Обучение экспертов для задачи прогнозирования со многими доменами

Н. А. Линдеманн, А. В. Грабовой
lindemann.na@phystech.edu; andriy.graboviy@mail.ru

Рассматривается задача аппроксимации выборки со многими доменами единой мультимоделью – смесью экспертов. Каждый домен аппроксимируется локальной моделью. В работе рассматривается двухэтапная задача оптимизации на основе EM-алгоритма. Используется выборка отзывов сайта Amazon для разных типов товара, которая содержит в себе несколько доменов. В качестве эксперта используется линейная модель, а в качестве признакового описания отзывов используются tf-idf вектора внутри каждого домена.

Ключевые слова: *Смесь экспертов, обучение экспертов, классификация текстов.*

1 Введение

На текущий момент в машинном обучении появляется все больше задач связанных с данными, которые взяты с разных источников. Часто появляются выборки, которые состоят из большого числа доменов. Под *доменом* понимается подмножество объектов выборки, которые обладают некоторыми одинаковыми признаками.

В работе рассматривается задача аппроксимации выборки со многими доменами смесью экспертов. Рассматривается задача бинарной классификации текстов, используется дополнительная информация о доменах. Прдполагается, что это позволит строить более простые и интерпретируемые модели.

Метод решения задачи состоит в построении мультимодели, являющейся смесью локальных моделей. Каждый домен аппроксимируется локальной линейной моделью, смесь которых является итоговым классификатором. Задача обучения модели сводится к двухэтапной задаче оптимизации на основе EM-алгоритма. Схема способа построения смеси экспертов представлена на рис. 1.

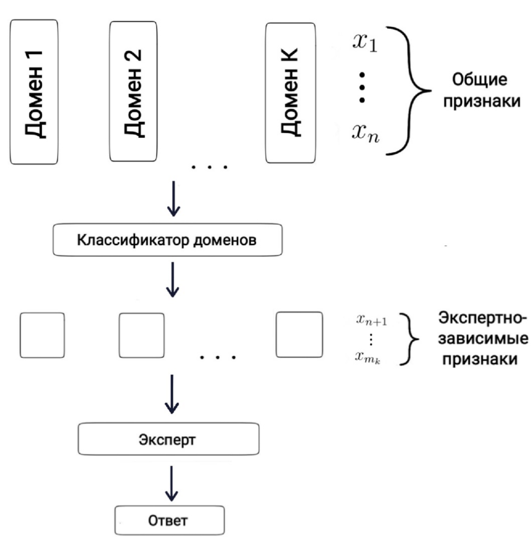

Рис. 1 Способ построения смеси экспертов.

Используется выборка отзывов сайта Amazon для разных типов товаров, которая содержит в себе несколько доменов. Каждый объект имеет экспертно-зависимое описание, которое определяется его пренадлежностью к тому или иному домену. В качестве признакового описания отзывов используется tf-idf вектора внутри каждого домена.

2 Постановка задачи

2.1 Постановка задачи обучнения одного эксперта

Задача бинарной классификации является задачей апроксимации целевой функции

$$\mathbf{f} : \mathbb{R}^n \to \{-1, +1\},$$

где $\mathbb{R}^n$ – пространство признакового описания объектов, а $\{-1, +1\}$ – метка класса объекта. Задачей локальной модели является апроксимация функции $\mathbf{f}$ на некотором домене. На основе общих признаков $(x_1, \ldots, x_n)$ эксперт генерирует экспертно-зависимые признаки $(x_{n+1}, \ldots, x_{m_k})$, количество которых зависит от конкретного домена, и с помощью признаков $(x_1, \ldots, x_n, x_{n+1}, \ldots, x_{m_k})$ локальная модель делает предсказание о принадлежности объекта к одному из двух классов.

