Parcial 1 de Teoría de Aprendizaje de Máquina 2025-1

(1). El modelo base es $t_n = \phi(x_n)^T w + \eta_n, \quad \eta_n \sim N(0, \sigma^2)$

El conjunto de datos: $\{ t_n \in \mathbb{R}, x_n \in \mathbb{R}^p \}_{n=1}^{N}$

$\phi: \mathbb{R}^p \to \mathbb{R}^q, \quad q \geq p$
Asumimos que los datos son i.i.d.
El vector $w \in \mathbb{R}^q$ son los pesos del modelo.

1.1 Mínimos cuadrados:

Queremos encontrar un $w \in \mathbb{R}^q$ tal que minimice el error cuadrático entre los valores reales t_n y los valores predichos por el modelo $\phi(x_n)^T w$.

La Función de costo (error cuadrático medio) es:

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

Llevado a forma matricial queda:

- $t \in \mathbb{R}^N$: Vector con las salidas reales $[t_1, t_2, \ldots, t_N]^T$
- $\Phi \in \mathbb{R}^{N \times q}$: matriz de diseño, donde cada fila es $\phi(x_n)^T$
- $w \in \mathbb{R}^q$: Vector de pesos.

Entonces:

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

Derivamos con respecto a w para minimizar; Aplicando la identidad matricial

$$\nabla w [(a - Aw)^T (a - Aw)] = -2A^T (a - Aw)$$

$\nabla w J(w) = -2\Phi^T (t - \Phi w)$. Haciendo $A = \Phi$ y $a = t$

Multiplicando be igualando a cero obtenemos:

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

Dividiendo entre 2 y despejando w se tiene:

$2\Phi^T \Phi w = 2\Phi^T t \rightarrow w = \frac{\Phi^T t}{\Phi^T \Phi}$, Aplicando inversa si $\Phi^T \Phi$ es invertible se obtiene

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

Donde Φ es la matriz de diseño, donde cada fila es un vector transformado de entrada:

$$\Phi = \begin{bmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_N)^T \end{bmatrix} \in \mathbb{R}^{N \times a}$$

12 Mínimos cuadrados Regularizados: (Ridge Regression)

En el modelo de mínimos cuadrados ordinarios la solución puede ser inestable si: las columnas de Φ están correlacionadas. si: $N < a$ (más parámetros que datos). si los datos tienen mucho ruido.

Entonces buscamos penalizar el tamaño de los coeficientes para evitar sobreajuste o inestabilidad. La función de costo para este modelo es:

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

donde $\lambda > 0$ es el parámetro de regularización (controla cuanto penalizar)

$\|w\|^2 = w^T w$ es la norma cuadrada de los pesos.

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

Derivamos con respecto a w utilizando las siguientes identidades matriciales.

$$\nabla_w [(t - \Phi w)^T (t - \Phi w)] = -2\Phi^T (t - \Phi w)$$

$$\nabla_w [w^T w] = 2w \quad . \quad \text{Entonces:}$$

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

Igualamos a cero $\rightarrow -2\Phi^T (t - \Phi w) + 2\lambda w = 0$

Dividimos entre 2 $\rightarrow -\Phi (t - \Phi w) + 2\lambda w = 0$

Distribuimos el producto $\rightarrow -\Phi t + \Phi^T \Phi w + \lambda w = 0$

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

Donde I es la matriz identidad de tamaño $Q \times Q$, Despejando w

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

1.3 Regresión por Máxima Verosimilitud (Maximum likelihood)

En el modelo base del problema, cada salida t_n se genera como una combinación lineal de las caracteristicas $\phi(x_n)$, con ruido gaussiano aditivo.

