio de características de dimensión Q .

Si $\phi(x_n) = \begin{pmatrix} 1 \\ x_n \\ x_n^2 \end{pmatrix}$ y $w = \begin{pmatrix} b \\ m \end{pmatrix}$, entonces $\phi(x_n)w = \begin{pmatrix} 1 \\ x_n \\ x_n^2 \end{pmatrix} \begin{pmatrix} b \\ m \end{pmatrix} = b + mx_n + mx_n^2$, que es la forma simple de la regresión.

La función de base ϕ permite que el modelo capture relaciones no lineales entre las entradas y la salida. Por ejemplo si:

$\phi(x_n) = \begin{pmatrix} 1 \\ x_n \\ x_n^2 \end{pmatrix}$, el modelo puede aprender una relación cuadrática.

w es el vector de pesos de dimensión Q . Cada componente de w corresponde a una de las características transformadas por $\phi(x_n)$.

$\phi(x_n)w^T$: Representa la predicción del modelo para la n -ésima observación, basada en las características transformadas y los pesos aprendidos. Análogo a $mx + b$.

w^T : Transpuesta para convertirlo en vector fila que así la multiplicación con $\phi(x_n)$ sea un producto escalar.

En la regresión lineal simple con regularización L₂ (Ridge) añadimos un término de penalización a la magnitud de los parámetros m y b :

$$\text{Eridge}(m, b) = \frac{1}{2} \sum_{n=1}^N (y_n - (mx_n + b))^2 + \frac{\lambda}{2} (m^2 + b^2)$$

donde $\lambda > 0$ por lo tanto

La analogía:

m (pendiente) \Rightarrow w (vector de pesos)

b (intersección) \Rightarrow w (m , b) — asumiendo que $t(x_n)$ incluye un término constante como 1^T)

$\frac{\lambda}{2} (m^2 + b^2)$ (Término de regularización L2 penalizando la magnitud de m y b)

$$\frac{1}{2} w^T w = \frac{\lambda}{2} \sum_{j=1}^Q w_j^2$$

Término de regularización L2: La clave de la analogía reside en el segundo término.

- * En la regresión lineal simple con Ridge, penalizamos la suma de los cuadrados de la pendiente (m^2) y la intersección (b^2). Esto empuja los valores de m y b hacia cero. Especialmente cuando λ es grande.
- * En el modelo RLS con L2, penalizamos la suma de los cuadrados de todos los elementos del vector de pesos w ($w^T w = \sum_{j=1}^Q w_j^2$)
 - Esto tiene el mismo efecto de encoger la magnitud de todos los pesos hacia cero.
- El parámetro de regularización (λ) controla la fuerza de la regularización.
- * Si λ es pequeño, la penalización es débil y el modelo se acerca a la solución de mínimos cuadrados.
- * Si λ es grande, la penalización es fuerte y los pesos se ven forzados a ser más pequeños. Esto ayuda a prevenir el sobreajuste (grande Q) o pocos datos.

la regularización. Llega el modelo RLS generalizado:

En lugar de solo penalizar la pendiente y la intersección penalizan la magnitud de todos los pesos, asociados con los características transformadas por $\Phi(X)$.

Esto tiene el efecto de simplificar el modelo y hacerlo más resistente al sobreajuste al evitar que los pesos tomen valores muy grandes.

La fuerza de esta simplificación está controlada por el parámetro de regularización λ .

Modelo de Máxima Verosimilitud (MV)

Teniendo en cuenta y recordando:

Las estimaciones de MCO (mínimos cuadrados ordinarios) se derivan de ajustar un modelo lineal que representa el promedio de los datos. El modelo se obtiene al minimizar la suma de las desviaciones o errores al cuadrado.

La función de verosimilitud indica que tan probable es reproducir la muestra observada en función de los posibles valores de los parámetros. Es decir, el fin es maximizar la función de verosimilitud con base en los parámetros que tienen más probabilidad de producir los datos observados. Técnicamente, se recomienda el modelo de máxima verosimilitud (MV) para muestras arbitrarias por su adaptabilidad a los diferentes tipos de datos, y producir estimaciones más precisas.

Ventajas de MV sobre MCO son:

- * Mayores estimaciones de los parámetros de distribución
- * Menor varianza de los parámetros (eficiencia)
- * Confiabilidad en la medición de los intervalos de confianza y en las pruebas de hipótesis de los parámetros
- * En los modelos de datos censurados o truncados el MV usa todos los datos individuos los censurados

Por lo tanto para los ESTIMADORES MV:

Sed una población representada por una V.A y función de densidad $f(x, \theta)$ parámetro = θ

θ Puede ser cualquier parámetro de la población
" es el " desconocido de la población

