Introduction to Gaussian Processes

Part 4: Active Learning
Discrete case

Ruben Martinez-Cantin

Defense University Center
Zaragoza, Spain
rmcantin@unizar.es

Outline

What you will see here:

- Gaussian process hyperparameters
- Regression
- Binary classification
- Active learning and experimental design
- Submodularity
- Bayesian optimization
- Stochastic bandits

Discrete experimental design

- Now we assume that we can only access a set of points $\mathcal{V}$
  - Image and sound processing
  - Sensor placement
  - Robot planning
  - Processing biological samples
- Pool-based active learning
  - It reduces the active learning bias
- We can compute non-greedy strategies
  - Although the complexity is NP-hard.

Non-greedy predictions

- We are going to generalize our prediction model to multiple outputs.
  - Previous inputs $\rightarrow$ set A.
  - Design inputs $\rightarrow$ set B.

$$\hat{y}_B | x_B, \mathbf{x}_A, y_A = K(x_B, \mathbf{x}_A)(K(\mathbf{x}_A, \mathbf{x}_A) + \sigma_n^2 I)^{-1} y_A$$
$$\Sigma_B | x_B, \mathbf{x}_A, y_A = K(x_B, x_B) - K(x_B, \mathbf{x}_A)(K(\mathbf{x}_A, \mathbf{x}_A) + \sigma_n^2 I)^{-1} K(\mathbf{x}_A, x_B)$$

- Now $\hat{y}_B$ is a vector and Σ_B is a matrix

Including hyperparameters

- If we add some extra parameters to the model such as

$$\begin{bmatrix} y_B \\ y_A \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{1} \\ \mathbf{1} \end{bmatrix} \mu, \sigma_s^2 \begin{bmatrix} K(\mathbf{x}_B, \mathbf{x}_B) & K(\mathbf{x}_B, \mathbf{x}_A) \\ K(\mathbf{x}_A, \mathbf{x}_B) & K(\mathbf{x}_A, \mathbf{x}_A) + \sigma_n^2 I \end{bmatrix} \right)$$

- where $\mu \sim \mathcal{N}(0, \delta^2 \sigma_s^2)$ and $\sigma_s^2 \sim \mathcal{IG}(a/2, b/2)$
- Then we can compute the predictions

$$\hat{y}_{B|A} = \hat{\mu} + K_{BA} K_{AAn}^{-1} (y_A - \mathbf{1}\hat{\mu})$$

$$\Sigma_{B|A} = \hat{\sigma}_s^2 \left[K_{BB} - (K_{BA} K_{AAn}^{-1} K_{AB}) \frac{\alpha^T \alpha}{\mathbf{1}^T K_{AAn}^{-1} \mathbf{1} + \delta^{-2}} \right]$$

- where

$$K_{AAn} = K_{AA} + \sigma_n^2 I \quad \alpha = \mathbf{1} - \mathbf{1}^T K_{AAn}^{-1} K_{AB}$$

$$\hat{\mu} = \frac{\mathbf{1}^T K_{AAn}^{-1} y_A}{\mathbf{1}^T K_{AAn}^{-1} \mathbf{1} + \delta^{-2}} \quad \hat{\sigma}_s^2 = \frac{b + y_A^T K_{AAn}^{-1} y_A - (\mathbf{1}^T K_{AAn}^{-1} \mathbf{1} + \delta^{-2}) \hat{\mu}^2}{n + a + 2}$$

Bayesian experimental design recap

- Let assume that $\{\lambda_i\}_{i=1}^{N}$ are the eigenvalues of $\Sigma_{B|A}$
- Bayesian A-optimality (expected mean squared error)

$$EMSE_{B|A} = \text{trace}(\Sigma_{B|A}) = \sum_{i=1}^{N} \lambda_i$$

- Bayesian D-optimality (entropy)

$$\mathcal{H}_{B|A} = \frac{1}{2} \log |\Sigma_{B|A}| + \frac{N}{2} \log \pi e \propto \log \prod_{i=1}^{N} \lambda_i$$

- Bayesian E-optimality (maximum predictive variance)

$$MPV_{B|A} = \max_i \lambda_i$$

- None of them depends on the outputs $y_{B|A}$!

Submodular functions

Definition

A set function $f : 2^V \to \mathbb{R}$ is called *submodular* if is satisfies

$$f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y) \qquad \forall\ X, Y \subseteq V$$

But this definition can be expressed in a more interesting way:

Definition

A set function $f : 2^V \to \mathbb{R}$ is called *submodular* if is satisfies

$$f(X + e) - f(X) \geq f(Y + e) - f(Y) \qquad \forall\ X \subset Y \subset Y + e \subseteq V$$

Submodular optimization

Theorem (Nemhauser et al., 1978)

Let F be a monotone submodular set function over a finite ground set V with $F(\emptyset) = 0$. Let A_G be the set of the first k elements chosen by the greedy algorithm, and let $OPT = \max_{A \subset V, |A|=k} F(A)$. Then

$$F(A_G) \geq OPT(1 - 1/e)OPT$$

- Example: Sensor localization using Mutual Information