Наивный байесовский классификатор

Скворцова Алина

Октябрь 2020

1 Теоретическая часть

a) Обратимся к формуле Байеса: $P(v_i \in d | d \in c_j) = \frac{P(d \in c_j | v_i \in d) P(v_i \in d)}{P(d \in c_j)}$. Заметим, что $P(d \in c_j | v_i \in d) = \frac{P(d \in c_j) P(v_i \in d)}{P(v_i \in d)} = P(d \in c_j)$. Отсюда имеем, что $P(v_i \in d | d \in c_j) = P(v_i \in d) = p_{ij}$. Если слово не лежит в классе, тогда его вероятность по умолчанию считается равной 0.

b) Из того, что признаки независимы по условию, то $P((k_1, k_2, ..., k_M)|c_j) = P(k_1|c_j)...P(k_M|c_j) = \prod_{i=1}^{M} p_{ij}$, где p_{ij}-вероятность выпадения того или иного слова в классе c_j. Как мы помним, кол-во выпадений слова в документе для модели Бернулли - несущественный параметр, поэтому приведенная нами формула - конечный ответ. Однако если учесть, что в документе нам могут предлагаться не только слова из обучающей выборки, то стоит дополнительно указать, что вероятность неизвестного слова равна не 0, а $\frac{1}{N}$ - используем аддитивное сглаживание(N - кол-во документов в классе).

c) Воспользуемся формулой Байеса: $P(c_j|d) = \frac{P(d|c_j)P(c_j)}{P(d)} = \frac{(\prod_{i=1}^{n} P(w_i|c_j))P(c_j)}{P(d)} = \frac{(\prod_{i=1}^{n} p_{ij})P(c_j)}{P(d)} \approx (\prod_{i=1}^{n} p_{ij})P(c_j)$ К сожалению, жизнь не идеальна и мы можем встретить слова, которые ранее нам не встречались. Тогда их вероятность по умолчанию будет равна 0 и все произведение также обратится в 0. Мы этого конечно не хотим, поэтому запишем формулу с использованием сглаживания: $P(c_j|d) = \prod_{i=1}^{n} \frac{\alpha + N_i}{|V| + N} P(c_j)$, где N_i - кол-во документов, в которых встречалось слово w_i, N - общее кол-во документов класса

d) Чтобы найти класс, которому принадлежит документ d, нам следует найти $argmax(P(c_j) * P(d|c_j))$. Тогда класс, при котором достигается максимум и будет искомым классом. Вероятность ошибки в таком случае равна $(1 - (P(c_j) * P(d|c_j)))$, где c_j - искомый класс.

2 Мультиномиальная модель

a) Вероятность того, что мы не встретим v_i слово ни разу в документе d_j длинны N равна $(1 - r_{ij})^N$. Тогда вероятность встретить v_i хотя бы один

раз можно посчитать как $(1-(1-r_{ij})^N)$.

b) Из того, что признаки независимы по условию, то $P((k_1,k_2,...,k_M)|c_j) = P(k_1|c_j)...P(k_M|c_j) = \prod_{i=1}^{M} r_{ij}$, где p_{ij}-вероятность выпадения того или иного слова в классе c_j. В мультиноминальной модели кол-во выпадений слова имеет значение. Поэтому вероятность того, что слово выпадет именно k раз, также должна учитываться. Тогда конечная формула будет выглядеть как $\prod_{i=1}^{M} r_{ij}^{k_i}$.

c) Воспользуемся формулой Байеса: $P(c_j|d) = \frac{P(d|c_j)P(c_j)}{P(d)}$. Вероятность класса считаем, как и раньше: $P(c_j) = \frac{N_{c_j}}{N}$, где N_{c_j} - кол-во документов в классе, а N - кол-во документов в тренировочной выборке. Также имеем, что $P(d|c_j) = |d|! \prod_{m=1}^{|V|} \frac{p(w_m|c_j)^{N_{im}}}{N_{im}!}$, где N_{im} - кол-во вхождений слова w_m в документ d.

d) Как и в модели Бернулли, чтобы определить класс для нераспределенного документа, нам нужно найти максимум из произведений $\prod P(d_j|c_k)P(c_k)$. Вероятность ошибки можно посчитать как $(1-max)$, max - вероятность $P(c_k|d_j)$.