(x_1, x_2, \dots, x_n) es una muestra aleatoria simple de tamaño n

El producto de cada función de densidad para cada elemento de la muestra es la función de MV dependiente del parámetro poblacional según los valores de la muestra

$$L(x_i, \theta) = f(x_1, x_2, \dots, x_n; \theta)$$

$$L(x, \theta) = f(x_1, x_2, \dots, x_n; \theta)$$

$$\hat{L}(x, \hat{\theta}) = \max L(x, \theta)$$

Facilitando cálculos:

$$\ln L(x, \hat{\theta}) = \max \ln L(x, \theta)$$

$$\frac{\partial \ln L(x, \theta)}{\partial \theta} - \sum_{i=1}^n \frac{\partial \ln L(x_i, \theta)}{\partial \theta} = 0$$

$$\left[\frac{\partial^2 \ln L(x, \theta)}{\partial \theta^2} \right] \hat{\theta} < 0$$

La función de verosimilitud para nuestro caso:

$$L(x, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_1 - \mu)^2}{2\sigma^2} \cdots \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x_n - \mu)^2}{2\sigma^2}$$
$$= \frac{1}{(\sigma^2)^{n/2} (2\pi)^{n/2}} \exp - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}$$

El log es una función creciente y por tanto los valores máximos obtenidos, serán los máximos para μ y σ^2

$$\ln L(\bar{x}, \hat{\mu}, \hat{\sigma}^2) = \frac{n}{2} \ln(\hat{\sigma}^2) - \frac{n}{2} \ln(\tau_n^2) - \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\hat{\sigma}^2}$$

C.P.O (condiciones de Primer orden) igualando a cero las derivadas respectivas y $\hat{\sigma}^2$

$$\frac{\partial \ln L(\cdot)}{\partial \mu} = \frac{n}{2\hat{\sigma}^2} \sum_{i=1}^n (x_i - \mu)^2 = 0$$

$$\frac{\partial \ln L(\cdot)}{\partial \sigma^2} = -\frac{n}{2\hat{\sigma}^4} + \frac{\sum_{i=1}^n (x_i - \mu)^2}{2\hat{\sigma}^4} = 0$$

Y la solución del sistema:

$$\hat{\mu} = \bar{x}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n} = S^2$$

C.S.O: Para máximo es que el Hessiano del logaritmo de la función de verosimilitud sea definido negativo

Por lo tanto, para nuestro problema:

$$p(t_n | x_n, w, \sigma_n^2) = N(t_n | \phi(x_n)w^\top, \tau_n^2)$$

$$\rightarrow = \frac{1}{\sqrt{2\pi\tau_n^2}} \exp\left(-\frac{(t_n - \phi(x_n)w^\top)^2}{2\tau_n^2}\right)$$

Debido a la asunción de datos i.i.d, la verosimilitud del conjunto de datos $D = \{x_n, t_n\}_{n=1}^N$

dado w y τ_n^2 es el producto de los probabilidades individuales como se habrá mostrado previamente

Así que teniendo en cuenta las propiedades de log y que además lo mostrado anteriormente

Podemos decir que forma nuestro problema:

$$N(t_n | \phi(x_n) \mathbf{w}^\top, \sigma^2). \text{ se tiene}$$

$$\text{Imp}(\tilde{\mathbf{x}} | \mathbf{x}, \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(\frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \phi(x_n) \mathbf{w})^2)$$

Solución

Para encontrar \mathbf{w}_{ML} que maximiza la log-verosimilitud, tomamos el gradiente con respecto a \mathbf{w} y lo igualamos a cero. Notamos que el primer término no depende de \mathbf{w} , por lo que maximizar la log-verosimilitud con respecto a \mathbf{w} es equivalente a minimizar el segundo término (aparte del factor constante $-\frac{N}{2\sigma^2}$), que es precisamente la función de error de suma de cuadrados utilizada en MINIMOS CUADRADOS.

Por lo tanto, la solución de máxima verosimilitud para \mathbf{w} es idéntica a la solución de mínimos cuadrados, donde

$$E = \|t_n - \phi(x_n) \mathbf{w}\|^2 = \langle (t_n - \phi(x_n) \mathbf{w}), (t_n - \phi(x_n) \mathbf{w}) \rangle$$

donde $\phi(x_n) = \mathbf{x}$ y $t_n = \mathbf{y}$, por lo tanto: recordando las propiedades de los matrices:

$$(\mathbf{A}^\top)^{-1} = (\mathbf{A}^{-1})^\top$$

$$(\mathbf{A}^\top)^\top = \mathbf{A}$$

$$(\mathbf{A} \cdot \mathbf{B})^\top = \mathbf{B}^\top \cdot \mathbf{A}^\top$$

Sean \mathbf{A} y \mathbf{B} 2 matrices $n \times n$ suponga que:

$\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, entonces \mathbf{B} se llama la inversa de \mathbf{A} y se denota \mathbf{A}^{-1}

Entonces

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1} \mathbf{A} = \mathbf{I}$$

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A} \quad (\mathbf{AB})^{-1} = \mathbf{B}^{-1} \mathbf{A}^{-1} \quad (\mathbf{KA})^{-1} = \frac{1}{K} \mathbf{A}^{-1}$$

Entonces:

$$E = \|\mathbf{y} - \mathbf{xw}\|^2 = \langle (\mathbf{x} - \mathbf{xw}), (\mathbf{y} - \mathbf{xw}) \rangle = (\mathbf{y} - \mathbf{xw})^T (\mathbf{y} - \mathbf{xw})$$

Para posibilitar el trabajo a una expresión algebraica más fácil de manipular, se eleva al cuadrado la norma.

