viernes, mayo 16, 2025 8:29 PM

1) Sea el modelo de Regresión $t_n = \phi(x_n)w^T + \eta_n$, con $\{t_n \in \mathbb{R}, x_n \in \mathbb{R}^P\}_{n=1}^N$, $w \in \mathbb{R}^Q$, $\phi: \mathbb{R}^P \to \mathbb{R}^Q$, $Q \geq P$, y $\eta_n \sim \mathcal{N}(\eta_n|0,\sigma_n^2)$. Presente el problema de optimización y la solución del mismo para los modelos mínimos cuadrados, mínimos cuadrados regularizados, máxima verosimilitud, máximo a-posteriori, Bayesiano con modelo lineal Gaussiano, regresión rígida Kernel y mediante procesos Gaussianos. Asuma datos i.i.d. Discuta las diferencias y similitudes entre otros modelos estudiados.

## ⇒ Definimos el modelo general. (Least Squares)

$$t_n = \phi(x_n)^T w + \eta_n, \quad \eta_n \sim \mathcal{N}(0,\sigma^2)$$

Donde
- $\phi(x_n) \in \mathbb{R}^Q$: función de transformación de características
- $w \in \mathbb{R}^Q$: Vector de Parámetros
- $t_n$ es el valor objetivo escalar: Variable de salida (escalares).
- Hay N muestras, todas i.i.d.
- $x_n \in \mathbb{R}^P$: Vector de Entrada
- $\Phi \in \mathbb{R}^{N\times Q}$: Matriz de diseño, donde cada fila es $\phi(x_n)^T$
- $t \in \mathbb{R}^N$: Vector de salidas
- Queremos encontrar $w$ que minimice el error cuadrático entre la predicción $\phi(x_n)^T w$ y el valor real $t_n$.

### 2. Función de Pérdida
- Función objetivo (error cuadrático medio):

$$J(w) = \sum_{n=1}^N (t_n - \phi(x_n)^T w)^2 = \|t - \Phi w\|^2$$

Expandimos el cuadrado.

$$J(w) = (t-\Phi w)^T(t-\Phi w)$$
$$J(w) = t^T t - 2w^T\Phi^T t + w^T\Phi^T\Phi w$$

### 3. Derivada con respecto a $w$.
- Buscamos el mínimo de $J(w)$, derivamos respecto a $w$ y lo igualamos a cero:

$$\nabla_w J(w) = -2\Phi^T t + 2\Phi^T\Phi w$$
$$0 = -2\Phi^T t + 2\Phi^T\Phi w$$

### 4. Solución cerrada.
Despejamos

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

- Si $\Phi^T\Phi$ es invertible

$$w^* = (\Phi^T\Phi)^{-1}\Phi^T t$$

Esta es la solución óptima del modelo de mínimos cuadrados.

### 5. Interpretación
- Esta solución se llama también la solución de la Pseudoinversa:

$$w^* = \Phi^+ t$$

- Supone que no hay regularización (riesgo de overfitting si hay colinealidad o pocos datos).

## ⇒ Modelo de Mínimos Cuadrados Regularizados (Ridge Regresion).

Planteamiento general

Este modelo extiende la regresión por mínimos cuadrados con un término de regularización $L_2$ que penaliza la magnitud de los coeficientes $w$. El modelo ayuda a evitar sobreajuste (overfitting) y a mejorar la estabilidad numérica

### 2. Función de pérdida
La nueva función de costo es.

$$J(w) = \sum_{n=1}^N (t_n - \phi(x_n)^T w)^2 + \lambda\|w\|^2 = \|t - \Phi w\|^2 + \lambda w^T w$$

Donde
- $\lambda > 0$ es el hiperparámetro de regularización
- $\|w\|^2 = w^T w$

### 3. Expansión algebraica
Expandimos la función

$$J(w) = (t-\Phi w)^T(t-\Phi w) + \lambda w^T w$$
$$J(w) = t^T t - 2w^T\Phi^T t + w^T\Phi^T\Phi w + \lambda w^T w$$