El modelo de probabilidad condicional es:

$$p(t_n \mid x_n, w) = N(t_n \mid \phi(x_n)^T w, \sigma^2)$$

Dado que los datos son i.i.d. la verosimilitud total es:

$$p(t_n \mid X_n, w) = \prod_{n=1}^{N} N(t_n \mid \phi(x_n)^T w, \sigma^2)$$

Para facilitar derivadas, usamos el logaritmo de la verosimilitud:

$$\text{Log } p(t \mid X, w) = \sum_{n=1}^{N} \log N(t_n \mid \phi(x_n)^T w, \sigma^2)$$

utilizando la fórmula del logaritmo de la Gaussiana univariada.

$$\log N(t_n \mid \mu_n, \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}(t_n - \mu_n)^2$$

Sustituyendo $\mu_n = \phi(x_n)^T w$:

$$\log p(t \mid X, w) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}(t_n - \phi(x_n)^T w)^2$$

Como queremos maximizar el log-verosimilitud con respecto a w, lo cual es equivalente a minimizar la función error cuadrático

$$w^* = \arg\min_w \sum_{n=1}^{N}(t_n - \phi(x_n)^T w)^2$$

En nota matricial $w^* = \arg\min_w \|t - \Phi w\|^2$

y como se vio en el modelo 1 de mínimos cuadrados, la solución analítica es:

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

En conclusión, mínimos cuadrados y máxima verosimilitud dan la misma solución si asumimos ruido Gaussiano de varianza constante. La diferencia q radica en que ML2 tiene una justificación probabilística.

1.4 Regresión Bayesiana: Máximo a Posteriori (MAP)

En ML solo maximizamos la verosimilitud $\max_w p(t \mid X, w)$

En MAP, usamos el teorema de Bayes

$$p(w \mid t, X) \propto p(t \mid X, w) \cdot p(w)$$

y buscamos el valor más probable de w dado los datos:

$$w_{MAP} = \arg\max_w p(w \mid t, X) = \arg\max_w p(t \mid X, w) \cdot p(w)$$

El modelo de regresión con ruido Gaussiano es:

$$p(t_n \mid x_n, w) = N(t_n \mid \phi(x_n)^T w, \sigma^2)$$

El Prior Gaussiano sobre los pesos es:

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

creemos que los valores de w estan centrados en cero, con varianza inversa α (mayor α = menos confianza en pesos grandes.

Queremos maximizar $\log p(w|t,x) \propto \log p(t|X,w) + \log p(w)$

El log-Verosimilitud es:

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

El log-Prior es: logaritmo de la gaussiana multidimensional.

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

Sumamos ambos e ignoramos terminos constantes

$$\log p(w|t,x) = -\frac{1}{2\sigma^2}\|t - \Phi w\|^2 - \frac{\alpha}{2}\|w\|^2 + \text{const}$$

La función a minimizar (negativo del log posterior) es:

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

Multiplicamos por $2\sigma^2$ para simplificar (sin afectar el mínimo):

$$E(w) = \|t - \Phi w\|^2 + \lambda\|w\|^2,$$

donde $\boxed{\lambda = \sigma^2\alpha}$

La Función error es la misma que en el modelo 2 minimos cuadrados regularizados, por lo que la solución es:

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

1.5 Regresión Bayesiana Completa:

La idea principal es que en lugar de buscar un único w que maximice la posterior (como MAP), ahora consideramos que los pesos tienen una distribución posterior completa:

$$P(w|D) = \frac{P(D|w)\,p(w)}{P(D)}$$

Donde $D = \{x, t\}$ es el conjunto de datos.

Luego, la predicción bayesiana de un nuevo punto x_* es:

$$p(t_*|x_*, D) = \int p(t_*|x_*, w)\, p(w|D)\, dw$$

Esta integral promedia todas las posibles predicciones dadas por distintos w, ponderadas por la creencia posterior que tenemos sobre ellos.

El modelo de observación es:

$$p(t_n|w) = N(t_n|\phi(x_n)^T w, \sigma^2)$$

El prior gaussiano sobre los pesos es:

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

La posterior será una distribución gaussiana multivariada. Esto se puede demostrar usando propiedades de productos de gaussianas

$$p(w|D) = N(w|m_N, S_N)$$

Donde $S_N = (\alpha I + \beta \Phi^T \Phi)^{-1}$ es la covarianza posterior.
$m_N = \beta S_N \Phi^T t$ es la media del posterior.