La norma al cuadrado de un vector es igual al producto punto del vector consigo mismo, que también se puede expresar como el producto \mathbf{H} del vector transpuesto por el vector original.

Entonces

$$(\mathbf{y} - \mathbf{xw})^T (\mathbf{y} - \mathbf{xw}) = (\mathbf{y}^T - (\mathbf{xw})^T) (\mathbf{y} - \mathbf{xw})$$

por propiedades de la transpuesta.

$$(\mathbf{y}^T - (\mathbf{xw})^T) (\mathbf{y} - \mathbf{xw}) = (\mathbf{y}^T - \mathbf{w}^T \mathbf{x}^T) (\mathbf{y} - \mathbf{xw}) \text{, expandiendo:}$$

$$\mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{xw} - \mathbf{w}^T \mathbf{x}^T \mathbf{y} + \mathbf{w}^T \mathbf{x}^T \mathbf{xw}$$

Recordemos que $\mathbf{y}^T \mathbf{xw}$ es un escalar. La transpuesta de un escalar es el mismo escalar. Por lo tanto:

$$(\mathbf{y}^T \mathbf{xw})^T = \mathbf{w}^T \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{y}$$

sin embargo esto no es lo que tenemos en el tercer término.

Reconociendo la relación entre $\mathbf{y}^T \mathbf{xw}$ y $\mathbf{w}^T \mathbf{x}^T \mathbf{y}$.

Consideremos el producto \mathbf{xw} . El resultado es un vector columna de N elementos (las predicciones para cada observación). Luego multiplicamos este vector por \mathbf{x}^T (un vector fila de N elementos). El resultado de esta multiplicación es un escalar.

Ahora consideremos el producto $\mathbf{x}^T \mathbf{y}$. La matriz \mathbf{x}^T tiene dimensiones $Q \times N$, y el vector \mathbf{y} tiene dimensiones $N \times 1$. El resultado $\mathbf{x}^T \mathbf{y}$ es un vector columna de Q elementos. Luego multiplicamos este vector por \mathbf{w}^T (fila de Q elementos). También es un escalar. De hecho:

$$\mathbf{y}^T \mathbf{xw} = (\mathbf{w}^T \mathbf{x}^T \mathbf{y}^T)^T = \mathbf{w}^T (\mathbf{x}^T \mathbf{y})^T = \mathbf{w}^T \mathbf{y}^T \mathbf{x}$$

Volviendo a la expresión

$$E^2 = \mathbf{y}^T \mathbf{y} - \mathbf{y}^T \mathbf{xw} - \mathbf{w}^T \mathbf{x}^T \mathbf{y} + \mathbf{w}^T \mathbf{x}^T \mathbf{xw} -$$

Sabemos que $(\mathbf{y}^T \mathbf{xw}) = \mathbf{w}^T \mathbf{x}^T \mathbf{y}$, entonces:

$$E^2 = \mathbf{y}^T \mathbf{y} - 2 \mathbf{y}^T \mathbf{xw} + \mathbf{w}^T \mathbf{x}^T \mathbf{xw}$$

Ahora, vamos a derivar esta forma simplificada con respecto a \mathbf{w} :

$$\frac{\partial}{\partial \mathbf{w}} (\mathbf{y}^T \mathbf{y}) = 0$$

$$\frac{\partial}{\partial \mathbf{w}} (-2 \mathbf{y}^T \mathbf{xw}) \quad \text{recordando reglas de derivadas matriciales:}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{a}^T \mathbf{x}) = \mathbf{a}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^T \mathbf{a}) = \mathbf{a}$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^T \mathbf{A} \mathbf{x}) = 2 \mathbf{A} \mathbf{x} \quad (\text{si } \mathbf{A} \text{ es simétrica})$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{B} \mathbf{x}) = \mathbf{B}^T$$

$$\frac{\partial}{\partial \mathbf{x}} (\mathbf{x}^T \mathbf{B}^T) = \mathbf{B}$$

Por lo tanto

$$\frac{\partial}{\partial \mathbf{w}} (-2 \mathbf{y}^T \mathbf{xw}) = -2 \mathbf{x}^T \mathbf{y}$$

$$\frac{\partial}{\partial \mathbf{w}} (\mathbf{w}^T (\mathbf{x}^T \mathbf{x}) \mathbf{w}) = 2 (\mathbf{x}^T \mathbf{x}) \mathbf{w}$$

