УДК 316.6

КУЗЬМИНА В. Е., студент группы АУ-114, III курс
Московский государственный технический университет
им. Н.Э. Баумана
г. Москва

СЭМПЛИРОВАНИЕ ВЫБОРКИ ПО НЕУВЕРЕННОСТИ НА ПРИМЕРЕ ТЕКСТОВОГО ДОКУМЕНТА

Пусть X – множество исследуемых объектов, Y – множество классов, $X \times Y$ – вероятностное пространство с плотностью распределения $p(x, y)$. Пусть множество X сформировано из статьи [1] и состоит из 9589 объектов-слов, т.е. X={"агрессия",“в”,“общении”,“медиапользователей”, “анализ”, “особенностей”, “поведения”, “и”, “взаимного”, “влияния”, “статья”, “основана”, “на”, “данных”, “многоаспектного”, “междисциплинарного”, “исследования”, “актуальной”, “проблемы”, “возможности”, “преодоления”, “экстремальных”, “ситуаций”, “в”, “конкретных”, “условиях”, ... “таким”, “образом”, “исследование”, “распространения”, “агрессии”, “и”,“языка”, “вражды”, “в”, “новых”, “медиа”, “важно”, “продолжать”, “на”, “стыке”, “самых”,“разных”, “областей”, “знания”, “включая”, “социологию”, “медиапсихологию”,“коммуникативистику”, культурологию”, математическое”, и”, “имитационное”, “моделирование”, “информационные”, “технологии"}.

Пусть множество Y состоит из 3 классов, относящихся к определенной предметной области: Y_1 – класс «геометрия», Y_2 – «математика», Y_3 – «искусственный интеллект».

Для оценки апостериорных вероятностей класса y для каждого объекта x используем формулу Байеса:

$$\mathrm{P(y|x)} = \frac{\mathrm{P(y)p(x|y)}}{\mathrm{p(x)}} = \frac{\mathrm{P(y)\ p_y(x)}}{\sum_{\mathrm{s \in Y}} \mathrm{p_s(x)P_s}}, \qquad (1)$$

где $p(x)$ – плотность распределения объектов x, которая представляет собой сумму плотности распределения объекта x на классе s, $s \in Y$, умноженную на априорную вероятность класса s, $p(x) = \sum_{s \in Y} p_s(x) P_s$;

$P(y|x)$ – апостериорная вероятность класса y, которая показывает, с какой вероятностью x принадлежит каждому из классов y;

$P(y)$ – априорная вероятность класса, которая отражает долю объектов класса y в выборке;

$p(x|y)$ – функция правдоподобия класса, которая отражает плотность распределения x при известном классе y, $p(x|y) = p_y(x)$

Если известно значение $P(y|x)$, то объект x необходимо относить к тому классу, для которого оценка вероятности выше:

$$\mathrm{a(x)} = \arg\max_{\mathrm{y \in Y}} \mathrm{P(y|x)}. \qquad (2)$$

Также из исходных данных необходимо удалить шум (предлоги, союзы, вводные слова и т.д.). Так как шумы являются нетипичными в контексте выборки объектами, то модель может быть не полностью уверена в их классификации, в то время как для решения основной задачи их классификация не очень полезна. Вокруг шумов плотность распределения мала, и вследствие этого применяется эвристика «взвешивание по плотности», где предпочтение отдается тем объектам, в которых плотность больше [1].

Проранжируем найденные апостериорные вероятности в порядке убывания. Если $p_1(x) \approx p_2(x) \approx p_3(x)$ или $p_1(x) \approx p_2(x) \ll p_3(x)$ то модель не сможет выбрать класс, к которому необходимо отнести x. В этом случае оценку x должен производить оракул. Для того, чтобы снизить количество обращений к оракулу, используются следующие принципы:

1. Принцип наименьшей достоверности:
$$x_i = \arg\min_{x \in X} p_1(x).$$
То есть чем меньше значение $p_1(x)$, тем больше распределение вероятностей похоже на равномерное; следовательно, объект должен попасть в выборку, направляемую для оценки оракулу.
2. Принцип наименьшей разности отступов:
$$x_i = \arg\min_{x \in X}(p_1(x) - p_2(x)).$$
То есть, чем меньше разница между $p_1(x)$ и $p_2(x)$, тем ближе объект находится к границе классов; следовательно, объект должен попасть в выборку, направляемую для оценки оракулу.
3. Принцип максимума энтропии:
$$x_i = \arg\min_{x \in X} \sum_m p_m(x) \ln p_m(x).$$

Оценка производится по энтропии распределения вероятностей, которая будет минимальна при равномерном распределении, т.е. $p_1(x) \approx p_2(x) \approx p_3(x)$.

