О глубоких гауссовских моделях в задачах машинного обучения

А. Р. Ибрагимова[1], А. К. Горшенин[2]

Работа посвящена исследованию глубоких нейросетевых архитектур с реализацией смесей нормальных распределений в скрытых слоях для решения задач кластеризации и регрессии. Проведено сравнение таких моделей с различными наборами гиперпараметров относительно классических методов: k-средних, линейной регрессии, смешанных гауссовских моделей GMM и других.

Ключевые слова: глубокие нейронные сети, смеси нормальных распределений, EM-алгоритм.

1. Введение

Модели глубокого обучения в последнее время становятся все более популярными в задачах машинного обучения. В статье [1] для решения задачи кластеризации был предложен метод на основе глубокой гауссовской модели DGMM, обобщающей классические конечные нормальные смешанные распределения на несколько слоев нейронной сети.

Ключевым является представление k-компонентной смеси для некоторого p-мерного случайного вектора y $f(y;\theta) = \sum\limits_{j=1}^{k} p_j \varphi^{(\mathrm{p})}(y;\mu_j,\Sigma_j)$, где $\theta = (p_1,\ldots,p_k,\mu_1,\ldots,\mu_k,\Sigma_1,\ldots,\Sigma_k)$, $p_1,\ldots,p_k > 0$ — веса компонент, такие что $\sum_{j=1}^{k} p_j = 1$, μ_j,Σ_j — математическое ожидание и ковариационная матрица j-ой компоненты смеси в виде смеси факторных анализаторов [2]: $y = \mu_j + \Lambda_j z + u \quad$ с вероятностью p_j, $j = \overline{1,k}$, где z — латентная p-мерная переменная, $u \sim N(0,\Psi_j)$ — вектор случайной ошибки, Λ_j — квадратная матрица факторных нагрузок размерности p, Ψ_j — диагональная матрица, такая что $\Sigma_j = \Lambda_j \Lambda_j^{\mathrm{T}} + \Psi_j$.

---

[1] *Ибрагимова Альфия Рустемовна* — магистрант кафедры математической статистики факультета вычислительной математики и кибернетики МГУ имени М.В. Ломоносова, e-mail: alfiyakzn@gmail.com.

Ibragimova Alfiya Rustemovna — MSc student, Lomonosov Moscow State University, Faculty of Computational Mathematics and Cybernetics.

[2] *Горшенин Андрей Константинович* — д.ф.-м.н., доцент, руководитель отдела Федерального исследовательского центра «Информатика и управление» Российской академии наук, e-mail:agorshenin@frccsc.ru.

Gorshenin Andrey Konstantinovich — Doctor of Science, Associate Professor, Head of Department, Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences.

Для оценивания неизвестных параметров модели применяется EM-алгоритм [3] и широко его использующаяся стохастическая модификация [4, 5].

В данной работе проведено сравнение рассматриваемой модели с классическими методами кластеризации без учителя и с учителем, включая алгоритмы k-средних, экстремального градиентного бустинга, опорные вектора и случайные леса, на основе тестирования на наборах открытых данных репозитория UCI Machine Learning.

Рассмотрены возможные подходы к построению аналога подобного метода для решения регрессионных задач.

2. Архитектура нейронной сети

Глубокая гауссовская смесь из h слоев случайного вектора y имеет вид:

$$\begin{array}{ll}
(1) & y = \eta^{(1)}_{s_1} + \Lambda^{(1)}_{s_1} z^{(1)} + u^{(1)} \ \text{ с вероятностью } \ p^{(1)}_{s_1},\ s_1 = \overline{1, k_1}, \\
(2) & z^{(1)} = \eta^{(2)}_{s_2} + \Lambda^{(2)}_{s_2} z^{(2)} + u^{(2)} \ \text{ с вероятностью } \ p^{(2)}_{s_2},\ s_2 = \overline{1, k_2}, \\
 & \dots \\
(h) & z^{(h-1)} = \eta^{(h)}_{s_h} + \Lambda^{(h)}_{s_h} z^{(h)} + u^{(h)} \ \text{ с вероятностью } \ p^{(h)}_{s_h},\ s_h = \overline{1, k_h},
\end{array}$$

где $z^{(1)}, \dots, z^{(h)}$ — скрытые переменные на слоях, имеющие размерности $r_1, \dots, r_h$, при $z^{(h)} \sim N(0, I_p)$, $\eta^{(1)}_{s_1}, \dots, \eta^{(h)}_{s_h}$ — векторы средних длины p, $\Lambda^{(1)}_{s_1}, \dots, \Lambda^{(h)}_{s_h}$ — матрицы факторных нагрузок размерности $r_{h-1} \times r_h$, при $r_0 = p$, $u^{(l)} \sim N(0, \Psi^{(l)}_{s_l})$, $l = 1, \dots, h$ — векторы случайных ошибок.