3 Практическая часть

a) Данные для позитивных отзывов:
Длина минимального отзыва = 70
Длина максимального отзыва = 10363
Длина среднего документа = 1361(с округлением)
Медианная длина = 1076

Данные для негативных отзывов:
Длина минимального отзыва = 52
Длина максимального отзыва = 8969
Длина среднего документа = 1316(с округлением)
Медианная длина = 1065

c) Важно, чтобы тестовая и обучающая выборки обрабатывались одинаково, т.к. это поможет разработчику усовершенствовать формулы, а значить

и усовершенствовать точность вычислений.

d.1) Таблица 30 слов с минимальным наивными байесовскими весами

Слово	Мин.НБВ	Вес в pos	Вес в neg	Абс. частота в pos	Абс. частота в neg
boll	-4.52	$3.85*10^{-7}$	$3.55*10^{-5}$	0	90
uwe	-4.23	$3.85*10^{-7}$	$2.65*10^{-5}$	0	67
thunderbirds	-3.90	$3.85*10^{-7}$	$1.91*10^{-5}$	0	48
dreck	-3.88	$3.85*10^{-7}$	$1.87*10^{-5}$	0	47
seagal	-3.87	$7.70*10^{-7}$	$3.70*10^{-5}$	1	94
dahmer	-3.77	$3.85*10^{-7}$	$1.67*10^{-5}$	0	42
arquette	-3.72	$3.85*10^{-7}$	$1.59*10^{-5}$	0	40
ajay	-3.70	$3.85*10^{-7}$	$1.56*10^{-5}$	0	39
beowulf	-3.67	$3.85*10^{-7}$	$1.52*10^{-5}$	0	38
deathstalker	-3.67	$3.85*10^{-7}$	$1.52*10^{-5}$	0	38
grendel	-3.65	$3.85*10^{-7}$	$1.48*10^{-5}$	0	37
hackenstein	-3.65	$3.85*10^{-7}$	$1.48*10^{-5}$	0	37
unwatchable	-3.59	$7.70*10^{-7}$	$2.80*10^{-5}$	1	72
kareena	-3.56	$3.85*10^{-7}$	$1.36*10^{-5}$	0	34
stinker	-3.47	$7.70*10^{-7}$	$2.49*10^{-5}$	1	63
lordi	-3.47	$3.85*10^{-7}$	$1.24*10^{-5}$	0	31
slater	-3.47	$3.85*10^{-7}$	$1.24*10^{-5}$	0	31
hammerhead	-3.44	$3.85*10^{-7}$	$1.20*10^{-5}$	0	30
wayans	-3.44	$3.85*10^{-7}$	$1.20*10^{-5}$	0	30
ariel	-3.41	$7.70*10^{-7}$	$2.34*10^{-5}$	1	59
yawn	-3.39	$7.70*10^{-7}$	$2.30*10^{-5}$	1	58
blah	-3.39	$1.92*10^{-6}$	$5.73*10^{-5}$	4	146
welch	-3.37	$3.85*10^{-7}$	$1.31*10^{-5}$	0	28
hobgoblins	-3.37	$3.85*10^{-7}$	$1.31*10^{-5}$	0	28
ripley	-3.34	$3.85*10^{-7}$	$1.09*10^{-5}$	0	27
varma	-3.34	$3.85*10^{-7}$	$1.09*10^{-5}$	0	27
turgid	-3.34	$3.85*10^{-7}$	$1.09*10^{-5}$	0	27
bigelow	-3.34	$3.85*10^{-7}$	$1.09*10^{-5}$	0	27
gamera	-3.32	$7.70*10^{-7}$	$2.14*10^{-5}$	1	54
revolting	-3.30	$3.85*10^{-7}$	$1.05*10^{-5}$	0	26

