Байесовское мультимоделирование: гауссовские процессы

Московский Физико-Технический Институт

2021

Определение (wiki)

- Случайный процесс f_t с непрерывным временем является гауссовским тогда и только тогда, когда для любого конечного множества индексов $t_1, \ldots, t_k$: $f_{t_1}, \ldots, f_{t_k}$ — многомерная гауссовская случайная величина.
- Всякая линейная комбинация имеет $f_{t_1}, \ldots, f_{t_k}$ одномерное нормальное распределение.

Определение (упрощенное)

Назовем гауссовским процессом $\mathcal{GP}(\boldsymbol{m(x)}, \boldsymbol{k(x, x')})$ распределение на множестве функций, такое что для каждых $\boldsymbol{x}, \boldsymbol{x'}$: $\mathcal{GP}(\boldsymbol{m(x)}, \boldsymbol{k(x, x')})$ — гауссовское распределение.

Пример: регрессия

$$\boldsymbol{f} \sim \mathcal{N}(0, \boldsymbol{K}),$$

где $\boldsymbol{K}$ — ковариационная матрица для объектов выборки $\boldsymbol{X}$.

$$y \sim f + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

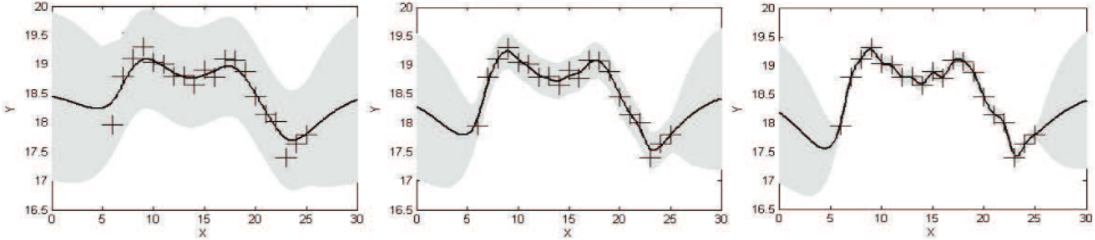
Тогда $\boldsymbol{y} \sim \mathcal{N}(0, \boldsymbol{K} + \sigma^2 \boldsymbol{I})$, предсказание для новых объектов $\hat{\boldsymbol{y}}$:

$$\hat{\boldsymbol{y}} \sim \mathcal{N}({\boldsymbol{K}'}^{\mathsf{T}} (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}, \boldsymbol{K}'' - {\boldsymbol{K}'}^{\mathsf{T}} (\boldsymbol{K} + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{K}').$$

Отличие гауссовых процессов от других подходов

- Модель непараметрическая
  - параметризация только ковариационной функции и уровня шума
  - оптимизация — через ОМП
- Априорное распределение задается для функции, а не для параметров
  - задание гауссового априорного распределения на параметры: тоже получаем гауссовый процесс с вырожденной матрицей ковариации
- Сложность предсказания: $O(N^3)$.

Влияние σ^2 на предсказательную способность



(McDuff, 2010)

Ковариационная функция

Основные свойства:

- Симметричность $\boldsymbol{K}(x, x') = \boldsymbol{K}(x', x)$
- Положительная полуопределенность: $\boldsymbol{v}^{\mathsf{T}} \boldsymbol{K} \boldsymbol{v} \geq 0$
- Стационарность: $K(x, x') = K(x + a, x' + a)$
- Изотропность: зависимость только от $||x - x'||$

