Рандомизированный алгоритм классификации в условиях малого количества примеров.

Бояров А.А., a.boiarov@spbu.ru.

Аннотация

Одним из основных ограничений, мешающих эффективно использовать алгоритмы машинного обучения, является необходимость наличия большого количества данных в обучающей выборке. Одним из возможных решений этой проблемы является использование рандомизированных методов, способных обучаться всего по нескольким примерам. В статье рассмотрен подход, основанный на использовании рандомизированного алгоритма стохастической аппроксимации для кластеризации и метода SPS. В такой парадигме неопределённостей, которые возникают при классификации всего по нескольким примерам представляется эффективным использование итеративных рандомизированных подходов для оценивания центроидов классов.

Введение

Последние успехи в распознавании образов на изображениях во многом связаны с парадигмой обучения с учителем. Для такого успешного обучения необходимо очень большое количество размеченных данных. Однако, в реальных задачах такие данные далеко не всегда есть в наличии. Перспективным решением этой проблемы представляется обучение по нескольким примерам на класс. Одним из самых успешных методов такого обучения является метод Prototypical Networks [1]. Однако, данный метод имеет ряд существенных недостатков, таких как скорость работы и неустойчивость к неопределённостям на входе.

Рандомизированный алгоритм классификации в условиях малого количества примеров

В качестве метода принятия решения в Prototypical Networks будем рассматривать рандомизированный алгоритм стохастической аппроксимации (РАСА) для кластеризации, описанный в [3].

Пусть определены натуральное число $k > 1$, множество входных данных $X = \{\mathbf{x}^1, \mathbf{x}^2, \ldots\}$, являющееся подмножеством Евклидова пространства R^d, и заданное на X вероятностное распределение $\mathrm{P}(X)$. Обозначим через $1, \ldots, k$ множество индексов $\{1, 2, \ldots, k\}$. Будем считать, что множество входных данных X разбивается на k неизвестных подмножеств $\{\mathbf{X}_1^\star, \ldots, \mathbf{X}_k^\star\} : X = \cup_{i \in 1..k} \mathbf{X}_i^\star$ таким образом, что вероятностное распределение $\mathrm{P}(X)$ можно представить с помощью смеси распределений: $\mathrm{P}(X) = \sum_{i=1}^k p_i \mathrm{P}(\mathbf{X}_i^\star)$, где p_i $(p_i > 0)$ и $\mathrm{P}(\mathbf{X}_i^\star)$, $i \in 1, \ldots, k$, — соответствующие вероятности и распределения. Пусть векторы $\theta_i,\ i \in 1, \ldots, k$, — *центры кластеров* или *центроиды*, а матрицы $\Gamma_i,\ i \in 1, \ldots, k$, — *ковариационные матрицы*, тогда функционал качества кластеризации принимает вид

$$F(\Theta, \Gamma) = \sum_{i=1}^{k} \sum_{\mathbf{x}^j \in \mathbf{X}_i} (\mathbf{x}^j - \theta_i)^{\mathrm{T}} \Gamma_i^{-1} (\mathbf{x}^j - \theta_i) \to \min_{\Theta, \Gamma}, \qquad j \in 1, \ldots, n. \quad (1)$$

Согласно [3] алгоритм для нахождения центроидов в решения этой задаче имеет вид:

$$\{\, \mathbf{y}_{\pm}^{n} = \mathbf{y}^n(\widehat{\Theta}^{n-1} \pm \beta^n \Delta^n \mathbf{j}^{n\mathrm{T}}, \widehat{\Gamma}^{n-1}),$$

$$\widehat{\Theta}^n = \widehat{\Theta}^{n-1} - \mathbf{j}^{n\mathrm{T}} \alpha^n \frac{\mathbf{y}_+^n - \mathbf{y}_-^n}{2\beta^n} \Delta^n \mathbf{j}^{n\mathrm{T}},$$

Также перспективным является использование метода SPS [4] для нахождения доверительного интервала для сэмплирования центроида каждого класса.

Эксперименты были проведены на базе Omniglot [2]. В результате метод Prototypical Networks показывает результат 0.73 точности, PACA для кластеризации — 0.75, SPS — 0.74. Кроме того, PACA обладает высокой скоростью работы и устойчивостью ко внешним возмущениям.

Литература

[1] Snell J., Swersky K., Zemel R. Prototypical networks for few-shot learning //Advances in Neural Information Processing Systems. – 2017. – С. 4077-4087.

[2] Lake B. M., Salakhutdinov R., Tenenbaum J. B. Human-level concept learning through probabilistic program induction //Science. – 2015. – Т. 350. – №. 6266. – С. 1332-1338.

[3] Boiarov A., Granichin O., Wenguang H. Simultaneous perturbation stochastic approximation for clustering of a Gaussian mixture model under unknown but bounded disturbances //2017 IEEE Conference on Control Technology and Applications (CCTA). – IEEE, 2017. – С. 1740-1745.

[4] Csáji B. C., Campi M. C., Weyer E. Sign-Perturbed Sums: A new system identification approach for constructing exact non-asymptotic confidence regions in linear regression models //IEEE Transactions on Signal Processing. – 2015. – Т. 63. – №. 1. – С. 169-181.