Por lo tanto igualando a cero:

$$\frac{\partial E^2}{\partial \mathbf{w}} = 0 = 0 - 2 \mathbf{x}^T \mathbf{y} + 2 (\mathbf{x}^T \mathbf{x}) \mathbf{w} = 0$$

$$\mathbf{X}^T \mathbf{X} \mathbf{W} = \mathbf{X}^T \mathbf{Y}$$

Resolviendo para \mathbf{W}

$$\mathbf{X}^T \mathbf{X} \mathbf{W} = \mathbf{X}^T \mathbf{Y}$$

$$\mathbf{W} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Esta es la solución para el vector de pesos \mathbf{W} que minimiza la suma de los errores cuadráticos, que corresponde a la solución de mínimos cuadrados que presentamos anteriormente.

$$\mathbf{W}_{LS} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{E}$$

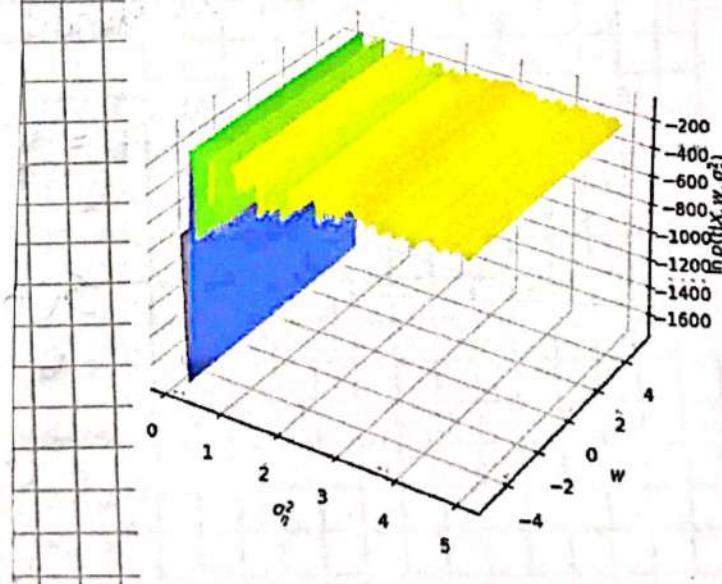
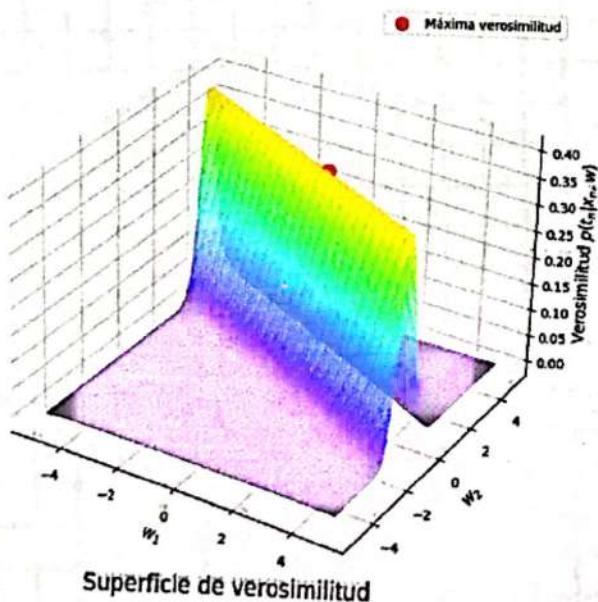
Como la solución de máxima verosimilitud para \mathbf{W} es idéntica a la solución de mínimos cuadrados

$$\mathbf{W}_{MV} = \mathbf{W}_{LS} = (\mathbf{\Phi}^T \mathbf{\Phi})^{-1} \mathbf{\Phi}^T \mathbf{E}$$

Si también queremos estimar $\sigma_n^2 M_L$:

$$\hat{\sigma}_n^2 M_L = \frac{1}{N} \sum_{n=1}^N (\mathbf{e}_n - \phi(\mathbf{x}_n) \mathbf{W}_{ML}^T)^2$$

Esta es la varianza muestral del error utilizando los pesos estimados por max. Verosimilitud



Modelo de Máximo A-Posteriori (MAP)

Recordando que θ : parámetro de interés para y_1 no será fijo sino una V.A que tiene distribución:

$$f(\theta) \rightarrow \text{distribución a priori}$$

con la distribución de los datos:

$$f(x|\theta)$$

y también la distribución a posteriori que la provee el TEOREMA DE BAYES que es:

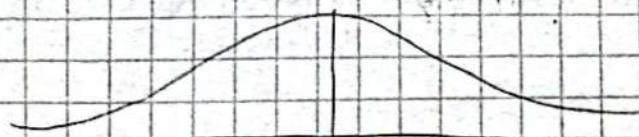
$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{m(x)}$$

donde $m(x) = \int f(x|\theta)f(\theta)d\theta \rightarrow$ Distribución a posteriori