Покажем формирование выборки по неуверенности на примере. Для каждого неразмеченного объекта рассчитаем дискретное распределение x по классам $P(y|x)$. Для этого оценим априорную вероятность каждого класса $P(y)$, а также плотность распределения x при известном классе y.

Априорную вероятность класса $P(y)$ можно оценить, зная количество запросов в поисковых системах, количество книг, статей, учебников на соответствующую тему («геометрия», «математика», «искусственный интеллект»). Однако в этом случае для нахождения каждого значения $p(x|y)$ для каждого слова необходимо будет проанализировать весь объем вышеперечисленных текстовых материалов. Поэтому для определения априорной вероятности класса $P(y)$ возьмем по одному учебнику-эталону: для класса «геометрия» – [2, 3], для класса «математика» – [4], для класса «искусственный интеллект» – [1].

Тогда $P(y_1)$ =100052/172230=0,58, $P(y_2)$ =64873/172230=0,376, $P(y_3)$ =7305/172230=0,042.

Часть объектов-слов из [3], например, таких, как «агрессия», «общение», «медиапользователь», «статья», «многоаспектное», «междисциплинарное», «актуальный», «преодоление», «экстремальный» не встречаются ни в одном классе-эталоне, — следовательно, их необходимо отправить на оценку оракулу.

Часть объектов-слов из [3] можно отнести к определенному классу, используя формулу Байеса (1). Класс для размеченных объектов указан в столбец «Класс» (см. табл. 1).

Таблица 1. Результаты формирования выборки по неуверенности

x	m_1	m_2	m_3	$P(y_1\|x)$	$P(y_2\|x)$	$P(y_3\|x)$	Sample	Класс
анализ	22	14	8	0,50077	0,31861737	0,18061	FALSE	1
особенность	0	0	5	0	0	1	FALSE	3
поведение	0	4	17	0	0,1917201	0,80828	FALSE	3
взаимное	22	9	0	0,70971	0,29028541	0	FALSE	1
влияние	1	0	0	1	0	0	FALSE	1
основан	2	0	4	0,33516	0	0,66484	FALSE	3
данные	128	11	11	0,85386	0,07336516	0,07278	FALSE	1
исследование	19	4	16	0,48883	0,10289338	0,40827	TRUE	?
проблема	3	0	4	0,43059	0	0,56941	FALSE	3
возможность	13	4	13	0,43489	0,13378753	0,43132	TRUE	?
ситуация	0	0	1	0	0	1	FALSE	3
конкретный	1	0	3	0,25155	0	0,74845	FALSE	3
условия	200	51	2	0,79059	0,20156505	0,00784	FALSE	1

Часть объектов-слов из [1, 3] попадает в выборку по неуверенности. 1 и 3 принципы формирования выборки по неуверенности в рассматриваемом фрагменте не используются, т.к. нет таких объектов, у которых $p_1(x) \approx p_2(x) \approx p_3(x)$. Те объекты, в которых модель сомневается, отправляются в выборку по неуверенности согласно принципу 2. К таким объектам, в частности, относятся слова «исследование» и «возможность», для которых $P(y_1|x) \approx P(y_3|x)$.

В заключение следует отметить, что с помощью вышеприведённого способа можно сформировать выборку по неуверенности из любого набора данных.

Список литературы:

1. *Кажберова, В.В., Чхартишвили, А.Г., Губанов, Д.А., Козицин, И.В., Белявский, Е.В., Федянин, Д.Н., Черкасов, С.Н., Мешков, Д.О.* Агрессия в общении медиапользователей: анализ особенностей поведения и взаимного влияния // Вестник Моск. ун-та. Серия 10. Журналистика. 2023. № 3. С. 26–56. DOI: 10.30547/vestnik.journ.3.2023.2656
2. Системы искусственного интеллекта: учеб. пособие. В 2-х частях./ С. Н. Павлов. — Томск: Эль Контент, 2011. — Ч. 1. — 176 с.
3. *Пылов, П. А.* Основы работы с моделями машинного и глубокого обучения / П. А. Пылов, Р. В. Майтак, А. В. Дягилева. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 256 с. – ISBN 978-5-9729-1547-7. – EDN HSSPQH.
4. *Протодьяконов, А. В.* Асимптотический анализ поведения прикладных моделей машинного обучения : Учебное пособие / А. В. Протодьяконов, А. В. Дягилева, П. А. Пылов. – Вологда : Общество с ограниченной ответственностью "Издательство "Инфра-Инженерия", 2023. – 144 с. – ISBN 978-5-9729-1455-5. – EDN APHQME.