### 4. Derivada con respecto a $w$
Derivamos la función de costo con respecto a $w$

$$\nabla_w J(w) = -2\Phi^T t + 2\Phi^T\Phi w + 2\lambda w$$

Igualamos a cero.

$$-2\Phi^T t + 2\Phi^T\Phi w + 2\lambda w = 0$$

Dividimos por 2.

$$-\Phi^T t + \Phi^T\Phi w + \lambda w = 0$$

### 5. Solución Cerrada.
Agrupamos términos

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

si $(\Phi^T\Phi + \lambda I)$ es invertible

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

### 6. Interpretación.
- Si $\lambda = 0$, recuperamos la solución de mínimos cuadrados ordinarios
- para $\lambda > 0$, los coeficientes se "encogen" hacia 0 (efecto de Regularización)
- Reduce la varianza del modelo, a costa de introducir un poco de sesgo (bias-variance tradeoff)

## ⇒ Modelo de Máxima Verosimilitud (MLE).

### 1 Planteamiento probabilístico
Partimos del mismo modelo de regresión lineal:

$$t_n = \phi(x_n)^T w + \eta_n, \quad \eta_n \sim \mathcal{N}(0,\sigma^2)$$

Entonces

$$t_n \sim \mathcal{N}(\phi(x_n)^T w, \sigma^2)$$

Es decir el valor observado $t_n$ se modela como una variable aleatoria Gaussiana, con media $\phi(x_n)^T w$ y varianza $\sigma^2$

### 2. Verosimilitud conjunta de los datos.
Dado que los datos son i.i.d (independientes e idénticamente distribuidos), la verosimilitud conjunta es:

$$p(t|\Phi, w, \sigma^2) = \prod_{n=1}^N \mathcal{N}(t_n|\phi(x_n)^T w, \sigma^2)$$

En forma Vectorial

$$p(t|w,\sigma^2) = \mathcal{N}(t|\Phi w, \sigma^2 I)$$

### 3 Log-Verosimilitud
Usamos la Log-Verosimilitud para facilitar el cálculo.

$$\log p(t|w,\sigma^2) = \log\left[\frac{1}{(2\pi\sigma^2)}\exp\left(-\frac{1}{2\sigma^2}\|t-\Phi w\|^2\right)\right]$$
$$= -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|t-\Phi w\|^2$$

### 4. Estimación de $w$ por MLE
Como queremos encontrar el $w$ que maximice la verosimilitud, también podemos minimizar el término cuadrático negativo (el log-likelihood negativo):

$$\hat{w}_{MLE} = \arg\min_w \|t - \Phi w\|^2.$$

### 5. Solución.
Ya sabemos que:

$$\hat{w}_{MLE} = (\Phi^T\Phi)^{-1}\Phi^T t$$

### 6. Estimación de $\sigma^2$
Además, podemos encontrar el MLE de $\sigma^2$ (La varianza del ruido).

$$\hat{\sigma}^2 = \frac{1}{N}\|t - \Phi\hat{w}\|^2$$

## ⇒ Modelo de Máximo A posteriori (MAP).

### 1. Planteamiento general.
Sabemos del modelo.

$$t_n = \phi(x_n)^T w + \eta_n, \quad \eta_n \sim \mathcal{N}(0,\sigma^2)$$

por tanto:

$$p(t|w,\sigma^2) = \mathcal{N}(t|\Phi w, \sigma^2 I)$$

Ahora, introducimos una distribución a priori sobre $w$:

$$p(w) = \mathcal{N}(w|0, \alpha^{-1}I)$$

Esto significa que a priori creemos que $w$ está centrado en 0 y tiene varianza $\alpha^{-1}$ (hiperparámetro de control de regularización)

### 2 objetivo MAP
El estimador MAP maximiza la distribución posterior $p(w|t)$, usando el teorema de Bayes:

$$p(w|t) \propto p(t|w)p(w).$$

Entonces.