В качестве локальной модели будем использовать логистическую регрессию, которая будет предсказывать вероятность того, что объект с признаковым описанием $\boldsymbol{x_i}$ принадлежит классу y_i:

$$p\left(y = y_i \mid \boldsymbol{x_i}, \mathbf{w}\right) = \sigma(y_i \mathbf{w} \cdot \boldsymbol{x_i}).$$

Рассмотрим правдоподобие выборки, а именно, вероятность наблюдать данный вектор $\boldsymbol{y}$ у домена $\mathbf{C}$ (выборка размера N). В предположении, что объекты выборки внутри одного домена независимы и из одного распределения, получим:

$$p\left(\boldsymbol{y} \mid \mathbf{C}, \mathbf{w}\right) = \prod_{i=1}^{N} p\left(y = y_i \mid \boldsymbol{x_i}, \mathbf{w}\right).$$

Далее рассмотрим логарифм правдоподобия:

$$\log p\left(\boldsymbol{y} \mid \mathbf{C}, \mathbf{w}\right) = \log \prod_{i=1}^{N} \sigma(y_i \mathbf{w} \cdot \boldsymbol{x_i}) = \sum_{i=1}^{N} \log \frac{1}{1 + e^{-y_i \mathbf{w} \cdot \boldsymbol{x_i}}} = -\sum_{i=1}^{N} \log(1 + e^{-y_i \mathbf{w} \cdot \boldsymbol{x_i}}).$$

Значит, в даном случае принцип максимального правдоподобия приводит к минимизации логистической функции потерь по всем объектам из данного домена:

$$\mathcal{L}(\mathbf{C}, \boldsymbol{y}, \mathbf{w}) = \sum_{i=1}^{N} \log(1 + e^{-y_i \mathbf{w} \cdot \boldsymbol{x_i}}) \to \min_{\mathbf{w}}.$$

2.2 Постановка задачи построения смеси экспертов

Обобщим подход аппроксимации одного домена на случай, когда в данных присутствует несколько доменов. Пусть всего имеется K доменов в выборке, тогда всю выборку $\mathbf{C}$ можно представить в виде:

$$\mathbf{C} = \bigsqcup_{k=1}^{K} \mathbf{C}'_k,$$

где $\mathbf{C}'_k$ множество объектов, принадлежащих k-му домену. Множеству объектов из домена $\mathbf{C}'_k \subset \mathbf{C}$ соответствует задача линейной регрессии для выборки $\mathbf{X}'_k \subset \mathbf{X}, \mathbf{y}'_k \subset \mathbf{y}$. Модель $\mathbf{g}_k$ аппроксимирующая выборку $\mathbf{X}'_k, \mathbf{y}'_k$ является локальной моделью для выборки $\mathbf{X}$, $\mathbf{y}$.

Определение 1. Модель $\mathbf{g}$ называется локальной моделью для выборки $\mathbf{U}$, если $\mathbf{g}$ аппроксимирует некоторое не пустое подмножество $\mathbf{U}' \subset \mathbf{U}$.

Определение 2. Мультимодель $\mathbf{f}$ называется смесью экспертов, если:

$$\mathbf{f} = \sum_{k=1}^{K} \pi_k \mathbf{g}_k\bigl(\mathbf{w}_k\bigr), \qquad \pi_k\bigl(\mathbf{x}, \mathbf{V}\bigr) : \mathbb{R}^{n\times|\mathbf{V}|} \to [0, 1], \qquad \sum_{k=1}^{K} \pi_k\bigl(\mathbf{x}, \mathbf{V}\bigr) = 1,$$

где $\mathbf{g}_k$ является k-й локальной моделью, π_k — шлюзовая функция, вектор $\mathbf{w}_k$ является параметрами k-й локальной моделью, а $\mathbf{V}$ — параметры шлюзовой функции.

