Нейросетевая аппроксимация плотности для построения вероятностно-метрического пространства

Вареник Наталия Викторовна

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Научный руководитель: д.ф.-м.н. В.В. Стрижов

Москва, 2020

Аппроксимация плотности молекулярной конфигурации

Задача

Разработать нейросетевую модель восстановления плотности распределения пространственной молекулярной конфигурации.

Требования к функции плотности распределения

- Дифференцируема по носителю, носитель – сфера
- Имеет явную функциональную зависимость

Проблема

Гистограммный метод, метод окна Парзена-Розенблатта, смесь гауссиан не удовлетворяют данным требованиям.

Решение

Использование монотонного перцептрона для аппроксимации функции распределения пространственной конфигурации и получение плотности путем дифференцирования.

Текущее состояние исследования и прикладной задачи

Исследование молекулярной конфигурации:

- *M. Kadukova, S. Grudinin*. 2017. Convex-PL: a novel knowledge-based potential for protein-ligand interactions deduced from structural databases using convex optimization// J Comput Aided Mol Des.
- *J. R. Lopez-Blanco and P. Chacon*. 2019. Korp: knowledge-based 6d potential for fast protein and loop modeling// Bioinformatics.

Существующие решения нейросетевой аппроксимации плотности:

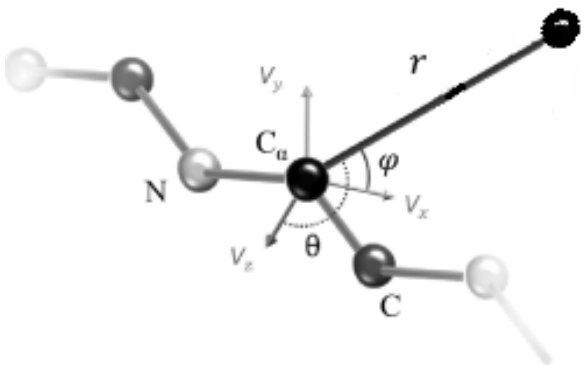
- *M. Magdon-Ismail and A. Atiya*. 2002. Density estimation and random variate generation using multilayer networks// IEEE Transactions on Neural Networks.
- *S. Zhang*. 2018. From CDF to PDF - A density estimation method for high dimensional data// CoRR.

Пространственная конфигурация пары белок-лиганд

Локальная
система координат аминокислоты:

$$\begin{cases} \mathbf{V}_z = (\mathbf{r}_{CC_\alpha} + \mathbf{r}_{NC_\alpha})/|\mathbf{r}_{CC_\alpha} + \mathbf{r}_{NC_\alpha}|, \\ \mathbf{V}_y = (\mathbf{V}_z \times \mathbf{r}_{NC_\alpha})/|\mathbf{V}_z \times \mathbf{r}_{NC_\alpha}|, \\ \mathbf{V}_x = \mathbf{V}_y \times \mathbf{V}_z. \end{cases}$$

Векторы из C_α в C, N:
$\mathbf{r}_{CC_\alpha} = \mathbf{r}_C - \mathbf{r}_{C_\alpha}$, $\mathbf{r}_{NC_\alpha} = \mathbf{r}_N - \mathbf{r}_{C_\alpha}$.



Пространственная конфигурация аминокислоты и лиганда [Pablo Chacon, 2019]

Постановка задачи восстановления плотности

Дано

$\mathfrak{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, $\mathbf{x}_i = (r, \theta, \varphi)^T \in \Omega$, где r – расстояние между аминокислотой и лигандом, θ и φ – сферические координаты лиганда в с. к. аминокислоты, $\Omega = [3\text{Å}, 20\text{Å}] \times [0, \pi] \times [0, 2\pi]$, y_i – значение эмпирической функции распределения в точке $\mathbf{x}_i$.

Монотонный перцептрон

$\mathfrak{F} = \{\mathbf{f} : (\mathbf{X}, \mathbf{w}) \mapsto \mathbf{y}\}$, $\mathbf{w} \in \mathbb{W}$ – параметры модели.

Аппроксимация функции распределения

- $S(\mathbf{y}, \mathbf{X}, \mathbf{w}) = ||\mathbf{y} - \mathbf{f}(\mathbf{X}, \mathbf{w})||_2^2$ – функция потерь,
- $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{W}} S(\mathbf{w} \mid \mathbf{f}, \mathbf{y}, \mathbf{X})$ – оптимизация параметров.

Плотность распределения

$$\hat{p}(\mathbf{x}|\mathbf{w}) = \frac{\partial^3}{\partial r \partial \theta \partial \varphi} f(\mathbf{x}, \mathbf{w}).$$

Эмпирическая функция распределения

Оценка эмпирической функции распределения

$$\hat{y}_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\Theta(\mathbf{x} - \mathbf{x}_i),$$

где $\Theta(\mathbf{x}) = 1$, если $x_j \geq 0$, $j = \overline{1, d}$, иначе 0.