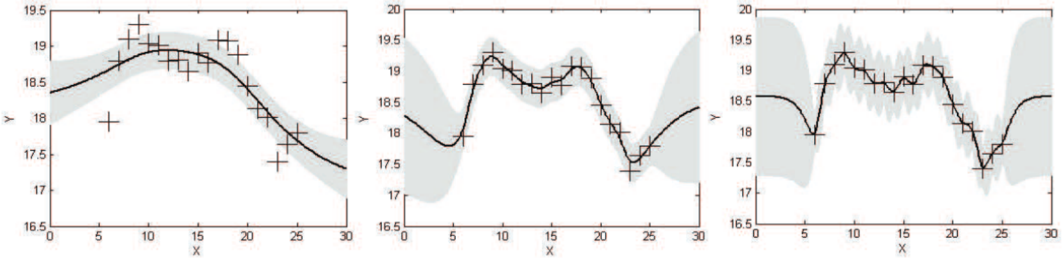
Ковариационная функция: виды

Экспоненциальная:

$$K = \sigma_0^2 \text{exp}\left(\frac{-(x-x')^2}{2\lambda}\right)$$

- Линейная: $K = \sigma_0^2 + xx'$
- Броуновская: $K = \min(x, x')$
- Периодическая: $K = \text{exp}(\frac{-2\text{sin}^2(\frac{x-x'}{2})}{\lambda^2})$
- Задаваемая нейросетью

Влияние λ на предсказательную способность



(McDuff, 2010)

Ковариация Матерна

$$K = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x - x'|}{\lambda} \right)^{\nu} K_{\nu} \left(\frac{\sqrt{2\nu}|x - x'|}{\lambda} \right),$$

K_{ν} — модифицированная функция Бесселя.

- Стационарная и изотропная функция
- При $\nu \to \infty$: экспоненциальная функция
- При конечных ν: предсказательная функция менее гладкая

Оптимизация гиперпараметров модели с применением гауссовых процессов

(Snoek et al., 2012)

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{w}(\boldsymbol{h}), \boldsymbol{h}|\boldsymbol{x}).$$

Модель $\boldsymbol{f}$ рассматривается как функция от гиперпараметров:

$$\boldsymbol{f} \sim \mathcal{GP}, \quad \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{f}, \sigma^2).$$

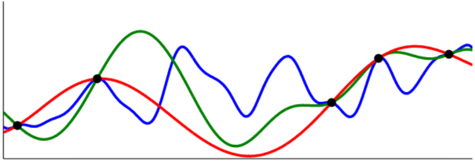
Используется ковариация Матерна, с $\nu = 2.5$.

Выбор следующей точки для оценки

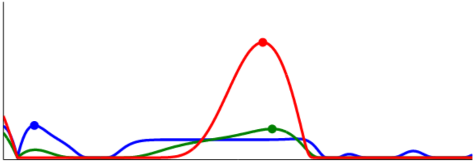
Выбор производится с применением одной из следующий функций (Acquisition function):

- Probability of Improvement: $\Phi(\gamma), \gamma = \frac{f(\boldsymbol{h}^*)-\boldsymbol{\mu}(\boldsymbol{h})}{\boldsymbol{\sigma}(\boldsymbol{h})}$
- Expected Improvement: $\approx \mathsf{E}\gamma$
- Confidence Bound: $\boldsymbol{\mu}(\boldsymbol{h}) - k\boldsymbol{\sigma}(\boldsymbol{h})$.

Влияние λ на оптимизацию гиперпараметров

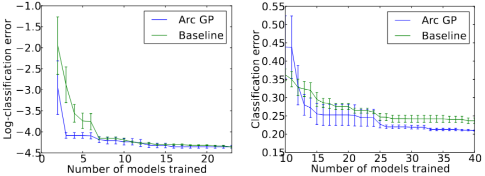


(a) Posterior samples under varying hyperparameters



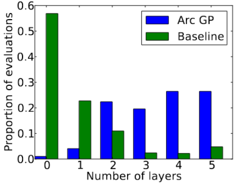
(b) Expected improvement under varying hyperparameters

Результаты



(a) MNIST (b) CIFAR-10

Figure 2: Bayesian optimization results using the arc kernel.