$$\hat{w}_{MAP} = \arg\max_w[\log p(t|w) + \log p(w)]$$

### 3. Desarrollamos los términos
Verosimilitud (como en MLE):

$$\log p(t|w) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\|t - \Phi w\|^2$$

Priori:

$$\log p(w) = -\frac{Q}{2}\log(2\pi\alpha^{-1}) - \frac{\alpha}{2}\|w\|^2$$

### 4. Función a minimizar (negativa de log posterior)
Ignorando constantes (no afectan el mínimo), la función a minimizar es:

$$J(w) = \frac{1}{2\sigma^2}\|t - \Phi w\|^2 + \frac{\alpha}{2}\|w\|^2$$

Podemos reescribirlo como

$$J(w) = \|t - \Phi w\|^2 + \lambda\|w\|^2, \text{ donde } \lambda = \sigma^2\alpha$$

¡Esto es exactamente el mismo costo que en Ridge Regresion!

### 5. Derivación y solución
Como ya derivamos esta función al resolver Ridge, usamos directamente:

$$\hat{w}_{MAP} = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T t.$$

Con:

$$\lambda = \sigma^2\alpha.$$

## * Modelo Bayesiano con Modelo Lineal Gaussiano

### 1. Planteamiento del modelo
Tenemos.

$$t_n = \phi(x_n)^T w + \eta_n, \quad \eta_n \sim \mathcal{N}(0,\sigma^2)$$

Vectorialmente.

$$t \sim \mathcal{N}(\Phi w, \sigma^2 I)$$

Y asumimos un prior Gaussiano sobre $w$:

$$p(w) = \mathcal{N}(w|0, \alpha^{-1}I)$$

### 2. Objetivo
Nuestro objetivo es calcular el posterior completo

$$p(w|t) = \frac{p(t|w)p(w)}{p(t)}$$

Donde
- $p(t|w)$ es la verosimilitud
- $p(w)$ es el prior
- $p(t)$ es la evidencia (constante de normalización).

Como ambos términos son distribuciones normales, el resultado es también una distribución normal.

$$p(w|t) = \mathcal{N}(w|\mu_N, \Sigma_N)$$

Vamos a derivar $\mu_N$ y $\Sigma_N$ a continuación

### 3. Derivación paso a paso del posterior $p(w|t)$
- Verosimilitud

$$p(t|w) \propto \exp\left(-\frac{1}{2\sigma^2}\|t - \Phi w\|^2\right)$$

- Prior:

$$p(w) \propto \exp\left(-\frac{\alpha}{2}w^T w\right)$$

- Producto (sin constantes):

$$p(w|t) \propto \exp\left(-\frac{1}{2\sigma^2}\|t - \Phi w\|^2 - \frac{\alpha}{2}w^T w\right.$$

- Expandimos el cuadrado

$$\|t - \Phi w\|^2 = t^T t - 2w^T\Phi^T t + w^T\Phi^T\Phi w$$

- Sustituimos en la exponencial

$$p(w|t) \propto \exp\left(-\frac{1}{2}\left[w^T\left(\frac{1}{\sigma^2}\Phi^T\Phi + \alpha I\right)w - \frac{2}{\sigma^2}w^T\Phi^T t\right]\right)$$

Completamos cuadrados para expresar esto como una Gaussiana en $w$:

La forma general de una gaussiana es

$$\exp\left(-\frac{1}{2}(w - \mu_N)^T\Sigma_N^{-1}(w - \mu_N)\right)$$

Comparando, obtenemos.
- Matriz de covarianza

$$\Sigma_N = \left(\alpha I + \frac{1}{\sigma^2}\Phi^T\Phi\right)^{-1}$$

- Media posterior

$$\mu_N = \frac{1}{\sigma^2}\Sigma_N\Phi^T t$$

### 4. Predicción bayesiana
Queremos predecir el valor $t_*$ para una nueva entrada $x_*$, con $\phi_* = \phi(x_*)$

Como el modelo es probabilístico, el valor de la salida también tendrá distribución Gaussiana.

$$p(t_*|x_*, t) = \mathcal{N}(t_*|\mu(x_*), \sigma^2(x_*))$$

Donde:
- Media predictiva

$$\mu(x_*) = \phi(x_*)^T\mu_N$$

- Varianza predictiva

$$\sigma^2(x_*) = \sigma^2 + \phi(x_*)^T\Sigma_N\phi(x_*)$$

Esto nos da predicciones con incertidumbre, lo cual es clave en contextos donde la confianza del modelo importa.