Свойство эмпирической функции распределения

По усиленному закону больших чисел $\hat{y}_n(\mathbf{x})$ сходится почти наверное к теоретической функции распределения $y(\mathbf{x})$:

$$\hat{y}_n(\mathbf{x}) \overset{\text{п.н.}}{\to} y(\mathbf{x}), n \to \infty.$$

Многослойный перцептрон как универсальный аппроксиматор монотонной функции

Монотонный многослойный перцептрон

- $f(\mathbf{x}) = \tilde{\mathbf{W}}_N \sigma\left(\ldots \sigma\left(\tilde{\mathbf{W}}_2 \sigma\left(\tilde{\mathbf{W}}_1 \mathbf{x} + \mathbf{b}_1\right) + \mathbf{b}_2\right)\ldots\right) + b_N$,
- $\tilde{\mathbf{W}}_i = \exp(\mathbf{W}_i)$,
- $\sigma(\cdot)$ – сигмоидная функция активации.

Получаем монотонно возрастающую модель как композицию монотонно возрастающих функций.

Теорема (Bernhard Lang)

Определенный выше многослойный перцептрон является универсальным аппроксиматором неубывающей функции.

Оценка плотности распределения

Многомерная плотность распределения

Смешанная производная d-го порядка от d-мерной функции совместного распределения: $p(\mathbf{x}) = \dfrac{\partial^d}{\partial x_1, \ldots, \partial x_d} F(\mathbf{x})$.

Оценка плотности дифференцированием модели

Формула для перцептрона с одним скрытым слоем:

$$\hat{p}(\mathbf{x}) = \frac{\partial^d}{\partial x_1, \ldots, \partial x_d} f(\mathbf{x}) =$$

$$\sum_{i=1}^{H} e^{w_i^{(2)}} \prod_{j=1}^{d} e^{W_{ij}^{(1)}} \sigma^{(d)} \left(\sum_{j=1}^{d} e^{W_{ij}^{(1)}} x_j + b_i^{(1)} \right) =$$

$$\sum_{i=1}^{H} e^{w_i^{(2)} + \sum_{j=1}^{d} W_{ij}^{(1)}} \sigma^{(d)} \left(\sum_{j=1}^{d} e^{W_{ij}^{(1)}} x_j + b_i^{(1)} \right).$$

Вычислительный эксперимент

Цель эксперимента

Восстановить плотность пространственной конфигурации пары аминокислота-лиганд с помощью предложенного монотонного многослойного перцептрона.

Данные

В базе данных 20 аминокислотных остатков. Аминокислотные остатки взаимодействуют с лигандами, их 40. Взаимодействие описывается 5 признаками: тип аминокислотного остатка, тип лиганда, расстояние и 2 угла θ и φ.

Гипотеза

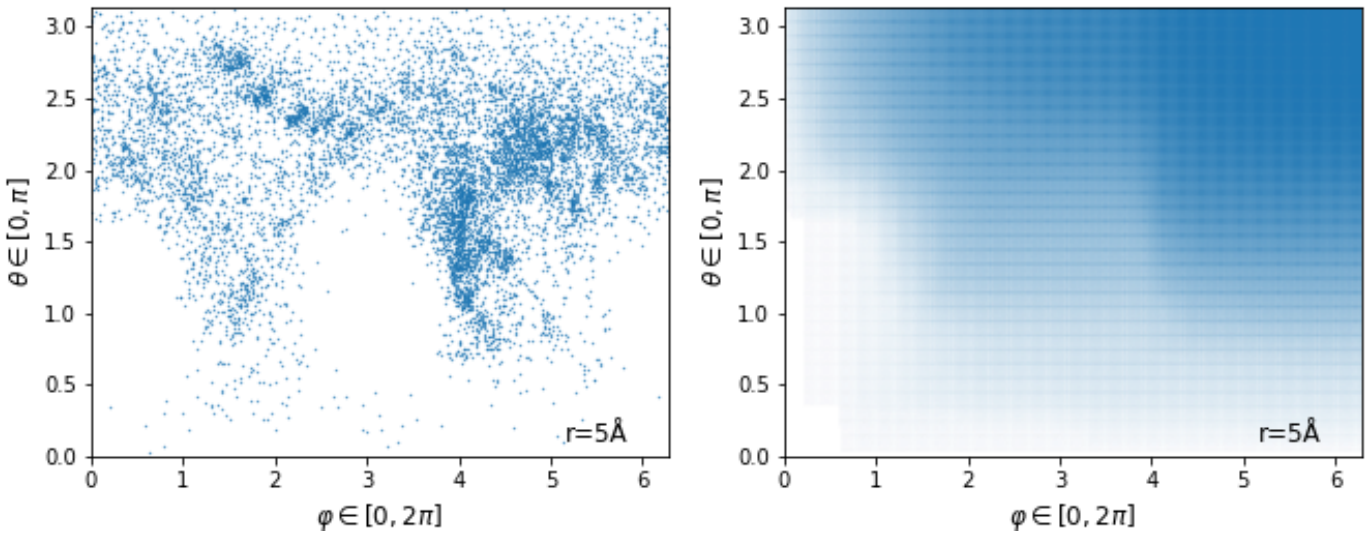
Максимумы полученной плотности распределения соответствуют устойчивым пространственным конфигурациям.