Пусть $\mathbf{w}_k$ является случайным вектором, который задается плотностью распределения $p^k\bigl(\mathbf{w}_k\bigr)$. Получим совместное распределения параметров локальных моделей и вектора ответов:

$$p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr) = \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k\bigl(y_i|\mathbf{w}_k, \mathbf{x}_i\bigr) \right),$$

где $\mathbf{W} = \bigl\{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\bigr\}$. Оптимальные параметры находятся при помощи максимизации правдоподобия:

$$\hat{\mathbf{V}}, \hat{\mathbf{W}} = \arg\max_{\mathbf{V},\mathbf{W}} p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr).$$

3 Вычислительный эксперимент

3.1 Анализ данных

Для проведения первого вычислительного эксперимента из всех отзывов с сайта Amazon были выбраны пять разных доменов: music, baby, kitchen_&_housewares, software, books. Выбор именно этих доменов был обусловлен тем, что получившаяся разнородная подвыборка содержала 9815 и была гиперсбалансирована:

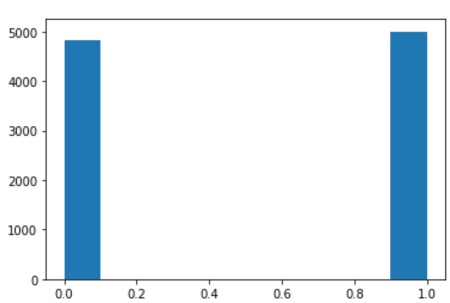

Рис. 2 Распределение подвыборки по классам.

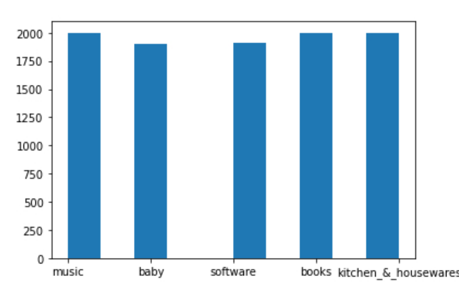

Рис. 3 Распределение подвыборки по доменам.

Далее полученная выборка была разделена на тестовую и обучающую в пропорции 30 : 70. После этого тексты отзывов были преобразованы к формату tf-idf размерности 25770.

Основными метриками, по которым мы будем оценивать качество работы модели, будут

$$\text{Accuracy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}, \quad \text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}, \quad \text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}.$$

3.2 Эксперимент с одной моделью

После подготовки данных была обучена одна модель. Эта модель представляла собой логистическую регрессию с логистической функцией потерь. Результаты обучения модели представлены на рисунках

	precision	recall	f1-score
0	0.81	0.77	0.79
1	0.79	0.83	0.81
accuracy			0.80

Рис. 4 Метрики обученной модели на тестовых данных.

	precision	recall	f1-score
0	0.81	0.77	0.79
1	0.79	0.83	0.81
accuracy			0.80

Рис. 5 Метрики обученной модели на обучающих данных.

3.3 Эксперимент с мультимоделью

4 Анализ задачи

По условию задачи:

$$p(\mathbf{X}, \mathbf{y}, \mathbf{w}|\mathbf{A}) = \prod_i N(\mathbf{x}_i|\mathbf{0}, \sigma^2\mathbf{I}_n) N(\mathbf{w}|\mathbf{0}, \mathbf{A}^{-1}) \prod_j p(y_j|\mathbf{x}_j, \mathbf{w}), \tag{1.1}$$

где $p(y_j = 1|\mathbf{x}_j, \mathbf{w}) = \frac{1}{1+\exp(-\mathbf{w}^\mathsf{T}\mathbf{x}_j)}$

Для простоты запишем (1.1) в следующем общем виде:

$$p(\mathbf{X}, \mathbf{y}, \mathbf{w}|\mathbf{A}) = p(\mathbf{X}) p(\mathbf{w}|\mathbf{A}) p(\mathbf{y}|\mathbf{X}, \mathbf{w}). \tag{1.2}$$

По формуле Байеса:

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}, \mathbf{A}) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\mathbf{A})}{\int_{\mathbf{w}' \in \mathbb{R}^n} p(\mathbf{y}|\mathbf{X}, \mathbf{w}') p(\mathbf{w}'|\mathbf{A}) d\mathbf{w}'} = \frac{\mathcal{Q}(\mathbf{w})}{\int_{\mathbf{w}' \in \mathbb{R}^n} \mathcal{Q}(\mathbf{w}')}, \tag{1.3}$$

где введено обозначение $\mathcal{Q}(\mathbf{w}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\mathbf{A})$.