Тогда функция плотности распределения y имеет вид:

$$f(y|\theta) = \sum_{s_1=1}^{k_1} p^{(1)}_{s_1} \dots \sum_{s_h=1}^{k_h} p^{(h)}_{s_h} \varphi^{(\mathrm{p})} \left(\eta^{(1)}_{s_1} + \sum_{l=2}^{h} \left(\prod_{m=1}^{l-1} \Lambda^{(m)}_{s_m} \right) \eta^{(l)}_{s_l}, \right.$$
$$\left. \Psi^{(1)}_{s_1} + \sum_{l=2}^{h} \left(\prod_{m=1}^{l-1} \Lambda^{(m)}_{s_m} \right) \Psi^{(l)}_{s_l} \left(\prod_{m=1}^{l-1} \Lambda^{(m)}_{s_m} \right)^{\top} \right).$$

Например, при $h = 2$ получим:

$$f(y|\theta) = \sum_{s_1=1}^{k_1} p^{(1)}_{s_1} \sum_{s_2=1}^{k_2} p^{(2)}_{s_2} \varphi^{(\mathrm{p})} (\eta^{(1)}_{s_1} + \Lambda^{(1)}_{s_1} \eta^{(2)}_{s_2}, \Psi^{(1)}_{s_1} + \Lambda^{(1)}_{s_1} \Psi^{(2)}_{s_2} \Lambda^{(1)\top}_{s_1}).$$

3. Результаты

Качество работы метода DGMM для решения задачи кластеризации продемонстрировано на рисунке 1 на примере 50 тестовых выборок репозитория UCI. Использована метрика Adjusted Rand Index (ARI), зависящая только от разбиения выборки на кластеры. Видно, что данный метод имеет точность, до 70% превышающую результаты для алгоритмов из аналогичного класса (k-средних и GMM), и всего на 9% в среднем хуже по сравнению с ситуацией обучения с учителем.

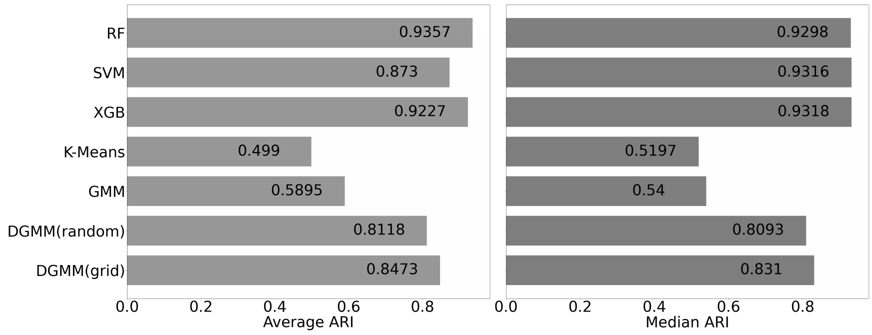

Рис. 1. Сравнение DGMM с другими алгоритмами

Также показано на основе анализа ряда тестовых выборок, что для регрессионных задач использование подобных моделей позволяет добиться точности, сопоставимой с GMM в терминах метрики RMSE и превзойти показатели для линейной регрессии на 22%. Развитие регрессионного варианта DGMM представляет направление дальнейших исследований.

Исследование выполнено при поддержке Междисциплинарной научно-образовательной школы Московского университета «Мозг, когнитивные системы, искусственный интеллект».

Список литературы

[1] Viroli C., McLachlan G.J., “Deep Gaussian Mixture Models”, *Statistics and Computing*, **29**:1 (2006), 43–51.

[2] Baek J., McLachlan J.G., Flack L., “Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**:7 (2010), 1298–1309.

[3] Wu X., Kumar V., Quinlan J., et al., “Top 10 algorithms in data mining”, *Knowledge and Information Systems*, **14**:1 (2008), 1–37.

[4] Горшенин А.К., Королев В.Ю., Турсунбаев А. М., “Медианные модификации EM- и SEM-алгоритмов для разделения смесей вероятностных распределений и их применение к декомпозиции волатильности финансовых временных рядов”, *Информатика и ее применения*, **2**:4 (2008), 12–47.

[5] Горшенин А.К., “О сходимости последовательности SEM-оценок в задаче статистического разделения смесей”, *Вестник Тверского государственного университета. Серия: Прикладная математика*, **4**:23 (2011), 39–49.

On Deep Gaussian Mixture Models in Machine Learning Problems

Ibragimova A.R., Gorshenin A.K.

The work is concentrated on the study of deep neural network architectures with implementation of mixtures of normal distributions in hidden layers for solving clustering and regression problems. The model with different sets of hyperparameters was compared to classical methods: k-means, linear regression, Gaussian mixture models (GMM), etc.

Keywords: deep neural networks, mixtures of normal distributions, EM algorithm.

References

[1] Viroli C., McLachlan G.J., "Deep Gaussian Mixture Models", *Statistics and Computing*, **29**:1 (2006), 43–51.

[2] Baek J., McLachlan J.G., Flack L., "Mixtures of factor analyzers with common factor loadings: applications to the clustering and visualization of high-dimensional data", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **32**:7 (2010), 1298–1309.

[3] Wu X., Kumar V., Quinlan J., et al., "Top 10 algorithms in data mining", *Knowledge and Information Systems*, **14**:1 (2008), 1–37.

[4] Gorshenin A.K., Korolev V.Yu., Tursunbayev A.M., "Median modification of EM- and SEM-algorithms for separation ofmixtures of probability distributions and their application to the decomposition of volatility of financial time series", *Inform. Primen.*, **2**:4 (2008), 12–47 (In Russian).

[5] Gorshenin A.K., "On convergence of SEM estimates sequence for statistical separation of mixture distribution", *Herald of Tver State University. Ser. Appl. Math.*, **4**:23 (2011), 39–49 (In Russian).