d.2)Таблица 30 слов с максимальным наивными байесовскими весами

Слово	Мин.НБВ	Вес в pos	Вес в neg	Абс. частота в pos	Абс. частота в neg
sabu	3.57	$1.38*10^{-5}$	$3.90*10^{-7}$	35	0
stardust	3.57	$1.38*10^{-5}$	$3.90*10^{-7}$	35	0
kriemhild	3.57	$1.38*10^{-5}$	$3.90*10^{-7}$	35	0
gunga	3.59	$1.42*10^{-5}$	$3.90*10^{-7}$	36	0
anchors	3.59	$1.42*10^{-5}$	$3.90*10^{-7}$	36	0
carface	3.59	$1.42*10^{-5}$	$3.90*10^{-7}$	36	0
clara	3.62	$1.46*10^{-5}$	$3.90*10^{-7}$	37	0
luzhin	3.65	$1.50*10^{-5}$	$3.90*10^{-7}$	38	0
mclaglen	3.67	$1.54*10^{-5}$	$3.90*10^{-7}$	39	0
giovanna	3.67	$1.54*10^{-5}$	$3.90*10^{-7}$	39	0
panahi	3.67	$1.54*10^{-5}$	$3.90*10^{-7}$	39	0
gino	3.67	$1.54*10^{-5}$	$3.90*10^{-7}$	39	0
creasy	3.67	$1.54*10^{-5}$	$3.90*10^{-7}$	39	0
trier	3.70	$1.57*10^{-5}$	$3.90*10^{-7}$	40	0
khouri	3.70	$1.57*10^{-5}$	$3.90*10^{-7}$	40	0
ossessione	3.70	$1.65*10^{-5}$	$3.90*10^{-7}$	42	0
¨	3.70	$1.69*10^{-5}$	$3.90*10^{-7}$	43	0
deathtrap	3.81	$1.77*10^{-5}$	$3.90*10^{-7}$	45	0
din	3.87	$1.88*10^{-5}$	$3.90*10^{-7}$	48	0
biko	3.87	$1.88*10^{-5}$	$3.90*10^{-7}$	48	0
visconti	3.91	$1.96*10^{-5}$	$3.90*10^{-7}$	50	0
sox	4.047	$2.23*10^{-5}$	$3.90*10^{-7}$	57	0
flavia	4.047	$2.23*10^{-5}$	$3.90*10^{-7}$	57	0
kolchak	4.16	$2.50*10^{-5}$	$3.90*10^{-7}$	64	0
corbett	4.19	$2.58*10^{-5}$	$3.90*10^{-7}$	66	0
gundam	4.27	$2.81*10^{-5}$	$3.90*10^{-7}$	72	0
mildred	4.29	$2.85*10^{-5}$	$3.90*10^{-7}$	73	0
edie	4.48	$3.46*10^{-5}$	$3.90*10^{-7}$	89	0
paulie	4.50	$3.54*10^{-5}$	$3.90*10^{-7}$	91	0
antwone	4.57	$3.77*10^{-5}$	$3.90*10^{-7}$	97	0

e) Модель Бернулли

Тест	Время(s)	Точность
trained	5491.36	-
train set	355.39	89.99
dev-b set	0.69	73.30
dev set	211.98	83.88

Мультиноминальная модель

Тест	Время(s)	Точность
trained	359.66	-
train set	376.02	89.70
dev-b set	0.75	72.55
dev set	233.45	83.94

f)

Модель	Унаграммы	Биграммы	Триграммы	n > 3
Мультиноминальная	83.94	83.90	82.99	-
Бернулли	83.88	82.30	79.60	-

Как видно из таблицы, использование словарей, состоящих не из одного слова, не дает увеличения точности (при обучении на заданной выборке отзывов).