Выполним аппроксимацию Лапласа:

$$\log\mathcal{Q}(\mathbf{w}) \approx \log\mathcal{Q}(\mathbf{w}_{\text{MAP}}) + \nabla\log\mathcal{Q}(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\mathsf{T} \nabla\nabla^\mathsf{T}\log\mathcal{Q}(\mathbf{w}_{\text{MAP}})(\mathbf{w} - \mathbf{w}_{\text{MAP}}) =$$

$$= \log\mathcal{Q}(\mathbf{w}_{\text{MAP}}) - \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\mathsf{T} \mathbf{H}^{-1} (\mathbf{w} - \mathbf{w}_{\text{MAP}}), \tag{1.4}$$

где введено обозначение $\mathbf{H}^{-1} = -\nabla\nabla^\mathsf{T}\log\mathcal{Q}(\mathbf{w}_{\text{MAP}})$.

Для нашей задачи найдем $\mathbf{H}^{-1}$:

$$\begin{aligned}\mathbf{H}^{-1} &= -\nabla\nabla^{\mathsf{T}}(-\frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w} - \mathbf{1}^{\mathsf{T}}\log(1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w}))) = \\ &= \mathbf{A}^{-1} + \nabla\nabla^{\mathsf{T}}\mathbf{1}^{\mathsf{T}}\log(1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})) = \\ &= \mathbf{A}^{-1} + \sum_{i=1}^{m}\nabla\nabla^{\mathsf{T}}\log(1+\exp(-\mathbf{x}_i^{\mathsf{T}}\mathbf{w})) = \\ &= \mathbf{A}^{-1} + \sum_{i=1}^{m}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\frac{\exp(-\mathbf{x}_i^{\mathsf{T}}\mathbf{w})}{1+\exp(-\mathbf{x}_i^{\mathsf{T}}\mathbf{w})} - \mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\frac{\exp(-2\mathbf{x}_i^{\mathsf{T}}\mathbf{w})}{(1+\exp(-\mathbf{x}_i^{\mathsf{T}}\mathbf{w}))^2}.\end{aligned} \tag{1.5}$$

Тогда получаем:

$$\mathcal{Q}(\mathbf{w}) \approx \mathcal{Q}(\mathbf{w}_{\text{MAP}})\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{H}^{-1}(\mathbf{w}-\mathbf{w}_{\text{MAP}})\right). \tag{1.6}$$

86 Подставляя (1.6) в (1.3) получим:

$$\begin{aligned}p(\mathbf{w}|\mathbf{X},\mathbf{y},\mathbf{A}) &\approx \frac{\mathcal{Q}(\mathbf{w}_{\text{MAP}})\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{H}^{-1}(\mathbf{w}-\mathbf{w}_{\text{MAP}})\right)}{\int_{\mathbf{w}'\in\mathbb{R}^n}\mathcal{Q}(\mathbf{w}_{\text{MAP}})\exp\left(-\frac{1}{2}(\mathbf{w}'-\mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{H}^{-1}(\mathbf{w}'-\mathbf{w}_{\text{MAP}})\right)d\mathbf{w}'} = \\ &= \frac{\exp\left(-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{H}^{-1}(\mathbf{w}-\mathbf{w}_{\text{MAP}})\right)}{\int_{\mathbf{w}'\in\mathbb{R}^n}\exp\left(-\frac{1}{2}(\mathbf{w}'-\mathbf{w}_{\text{MAP}})^{\mathsf{T}}\mathbf{H}^{-1}(\mathbf{w}'-\mathbf{w}_{\text{MAP}})\right)d\mathbf{w}'} = \\ &= N(\mathbf{w}_{\text{MAP}},\mathbf{H}).\end{aligned} \tag{1.7}$$