Densidad marginal \rightarrow Distribución (verosimilitud condicional)

En este contexto queremos obtener la distribución a posteriori:

$$f(\theta|x) \quad \text{donde}$$



$$\hat{\theta}_B = E(\theta|x) \Rightarrow \text{Estimador de bayes}$$

$$\text{Mediana: } \hat{\theta} : \int_{-\infty}^{\hat{\theta}} f(\theta|x)d\theta = \frac{1}{2}$$

$$\text{Moda: } \hat{\theta} : f(\hat{\theta}|x) = \sup_{\theta} f(\theta|x)$$

$$f(\theta) = 1 \quad \hat{\theta} = \hat{\theta}_{MV} \quad \text{Supremo}$$

\searrow
Máximo Verosimil

Estructuras para el PROBLEMA de Optimización

En la estimación MAP, además de la verosimilitud de los datos, introducimos un conocimiento previo (prior) sobre los parámetros w . Usando la regla de Bayes, la probabilidad a posteriori de w dado los datos es proporcional a la verosimilitud multiplicada por la probabilidad a priori:

$$p(w | t, \mathbf{x}, \Sigma_n^2) \approx p(t | \mathbf{x}, \mathbf{w}, \Sigma_n^2) p(w)$$

$$W_{MAP} = \arg \max_w \log \left(\prod_{n=1}^N G(t_n | \Phi(x_n) w, \Sigma_n^2) \right) - \frac{1}{2} \sum_{n=1}^N \frac{\|t_n - \Phi w\|^2}{\Sigma_n^2}$$

Asumiendo datos i.i.d:

$$W_{MAP} = \arg \max_w -\frac{1}{2\Sigma_n^2} \|t - \Phi w\|^2 + \frac{1}{2\Sigma_n^2} \|w\|^2$$

Teniendo en cuenta que los factores de escala no modifican el punto máximo/minimo en la optimización, podemos factorizar el problema equivalente MAP como:

$$W_{MAP} = \arg \min_w \|t - \Phi w\|^2 + \frac{\Sigma_n^2}{\Sigma_w^2} \|w\|^2$$

Solución

Bajo estas suposiciones, el problema de optimización de MAP asumiendo ruido y prior Gaussiano, es equivalente a la optimización de mínimos cuadrados regularizados con $\lambda = \frac{\Sigma_n^2}{\Sigma_w^2}$

$$\Phi^T (\mathbf{t} - \Phi \mathbf{w}) - \frac{\Sigma_n^2}{\Sigma_w^2} \mathbf{w} = \mathbf{0}$$

$$\Phi^T \mathbf{t} - \Phi^T \Phi \mathbf{w} - \lambda \mathbf{w} = \mathbf{0}$$

$$(\Phi^T \Phi + \lambda I) \mathbf{w} = \Phi^T \mathbf{t} \quad \text{donde } \lambda = \frac{\Sigma_n^2}{\Sigma_w^2}, \text{ es el}$$

Parámetro de regularización

Norma

La solución para WMAP es:

$$W_{MAP} = (\mathbf{D}^T \mathbf{D} + \lambda \mathbf{I})^{-1} \mathbf{D}^T \mathbf{L}$$

Notamos que la solución MAP con un prior Gaussiano sobre los pesos es identica a la solución de mínimos cuadrados regularizados (Ridge Regression) con un parámetro de regularización λ que es la razón de las varianzas del ruido y del prior de los pesos.

Modelo Bayesiano con modelo lineal Gaussiano

En el enfoque Bayesiano, nuestro objetivo no es encontrar un único valor "optimo" para los parámetros (como el vector de pesos w), sino obtener una distribución de probabilidad completa sobre estos parámetros, dados los datos observados.

Esta distribución se llama la distribución a posteriori, $p(w|t, X, \sigma_n^2)$.

La regla de Bayes para inferencia de parámetros:

$$p(w|t, X, \sigma_n^2) = \frac{P(t|X, w, \sigma_n^2) p(w)}{P(t|X, \sigma_n^2)}$$

donde $P(w|t, X, \sigma_n^2)$ es la distribución a posteriori de los pesos w dado los datos observados (t y X) y la varianza del ruido (σ_n^2). Es la distribución que queremos encontrar.

y $P(t|X, w, \sigma_n^2)$ es la verosimilitud de los datos observados dado los pesos w y la varianza del ruido. Ya la vimos en el modelo de MV. Bajo la asunción de ruido Gaussiano i.i.d; es:

$$P(t|X, w, \sigma_n^2) = \prod_{n=1}^N N(t_n | (X_n)^T w, \sigma_n^2)$$

En forma vectorial:

$$P(t|X, w, \sigma_n^2) = N(t | \mathbf{X}^T w, \sigma_n^2 I)$$

donde Φ es la matriz de diseño e \mathbb{I} es la matriz identidad de $N \times N$.