## Regresión Rígida con Kernel (Kernel Ridge Regression-KRR)

### 1 Motivación
En Ridge clásico usamos la matriz de diseño $\Phi$, con transformaciones explícitas $\phi(x_n)$ pero a veces:
- $\phi(x_n)$ es muy costoso o imposible de computar
- Queremos trabajar en espacios de características de alta o infinita dimensión.

Usamos entonces un truco de Kernel, que evita computar $\phi(x)$ directamente

### 2. Recordatorio: Ridge tradicional
La solución de Ridge era:

$$w^* = (\Phi^T\Phi + \lambda I)^{-1}\Phi^T t.$$

La predicción para un nuevo punto $x_*$:

$$\hat{t}_* = \phi(x_*)^T w^*$$

Reemplazamos $w^*$

$$\hat{t}_* = \phi(x_*)^T(\Phi^T\Phi + \lambda I)^{-1}\Phi^T t.$$

### 3. Forma dual.
Definimos

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

Usamos el echo que:

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

Entonces:

$$\hat{t}_* = \phi(x_*)^T\Phi^T\alpha = K(x_*)^T\alpha.$$

### 4. Definimos el Kernel
Definimos el Kernel como

$$K_{ij} = K(x_i, x_j) = \phi(x_i)^T\phi(x_j).$$
$$K(x_*) = [K(x_*, x_1), K(x_*, x_2), \dots, K(x_*, x_N)]^T$$

### 5. Solución final en forma dual.

$$\alpha = (K + \lambda I)^{-1}t$$
$$\hat{t}_* = K(x_*)^T\alpha$$

### 6. Interpretación
Toda la solución depende de productos internos. $\phi(x_i)^T\phi(x_j)$, que se calculan con un Kernel.
- No se necesita conocer $\phi(x)$ explícitamente
- Se puede usar Kernels como:
  - Lineal: $K(x, x') = x^T x'$
  - Polinomial: $K(x, x') = (x^T x' + c)^d$
  - RBF (Gaussiano): $K(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$

La matriz $K \in \mathbb{R}^{N\times N}$, así que el costo computacional crece con el número de muestras, es ideal para conjunto de datos pequeños a medianos

## Regresión con procesos Gaussianos (GPR).

### 1. ¿Qué es un proceso Gaussiano?
Un proceso Gaussiano (GP) es una distribución sobre funciones $f(x)$, tal que para cualquier conjunto finito de puntos $\{x_1, \dots, x_n\}$, los valores $\{f(x_1), \dots, f(x_N)\}$ tienen una distribución conjunta normal.

$$f(x) \sim \mathcal{GP}(m(x), K(x, x'))$$

- $m(x)$: función media (usualmente 0)
- $K(x, x')$: función de covarianza o Kernel

### 2. Modelo de observación.
Asumimos que las observaciones están afectadas por ruido Gaussiano

$$t_n = f(x_n) + \eta_n, \quad \eta_n \sim \mathcal{N}(0, \sigma^2)$$

### 3. Formulación del Modelo.
Sean.
- $X = [x_1, \dots, x_N]^T$: entradas de entrenamiento
- $t = [t_1, \dots, t_N]^T$: Salidas observadas.
- $x_*$: Nueva entrada.
- $f_* = f(x_*)$: Valor funcional a predecir