Оценим $\mathbf{w}_{\text{MAP}}$:

$$\mathbf{w}_{\text{MAP}} = \arg\max_{\mathbf{w}\in\mathbb{R}^n} p(\mathbf{w}|\mathbf{X},\mathbf{y},\mathbf{A}) = \arg\min_{\mathbf{w}\in\mathbb{R}^n}\{-\log p(\mathbf{y}|\mathbf{X},\mathbf{w}) - \log p(\mathbf{w}|\mathbf{A})\}, \tag{1.8}$$

где

$$p(\mathbf{y}|\mathbf{X},\mathbf{w}) = \prod_i \hat{p}_i^{y_i}(1-\hat{p}_i)^{1-y_i}; \quad -\log p(\mathbf{w}|\mathbf{A}) = \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}; \quad \hat{\mathbf{p}} = \frac{1}{1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})}, \tag{1.9}$$

Подставляя (1.9) в (1.8) получаем:

$$\begin{aligned}\mathbf{w}_{\text{MAP}} &= \arg\min_{\mathbf{w}\in\mathbb{R}^n}\{-\mathbf{y}^{\mathsf{T}}\log\hat{\mathbf{p}} - (1-\mathbf{y})^{\mathsf{T}}\log(1-\hat{\mathbf{p}}) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}\} = \\ &= \arg\min_{\mathbf{w}\in\mathbb{R}^n}\{-\mathbf{y}^{\mathsf{T}}\log\frac{1}{1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})} - (\mathbf{1}-\mathbf{y})^{\mathsf{T}}\log\frac{\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})}{1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})} + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}\}\} = \\ &= \arg\min_{\mathbf{w}\in\mathbb{R}^n}\{(\mathbf{1}-\mathbf{y})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{w} + \mathbf{1}^{\mathsf{T}}\log(1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}\},\end{aligned} \tag{1.10}$$

где введя обозначения $\mathcal{L}(\mathbf{w}|\mathbf{X},\mathbf{y},\mathbf{A}) = (\mathbf{1}-\mathbf{y})^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{w} + \mathbf{1}^{\mathsf{T}}\log(1+\exp(-\mathbf{X}^{\mathsf{T}}\mathbf{w})) + \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{w}$ получим следующую оптимизационую задачу для нахождения $\mathbf{w}_{\text{MAP}}$:

$$\mathbf{w}_{\text{MAP}} = \arg\min_{\mathbf{w}\in\mathbb{R}^n}\mathcal{L}(\mathbf{w}|\mathbf{X},\mathbf{y},\mathbf{A}). \tag{1.11}$$

Данная оптимизационная задача решается с помощью градиентного спуска.

5 Заключение

На текущий момент не существует полного теоретического обоснования построения смесей локальных моделей для аппроксимации такого рода выборок.

Литература

[1] *J. Jiang*. A Literature Survey on Domain Adaptation of Statistical Classifiers // ?????, 2007

[2] *А.В. Грабовой, В.В. Стрижов*. Анализ выбора априорного распределения для смеси экспертов // ?????, 2018

[3] *G. Wilson, D.J. Cook*. A Survey of Unsupervised Deep Domain Adaptation // ACM Transactions on Intelligent Systems and Technology, 2020

[4] *M. Wang, W. Deng*. Deep Visual Domain Adaptation: A Survey // Manuscript accepted by Neurocomputing, 2018

[5] *J. Guo, D.J. Shah, R. Barzilay*. Multi-Source Domain Adaptation with Mixture of Experts // Conference on Empirical Methods in Natural Language Processing, 2018

Поступила в редакцию