Вычислительный эксперимент

Для обучения монотонного перцептрона на носителе выборки вводится равномерная сетка и в каждом узле сетки вычисляется значение эмпирической функции рапсределения:

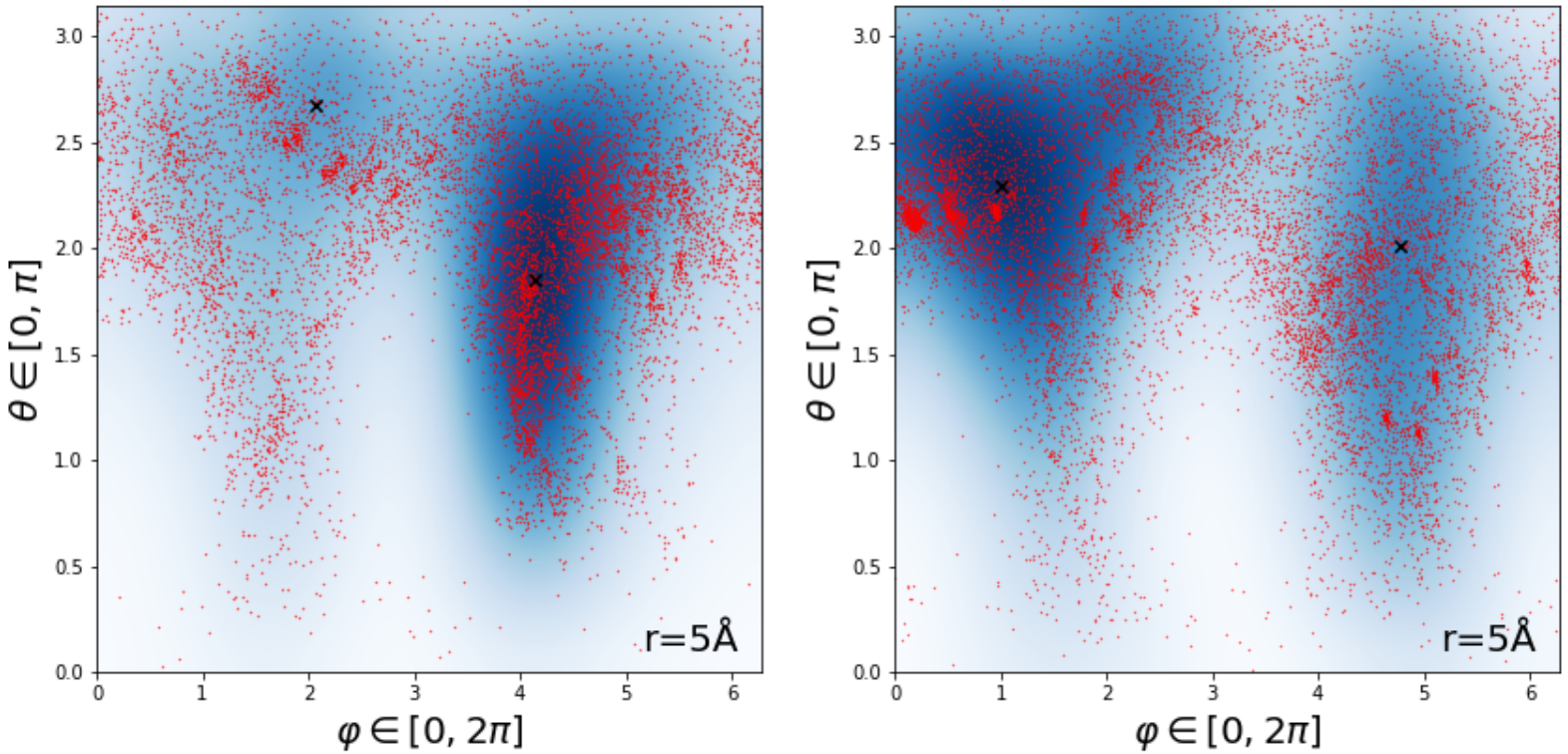
$$y(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\Theta(\mathbf{x} - \mathbf{x}_i)$$

где $\Theta(\mathbf{x}) = 1$, если $x_j \geq 0$ $j = \overline{1,3}$, иначе 0.



Посчитанная функция распределения для (C_ar6x, ALA).

Результаты восстановления плотности



Полученная плотность для пар (C_ar6x, ALA) и (C_ar6x, CYS).

Синим отмечена восстановленная плотность распределения. Красным – встречающиеся в базе данных пространственные конфигурации пары. Черным – найденные экстремумы.

Результаты, выносимые на защиту

- Предложен нейросетевой метод восстановления совместной плотности распределения пространственной конфигурации аминокислота-лиганд.
- Введена модификация многослойного перцептрона, являющаяся универсальным аппроксиматором монотонной функции.
- В результате эксперимента показана адекватность работы метода на исследуемой выборке.