Como $f(x)$ en un GP, el vector extendido

$$\begin{bmatrix} t \\ f_* \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) + \sigma^2 I & k(X, x_*) \\ k(x_*, X) & k(x_*, x_*) \end{bmatrix}\right)$$

donde
- $K(X, X) \in \mathbb{R}^{N\times N}$: Matriz de kernel entre dos puntos de entrenamiento.
- $k(X, x_*) \in \mathbb{R}^{N\times 1}$: Vector de covarianzas entre $x_*$ y puntos de entrenamiento

### 4. Predicción condicional
Usamos la regla de la condición de una gaussiana multivariada.

Entonces

$$p(f_*|x_*, X, t) = \mathcal{N}(\mu(x_*), \sigma^2(x_*)).$$

Con.

Media predictiva

$$\mu(x_*) = k(x_*, X)[K(X,X) + \sigma^2 I]^{-1}t.$$

Varianza Predictiva

$$\sigma^2(x_*) = k(x_*, x_*) - k(x_*, X)[K(X,X) + \sigma^2 I]^{-1}k(X, x_*)$$

### 5 Hiperparámetros.
- Se seleccionan típicamente mediante máxima verosimilitud marginal o métodos bayesianos.
- Por ejemplo, la varianza de kernel, la longitud de escala en el RBF Kernel, etc.

### 6 Elección del Kernel.
- Lineal: $k(x, x') = x^T x'$
- Polinomial: $k(x, x') = (x^T x' + c)^d$
- RBF (gaussiano): $k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2l^2}\right)$
- Matern, Rational Quadratic, etc.

### Características clave

| Características | GPR. |
|---|---|
| Tipo de salida | Distribución Gaussiana para cada predicción |
| Incertidumbre natural | ✓ Sí (Varianza predictiva) |
| No requiere estimar $w$ | ✓ |
| Regularización implícita | ✓ (a través de Kernel y $\sigma^2$) |
| Coste computacional | $O(N^3)$ por inversión de matriz |
| Ideal para. | Pocos datos, problemas donde la incertidumbre importa. |

## Comparación general final

| Modelo | Tipo | Prior | Regularización | Predicción probabilística | Coste |
|---|---|---|---|---|---|
| LS | Determinista | ✗ | ✗ | ✗ | Bajo |
| Ridge | Determinista | ✗ | ✓ (L2) | ✗ | Bajo |
| MLE | Probabilístico | ✗ | ✗ | ✗ | Bajo |
| MAP | Probabilístico | ✓ (sobre ww) | ✓ | ✗ | Bajo |
| Bayes lineal | Probabilístico | ✓ (sobre ww) | ✓ | ✓ (sobre ww y predicción) | Medio |
| Kernel Ridge | Determinista | ✗ | ✓ | ✗ | Medio |
| GP | Probabilístico | ✓ (sobre funciones) | ✓ | ✓ | Alto |

## Comparaciones clave entre modelos.

| Comparación | ¿Cuál Preferir? | Justificación. |
|---|---|---|
| LS vs Ridge | Ridge | Menor Varianza, más Estable |
| MLE vs MAP | MAP | MAP incluye regularización implícita |
| MAP vs Bayesiano Completo. | Bayesiano | MAP de puntos; Bayesiano de distribución Completa |
| Ridge vs Kernel Ridge | Kernel Ridge si hay no linealidad | Puede capturar relaciones más complejas |
| Kernel Ridge vs GP | GP si importa la incertidumbre | GP da predicción + Varianza; KR sólo punto. |
| Bayesiano vs GP | GP si no se quiere asumir $\phi(x)$. | GP es no paramétrico; modelo bayesiano lineal es paramétrico |

## Recomendación por situación:
- Muchos datos, relación Lineal → Ridge
- Pocos datos, se necesita incertidumbre → GP
- Modelo sencillo y explicable → LS o MLE
- Regularización + Conocimiento previo → MAP.
- Alta flexibilidad, no linealidad → Kernel Ridge o GP
- Predicción con probabilidad y covarianza → Modelo Bayesiano completo o GP.