Usamos la identidad de producto de gaussianas

Si $p(w) \sim N(w|0, \alpha^{-1} I)$ y $p(t|w) \sim N(t|\Phi w, \beta^{-1} I)$

Entonces la posterior también es gaussiana:

$$p(w|t) = N(m_N, S_N)$$

donde $\boxed{\beta = \frac{1}{\sigma^2}}$

Para predecir el valor t_*, calculamos la distribución predictiva, marginalizando los pesos:

$$p(t_* \mid x_*, D) = \int p(t_* \mid x_*, w)\, p(w \mid D)\, dw$$

Ambas distribuciones dentro del integrando son gaussianas:

$$p(t_* \mid x_*, w) = N(t_* \mid \phi(x_*)^T w, \beta^2)$$
$$p(w \mid D) = N(m_N, S_N)$$

La integral de una gaussiana con media y covarianza gaussiana da otra gaussiana.

$$\boxed{p(t_* \mid x_*, D) = N(t_* \mid \mu(x_*), \sigma^2(x_*))}$$

con media predictiva: $\boxed{\mu(x_*) = \phi(x_*)^T m_N}$

Con varianza predictiva: $\boxed{\sigma^2(x_*) = \frac{1}{\beta} + \phi(x_*)^T S_N \phi(x_*)}$

La predicción no es solo un número, sino una distribución: el modelo da la media y la varianza sobre la predicción.

1.6 Regresión Kernel Ridge (Regresión Ridge con núcleo)

Este modelo parte de la regresión Ridge, el modelo 2. Pero en vez de usar directamente una representación lineal de los datos, se emplea una transformación no lineal implícita mediante un Kernel.

En lugar de definir el modelo como: $Y(x) = \phi(x)^T w$

Lo expresamos de manera dual como una combinación de los valores del kernel entre los datos de entrenamiento y el nuevo dato x:

$$Y(x) = \sum_{n=1}^{N} a_n K(x_n, x)$$

Donde $K(x_n, x)$ es la función Kernel
$a = [a_1, \ldots, a_N]^T$ son los coeficientes del modelo en su forma dual.

El objetivo es encontrar el vector a, el cual se obtiene minimizando la función de coste regularizada en forma dual. A partir de:

$$\min_a \|Ka - t\|^2 + \lambda a^T K a$$

Donde:
$K \in \mathbb{R}^{N \times N}$ es la matriz kernel, con entradas $K_{ij} = K(x_i, x_j)$
λ es el parámetro de regularización.
$t \in \mathbb{R}^N$ es el vector de valores objetivo.

En la regresión ridge se busca

$$\min_w \|t - \Phi w\|^2 + \lambda \|w\|^2$$

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

En forma dual, definimos $w = \Phi^T a$

Entonces el modelo se convierte en:

$$y(x) = \phi(x)^T w = \phi(x)^T \Phi^T a = k(x)^T a$$

Donde:
$k(x) = [K(x_1, x), K(x_2, x), \ldots, K(x_N, x)]^T \in \mathbb{R}^N$

Reemplazando $w = \Phi^T a$ en la función de costo de ridge obtenemos:

$$E(a) = \|t - \Phi \Phi^T a\|^2 + \lambda a^T \Phi \Phi^T a$$

Definimos $K = \Phi \Phi^T$, la matriz kernel, Entonces

$$E(a) = \|t - Ka\|^2 + \lambda a^T K a$$

Derivamos con respecto a a, utilizando la siguiente identidad matricial:

$$\frac{d}{dx} \|Ax - b\|^2 = 2A^T(Ax - b) \rightarrow \frac{dE}{da} = -2K^T(t - Ka) + 2\lambda Ka$$

Como K es simétrica ($K^T = K$), tenemos:

$\frac{dE}{da} = -2K(t - Ka) + 2\lambda Ka$, Igualando a cero

$$-2Kt + 2K^2 a + 2\lambda Ka = 0$$

Dividiendo por 2 y factorizando K, obtenemos:

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

Multiplicamos ambos lados por K^{-1} (asumiendo que K es invertible):

$(K+\lambda I)a = t$, despejando a → $\boxed{a = (K+\lambda I)^{-1}t}$

Entonces, la predicción para un nuevo x es:

$$\boxed{y(x) = \sum_{n=1}^{N} a_n k(x_0, x) = k(x)^T a}$$

Esta predicción es muy útil cuando los datos no son linealmente separables en el espacio original. El modelo depende solo de los valores del Kernel, no de $\phi(x)$ directamente.

1.7 Regresión con proceso Gaussiano (GPR)

Un proceso gaussiano es una distribución conjunta infinita de variables aleatorias, en donde asumimos que los valores de salida $y(x)$ están generados por un proceso gaussiano con cierta media y covarianza entre puntos.

$$f(x) \sim GP(0, k(x, x'))$$

$f(x)$ es la función desconocida que queremos estimar
$k(x, x')$ es la función Kernel que define la covarianza entre los puntos x y x'. Se asume media cero por simplicidad.

El objetivo es obtener la distribución a posteriori de las predicciones f_* para un nuevo conjunto de entradas X_*, dados los datos de entrenamiento (x, t).

Dado los datos de entrenamiento $X = [x_1, x_2, \dots, x_n]$ y los nuevos puntos $X_* [x_*^1, x_*^2, \dots, x_*^M]$.

Asumimos una distribución conjunta gaussiana sobre los valores de salida verdaderos en el entrenamiento f y las predicciones en test f_*:

$$\begin{bmatrix} f \\ f_* \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} K(X,X) & K(X,X_*) \\ K(X_*,X) & K(X_*,X_*) \end{bmatrix} \right)$$

Donde:

"$K(X,X) \in \mathbb{R}^{N\times N}$: Covarianza entre los puntos de entrenamiento."
"$K(X,X_*) \in \mathbb{R}^{N\times M}$: Covarianza entre test y train."
"$K(X_*,X_*) \in \mathbb{R}^{M\times M}$: Covarianza entre los puntos test."

En la práctica, las observaciones t tienen ruido:

$$t = f + \varepsilon, \quad \varepsilon \sim N(0, \sigma_n^2 I)$$

Entonces la covarianza de las observaciones se vuelve:

$$\text{cov}(t) = K + \sigma_n^2 I$$

Dado que estamos en una distribución gaussiana conjunta, la distribución condicional también es gaussiana:

$$p(f_*|X_*, X, t) = N(\bar{f}_*, \text{cov}(f_*))$$

con
$$\bar{f}_* = K_*^T (K + \sigma_n^2 I)^{-1} t$$
$$\text{cov}(f_*) = K_{**} - K_*^T (K + \sigma_n^2 I)^{-1} K$$

Aplicando la identidad matricial de condicional de una distribución normal multivariada.

Si
$$\begin{bmatrix} x \\ y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix}, \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}\right)$$

Entonces:
$$p(y|x) \sim N(\mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x), \Sigma_{yy} - \Sigma_{yx}\Sigma_{xx}^{-1}\Sigma_{xy})$$

Aquí se aplica con
$$y = f_*$$
$$x = t$$
$$\Sigma_{xx} = K + \sigma_n^2 I$$
$$\Sigma_{xy} = K_*^T$$
$$\Sigma_{yy} = K_{**}$$

Media posterior: $\bar{f}_* = K_*^T (K + \sigma_n^2 I)^{-1} t$

Covarianza posterior: $K_{**} - K_*^T (K + \sigma_n^2 I)^{-1} K$

Esto da una distribución completa sobre las predicciones se puede usar solo la media $\bar{f}_*$ como predicción puntual o también se puede usar la varianza si se quieren intervalos de confianza.