Ковариационная функция для условных параметров

Задача выбора архитектуры нейросети:

- Количество нейронов на скрытом слое: **ок, можем задавать как вещественный гиперпараметр**
- Количество слоев на скрытом слое: **ок, можем задавать как вещественный гиперпараметр**
- **Как сравнивать две архитектуры с разным количеством слоев? Как сравнивать архитектуру [100, 100] и [100]?**

(Swersky et al., 2014): нужна специальная ковариационная функция!

Идея ковариационной функции

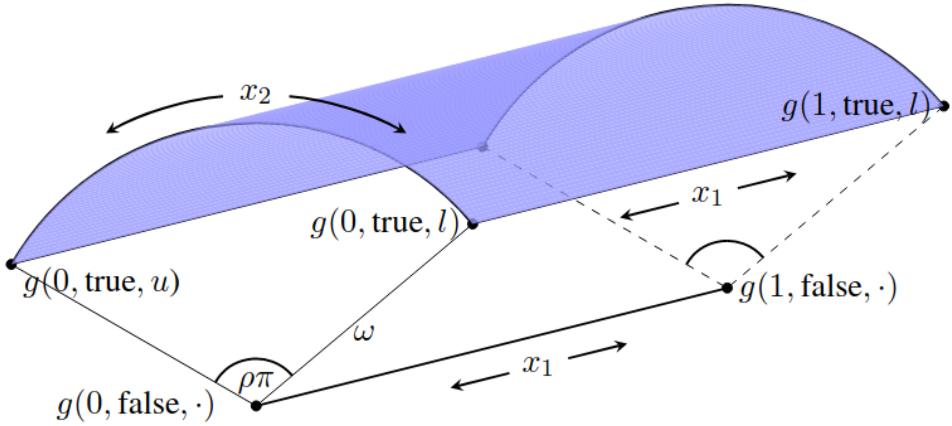
- If we are comparing two points for which the same parameters are relevant, the value of any unused parameters shouldn't matter,

$$k\big((x_1, \text{false}, x_2), (x'_1, \text{false}, x'_2)\big) = k\big((x_1, \text{false}, x''_2), (x'_1, \text{false}, x'''_2)\big), \ \forall x_2, x'_2, x''_2, x'''_2; \quad (1)$$

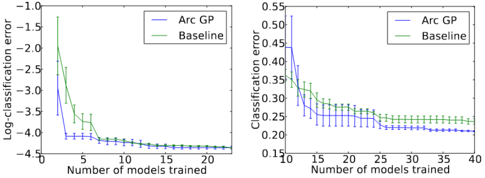
- The covariance between a point using both parameters and a point using only one should again only depend on their shared parameters,

$$k\big((x_1, \text{false}, x_2), (x'_1, \text{true}, x'_2)\big) = k\big((x_1, \text{false}, x''_2), (x'_1, \text{true}, x'''_2)\big), \ \forall x_2, x'_2, x''_2, x'''_2. \quad (2)$$

Идея ковариационной функции

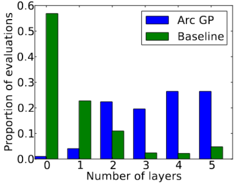


Результаты



(a) MNIST (b) CIFAR-10

Figure 2: Bayesian optimization results using the arc kernel.



Литература

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- Daniel McDuff, Gaussian Processes, https://courses.media.mit.edu/2010fall/mas622j/ProblemSets/slidesGP.pdf
- Chris Williams, Gaussian Processes for Machine Learning, https://www.newton.ac.uk/files/seminar/20070809140015001-150844.pdf
- Ed Snelson, utorial: Gaussian process models for machine learning, https://mlg.eng.cam.ac.uk/tutorials/06/es.pdf
- Snoek J., Larochelle H., Adams R. P. Practical bayesian optimization of machine learning algorithms //Advances in neural information processing systems. – 2012. – Т. 25.
- Swersky K. et al. Raiders of the lost architecture: Kernels for Bayesian optimization in conditional parameter spaces //arXiv preprint arXiv:1409.4011. – 2014.