Para la regla de Bayes, donde

$p(w)$: Es la distribución de prior de los pesos

Representa nuestro conocimiento o creencias sobre los valores de los pesos antes de observar los datos. La elección del prior es crucial en el enfoque Bayesiano.

y donde

$p(t | X, \sigma^2)$: Es la evidencia (o probabilidad marginal de los datos). Actúa como una constante de normalización para asegurar que la distribución a posterior sea una distribución de probabilidad válida (es decir, que su integral sobre todos los posibles valores de w sea igual a 1).

Se calcula integrando la verosimilitud multiplicada por el prior sobre todos los posibles valores de w :

$$p(t | X; \sigma^2) = \int p(t | X, w, \sigma^2) p(w) dw$$

Para que la inferencia Bayesiana seatractable, a menudo se eligen distribuciones conjugadas.

Para nuestro caso el modelo de regresión lineal con ruido Gaussiano, un prior Gaussiano sobre los pesos es conjugado.

*Prior Gaussiano sobre los pesos: Assumimos que nuestros pesos w provienen de un distribución Gaussiana con media m_0 y matriz de covarianza S_0 :

$$p(w) = N(w | m_0, S_0) = \frac{1}{(2\pi)^{N/2} |S_0|^{1/2}} \exp\left(-\frac{1}{2}(w - m_0)^T S_0^{-1} (w - m_0)\right)$$

Para simplificar o expresar una falta de preferencia inicial se elige una traza a priori de ruido ($m_0 = 0$) y una matriz de covarianza $S_0 = T_n^{-1}$ proporcional a la identidad ($S_0 = T_n^{-1}$), donde T_n^{-1} controla la dispersión de los pesos alrededor de cero.

Como ya mencionamos, la verosimilitud de los datos bajo el modelo lineal con ruido Gaussiano es:

$$P(t|X, w, T_n) = N(t | \tilde{\Omega}_n^{-1} w, T_n^{-1})$$

Cuando la verosimilitud es Gaussiana y el prior sobre los pesos también es Gaussiana, la distribución a posteriori $P(w|t, X, T_n)$ también será una Gaussiana.

Nuestra tarea es encontrar su media (m_N) y su matriz de covarianza (S_N).

Después de realizar la multiplicación (prior por verosimilitud) y completar el cuadro en el exponente (para obtener la forma de una Gaussiana), se llega a la siguiente forma:

$P(w|t, X, T_n) = N(w|m_N, S_N)$ donde la matriz de covarianza a posteriori S_N la media a posteriori m_N están dadas por:

$$S_N^{-1} = S_0^{-1} + \frac{1}{T_n} \tilde{\Omega}^T \tilde{\Omega} [1]$$

$$m_N = S_N (S_0^{-1} m_0 + \frac{1}{T_n} \tilde{\Omega}^T t) [2]$$

Desglosando la solución (Distribución a posteriori)

[1]: Nos da la inversa de la matriz de covarianza a posteriori.

Cuanto más datos (mayor N) y menor sea el ruido (menor T_n) mayor será la precisión a posteriori (menor será la incertidumbre).

(Continúa ...)

[2]: Esta ecuación nos da la media a posteriori.

Es una combinación de la media a priori (m_0) y la información proporcionada por los datos ($\Phi^T t$), ponderada por sus respectivas precisiones (inversas de las covarianzas).

Asumiendo un prior con medio cero ($m_0 = 0$) y covarianza $S_0 = \sigma_n^2 I$, las ecuaciones se simplifican:

$$[1] = \frac{1}{\sigma_n^2} \mathbb{I} + \frac{1}{\sigma_w^2} \Phi^T \Phi \quad [3]$$

$$[2] = \left(\frac{1}{\sigma_n^2} \mathbb{I} + \frac{1}{\sigma_w^2} \Phi^T \Phi \right)^{-1} \frac{1}{\sigma_n^2} \Phi^T t \quad [4]$$

Multiplicando por el numerador y el denominador de la última expresión por $\sigma_n^2 \sigma_w^2$:

$$m_N = \left(\frac{\sigma_n^2}{\sigma_w^2} \mathbb{I} + \Phi^T \Phi \right)^{-1} \sigma_w^2 \Phi^T t$$

Dividiendo entre σ_w^2 :

$$m_N = \left(\frac{\sigma_n^2}{\sigma_w^2} \mathbb{I} + \Phi^T \Phi \right)^{-1} \Phi^T t$$

Recordando que $\lambda = \frac{\sigma_n^2}{\sigma_w^2}$ en la regresión de Ridge (MAP con prior Gaussiana)

Vemos que la media a posteriori m_N coincide con la solución de la regresión del Ridge WTLS o WMAP.

Entonces, El modelo Bayesiano con modelo lineal Gaussiano, no solo nos da una estimación puntual de los pesos, sino toda una distribución de probabilidad sobre ellos.

