Técnicas para construir Kernels

Resolver por Lagrange para $q = 2$ y

$\|M\|_2 = 1$, $M \geq 0$

función $\hat{B}\left(\sum_{j=1}^{d} M_j * (x_j, x_y)\right)$

Restricciones: $\|M\|_q^2 \leq \{$

Maximizar:

$$\mathcal{L}(M, \lambda) = \hat{B}\left(\sum_{j=1}^{} M_j K(x_j, x_y)\right) + \lambda\left(1 - \sum_{j=1}^{} M_j^2\right)$$

- λ = multiplicador lagrange asociado a la restricción $q = 2$, es decir, norma L_2, implica $\|M\|_2^2 = 1$
- Término $\sum_{j=1}^{} M_j^2$ representa norma L_2 de los M_j

Valores optimos de M_j, lagrangiano respecto a M_j y λ.

$$\frac{\partial \mathcal{L}}{\partial M_j} = \frac{\partial \hat{B}}{\partial M_j} + \lambda(-2M_j) = 0$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = 1 - \sum_{j=1}^{} M_j^2 = 0$$

Kernel.

Hilbert Embeddings.

Si $\mathbb{E}_P \{ (K(x,x))^{1/2} \} < \infty$, $\rightarrow M_P \in \mathcal{F}$.

Operador T_P

$$T_P f = \mathbb{E}_P [K(x,x') f(x)]$$

$K(x,x')$ Kernel mide similitud x y x'.

T_P. operador acotado, funcion M_P tal que $T_P f = (f, M_P)$

Condicion Inicial:

Si $\mathbb{E}_P [\sqrt{K(x,x')}] < \infty$, M_P. pertenece espacio funciones $\mathcal{F}$.

$T_P f$.

$$T_P f = \mathbb{E}_P [K(x,x') f(x)]$$

$f(x)$ funcion actua operador

Desigualdad Jenson cota superior:

$$|T_P f| = |\mathbb{E}_P [K(x,x') f(x)]|$$

$$|T_P f| \leq \mathbb{E}_P |K(x,x')| |f(x)|]$$

$$|T_p f| \leq E_p\left[\|f(x)\| \sqrt{K(x,x')}\right]$$

$$T_p f = \langle f, M_p \rangle f.$$

$M_p(x) = E_p[K(x,x')]$, M_p. valor esperado Kernel,

GPR. Gaussian Process Regression.

Funcion: $f(x) \sim GP(m(x), K(x,x'))$

donde $m(x) = 0$ y $K(x,x')$, Kernel o funcion covarianza.

$\{x, y\}$ con $y = f(x) + \epsilon$, donde $\epsilon \sim N(0, \sigma^2)$ es el ruido gaussiano, el objetivo es predecir el valor de $f(x^*)$ para un nuevo punto x^*. La predicción $f(x^*)$ dada la observación y está dada por:

$$P(f^* | X, y, x^*) \sim N(M(x^*), \sigma^2(x^*))$$

donde.

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

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

Aquí, $K(X,X)$ es la matriz de covarianza construida con los puntos X.

Optimización:

hiperparámetros, longitud escala ℓ y varianza σ maximizando logaritmo de la verosimilitud marginal:

$$\log p(y|X) = -\frac{1}{2} y^T [K(X,X) + \sigma^2 I]^{-1} y - \frac{1}{2} \log |K(X,X) + \sigma^2 I| - \frac{n}{2} \log 2\pi$$

Optimiza respecto a los hiperparámetros del kernel usando métodos como L-BFGS o Adam.

GPC Gaussian Process Clasificatrion.

función $p(y=1|f(x)) = \phi(f(x))$.

ϕ función sigmoide o probit.

Optimización:

Métodos de aproximación.

Laplace: Aproxima verosimilitud marginal alrededor del valor máximo posterior.

Expectation Propagation (EP): Aproximación iterativa para computar la verosimilitud marginal.

VGP. Variational Gaussian Process

función: Aproximación variacional para conjunto datos grande.

Aproximar $p(f|X,y)$ con distribución variacional $q(f)$.

Evidencia variacional inferior. Cota inferior a la log-verosimilitud marginal:

$$ELBO = E_q[\log p(y|f)] - KL(q(f) \| p(f))$$

KL es divergencia Kullback-Leibler entre aproximación variacional $q(f)$ y la verdadera distribución posterior $p(f)$.

Optimización:

Elbo se maximiza respecto parámetros variacionales y los hiperparámetros del kernel. El gradiente de la Elbo se calcula mediante técnicas como la diferenciación automática y se optimiza con métodos de gradiente.

SGPR. Spase Gaussian Process Regression.

función.

Conjunto puntos ~~introducción~~ inducción z, que se aproximan a posterior.

Reducir complejidad computacional. de $O(n^3)$ a $O(m^2 n)$, donde m $(m \ll n)$.

ELBO en SGPR es:

$$ELBO = \mathbb{E}_q[\log p(y|f)] - KL(q(f) \| p(f|z))$$

Optimización:

Maximizar puntos de inducción z y los hiperparámetros del Kernel maximizando ELBO.

Se realiza con optimización de gradiente, (Adam).

SVGP. Sparse Variational Gaussian Process

Generaliza a clasificación y Regresión, manteniendo estructura y puntos de inducción para hacer aproximación variacional más eficiente.

$ELBO = \mathbb{E}_q[\log p(y|f)] - KL(q(u) \| p(u))$.

$q(u)$ es distribución variacional sobre puntos de inducción

Optimización.: Elbo respecto puntos inducción e hiperparámetros. del Kernel.