Esta distribución a posteriori, bajo nuestras condiciones) codifica nuestra incertidumbre sobre los valores de los pesos después de observar los datos. m_N a menudo coincide con las líneas obtenidas por otros métodos. Y S_N nos informa sobre la incertidumbre alrededor de esta media m_N

Norma

Regresión Ridge Kernel - (KRR)

PROBLEMA

La regresión (KRR) es una extensión de la regresión de Ridge que opera en un espacio de características implícito definido por una función Kernel

$$k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

El objetivo es minimizar la siguiente función de error con regularización L2 en el espacio de características implícito

$$E_{KRR}(w) = \frac{1}{2} \sum_{n=1}^N (t_n - \phi(x_n)^T w)^2 + \frac{\lambda}{2} w^T w$$

Sin embargo, en KRR, la solución para w se expresa en términos de una combinación lineal de las funciones de base evaluadas en los puntos de entrenamiento:

$$w = \sum_{i=1}^N \alpha_i \phi(x_i)$$

Sustituyendo esto en la función de predicción, obtenemos:

$$y(x) = \phi(x)^T w = \sum_{i=1}^N \alpha_i \phi(x)^T \phi(x_i) = \sum_{i=1}^N \alpha_i k(x, x_i)$$

El problema de optimización se reformula en términos de los coeficientes $\alpha = (\alpha_1, \dots, \alpha_N)^T$:

$$E_{KRR}(\alpha) = \frac{1}{2} \sum_{n=1}^N \left(t_n - \sum_{i=1}^N \alpha_i k(x_n, x_i) \right)^2 + \frac{\lambda}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j k(x_i, x_j)$$

En forma matricial con K siendo la matriz Kernel de $N \times N$ donde $K_{ij} = k(x_i, x_j)$ y $t = (t_1, \dots, t_N)^T$, El problema se escribe como:

$$E_{KRR}(\alpha) = \frac{1}{2} (t - K\alpha)^T (t - K\alpha) + \frac{\lambda}{2} \alpha^T K \alpha$$

Solución

Para encontrar α que minimiza $E_{KRR}(\alpha)$, tomamos el gradiente con respecto a α y lo igualamos a cero:

$$\nabla_{\alpha} E_{KRR}(\alpha) = -K^T(t - K\alpha) + \lambda K\alpha = 0$$

Como la matriz kernel K es simétrica ($K(x_i, x_j) = K^T$), por lo tanto:

$$-Kt + K^2\alpha + \lambda K\alpha = 0$$

$$(K^2 + \lambda K)\alpha = Kt$$

$$K(K + \lambda I)\alpha = Kt$$

Si K es invertible, podemos simplificar

$$(K + \lambda I)\alpha = t$$

por lo tanto la solución para α_{KRR} :

$$\alpha_{KRR} = (K + \lambda I)^{-1} t$$

La predicción para un nuevo punto x_* es entonces:

$$y(x_*) = \sum_{i=1}^N \alpha_i K(x_*, x_i) = K(x_*)^T (K + \lambda I)^{-1} t$$

donde $K(x_*) = (K(x_*, x_1), \dots, K(x_*, x_N))^T$

Regresión mediante Procesos Gaussiantos (GPR)

PROBLEMA: (infomedia)

En vez de aprender un vector de pesos w , modelamos directamente la distribución de probabilidad sobre las funciones $f(x)$.

Un proceso Gaussiano es completamente especificado por su función de media $m(x) = E[f(x)]$ y su función de covarianza (Kernel).

$$K(x, x') = E[(f(x) - m(x))(f(x') - m(x'))]$$

Asumiendo una media cero para simplificar, $f(x) \sim GP(0, K(x, x'))$.

Los valores objetivo observados t están relacionados con la función objetivo $f(x)$ (evaluada en los puntos de entrenamiento) a través del modelo de ruido aditivo:

$$t = f(x) + \eta, \text{ donde } \eta \sim N(0, \sigma^2 \mathbb{I})$$

La distribución conjunta de las observaciones t y las predicciones en un conjunto de nuevos puntos x_* (denotadas como $f_* = f(x_*)$) es Gaussiana:

$$\begin{pmatrix} t \\ f_* \end{pmatrix} \sim N\left(0, \begin{pmatrix} K(x, x) + \sigma^2 \mathbb{I} & K(x, x_*) \\ K(x_*, x) & K(x_*, x_*) \end{pmatrix}\right)$$

donde $K(A, B)$ es la matriz de covarianza evaluada entre los puntos en A y B utilizando la función Kernel K .

El objetivo es inferir la distribución predictiva a posteriori $p(f_* | t, X, X_*)$.

Solución (Distribución predictiva):

Utilizando las propiedades de las distribuciones Gaussiantos condicionales, la distribución a posteriori de f_* dado t es también Gaussiana:

$$\mu(f_{\pi}(x) + f_{\pi}(x_*, \Sigma_*)) = f_{\pi}(x, \mu_x, \Sigma_x)$$

Con media predicción:

$$\mu_n = K(x_n, X) K(X, X)^T K_n^{-1} t$$

y matriz de covarianza predictiva:

$$\Sigma_n = K(x_n, x_n) - K(x_n, X) [K(X, X) + K_n^{-1}]^{-1} K(X, x_n)$$

La predicción para un único nuevo punto x_* es entonces una distribución Gaussiana con media

$$\mu(x_*) = K(x_*, X) K(X, X) + K_n^{-1} t \quad \text{y Varianza:}$$

$$\Sigma^2(x_*) = K(x_*, x_*) - K(x_*, X) [K(X, X) + K_n^{-1}]^{-1} K(X, x_*)$$

Similitudes

- * Todos los modelos buscan predecir un valor objetivo t dado un conjunto de características X .

- A Dependencia de los datos: Todos los modelos aprenden de los datos de entrenamiento:

$$\{x_n, t_n\}_{n=1}^N$$

- B Uso de funciones de Base (Implicito o explicito):

+ Mínimos cuadrados (LS) y mínimos regularizados (RLS) utilizan explícitamente funciones de base $\phi(x)$.

+ KRR utiliza implicitamente funciones de base a través de la función Kernel

$$K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$$

+ GPR también utilizan una función Kernel que define la covarianza entre los valores de la función en diferentes puntos, lo que esto

Norma

relacionados con la similitud en un espacio de características implícito.

* Regularización:

- + RLS introduce explícitamente un término de regularización L_2 sobre los pesos w .
- + KRR también utiliza regularización L_2 sobre los pesos implícitos w (lo que se traduce en regularización sobre los coeficientes α).
- + En GPR, la elección de la función Kernel y sus hiperparámetros juega un papel similar al de la regularización, controlando la complejidad de la función aprendida.

* Modelo lineal en el ESPACIO DE CARACTERÍSTICAS:

LS, RLS, KRR aprenden una relación lineal en el espacio de características definido por $\phi(x)$ (explícita o implícitamente).

Diferencias

* Naturaleza de la salida:

- + LS, RLS, y KRR proporcionan una predicción puntual para un nuevo punto x^* .
- + GPR proporciona una distribución de probabilidad sobre la predicción para un nuevo punto x^* , lo que incluye una medida de incertidumbre (Varianza predictiva).

* Paramétrico vs No Paramétrico:

- + LS y RLS son modelos paramétricos: Una vez que se aprenden los pesos w , no se necesita almacenar los datos de entrenamiento para hacer predicciones (solo la función de base y los pesos). El # de parámetros es fijo (la dimensión de w).

+ KRR y EPR son todos no paramétricos

• Predicir para un nuevo punto depende directamente de los datos del entorno y a través de la función Kernel ($K(x)$) o la "infancia" de la distribución a posteriori (en EPR).

El # de "parámetros" crece con el tamaño del conjunto de datos.

* Enfoque:

+ LS, TLS, KRR se centran en encontrar los "mejores" pesos w que minimizan una función del error

+ GPR adopta un enfoque probabilístico, modelando directamente la distribución de funciones y realizando Inferencia Bayesiana para obtener la distribución predictiva.

* Manejo de la Incertidumbre:

GPR Proporciona una forma natural de cuantificar la incertidumbre en las predicciones, lo cual no es inherente a LS, TLS, KRR

(Aunque se podrían estimar intervalos de confianza a posteriori bajo ciertas asunciones).

* Complejidad Computacional:

- La complejidad de LS y TLS depende principalmente de la inversión de la matriz $\Phi^T \Phi + \lambda I$ (o

$\Phi^T \Phi + \lambda I^{-1}$), que es $\mathcal{O}(Q^3)$ si $N \geq Q$

- La complejidad de KRR depende de la inversión de la matriz Kernel $K + \lambda I$, que es $\mathcal{O}(N^3)$.

- De EPR también involucra la inversión de una matriz de tamaño $N \times N$ ($K(\mathbf{x}, \mathbf{x}) + \sigma_n^2 I$)

Lo que tambien es $O(N^3)$. Esto puede ser una limitacion para conjuntos de datos grandes.

Entonces, mientras que LS, RLS y KRR buscan un mapeo determinista de las entradas a las salidas minimizando un error, GPR proporciona una perspectiva probabilistica completa, incluyendo la incertidumbre en las predicciones!

KRR extiende la regularizacion de Ridge al espacio de caracteristicas implicito definido por un kernel, permitiendo modelos no lineales con una complejidad que depende del # de datos.

GPR tambien maneja modelos no lineales a traves del Kernel, pero su salida es una distribucion de probabilidad.

La eleccion depende de los requisitos de la tarea como la necesidad de cuantificar la incertidumbre, la interpretabilidad y la escalabilidad a grandes conjuntos de datos.