Логистическая регрессия
Лекция 6

Елена Тутубалина

Казанский федеральный университет

22 марта 2018 г.

Введение

Логистическая регрессия (или метод максимальной энтропии MaxEnt) относится к семейству логарифмически-линейных классификаторов.

- как и НБ применяется линейная комбинация взвешенных признаков
- НБ - генеративная модель, MaxEnt – дискриминативная.

Генеративная модель

Классификатор НБ:

$$\hat{y} = \operatorname*{argmax}_y P(y|x) = \operatorname*{argmax}_y P(x|y)P(y)$$

- оценивается генерация x из класса y
- правдоподобие признаков x при заданном y

Дискриминативная модель

- вычисляем $P(y|x)$ напрямую
- проводим различия между значениями y
- извлекаем признаки, комбинируем их, применяем функцию

Линейная комбинация 1

$$P(y = C|x) = \sum_{i=1}^{N} w_i f_i = w \cdot f$$

не годится, так как область значений от $-\infty$ до $+\infty$

Линейная комбинация 2

1) Делаем область значений положительной:

$$\exp \sum_i w_i f_i$$

2) Нормализуем

$$p(c|x) = \frac{1}{Z} \exp \sum_i w_i f_i$$

где $Z = \sum_c \exp(\sum_{i=1}^{N} w_i f_i)$

Логистическая регрессия

- считаем признаки бинарными: принимает значение 0 или 1 (индикаторная функция)
- признак – функция не только наблюдения x, но и класса c

$$p(c|x) = \frac{\exp(\sum_{i=1}^{N} w_i f_i(c, x))}{\sum_{c' \in C} \exp(\sum_{i=1}^{N} w_i f_i(c', x))}$$

Признаки в логистической регресии

Задача - анализ тональности отзывов о товарах
Примеры признаков:

$$f_1(x) = \begin{cases} 1, \text{если "удобный"} \in x \,\&\, c = + \\ 0, \text{иначе} \end{cases}$$

$$f_2(x) = \begin{cases} 1, \text{если "топорно"} \in x \,\&\, c = - \\ 0, \text{иначе} \end{cases}$$

Признаки в логистической регреcии 2

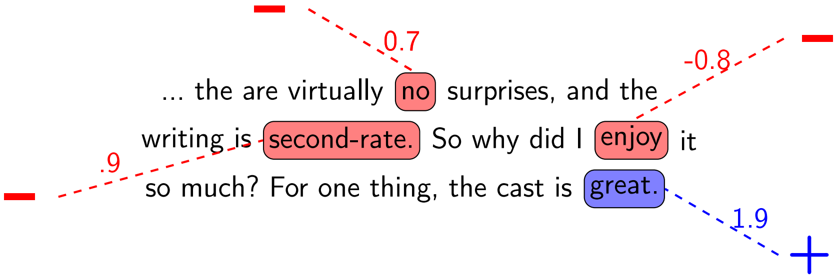
Задача - сегментация текста на предложения
Пример признаков:

- следующее слово в верхнем регистре и $c = \text{EOS}$ (End Of Sentence)
- следующее слово в нижнем регистре и $c = \text{not-EOS}$
- следующее слово примыкает непосредственно к точке и $c = \text{not-EOS}$
- предыдущее слово $\in$ словарю сокращений и $c = \text{not-EOS}$
- еще примеры ?

Программная реализация

- применяются шаблоны признаков
- при считывании обучающего множества по этим шаблонам генерируется индекс признаков
- при создании индекса применяется хэширование
- см. пример в Apache OpenNLP

Пример классификации



Вычисление наиболее вероятного класса

$$\hat{c} = \arg\max_{c \in C} P(c|x)$$

$$= \arg\max_{c \in C} \frac{\exp(\sum_{i=1}^{N} w_i f_i(c, x))}{\sum_{c' \in C} \exp(\sum_{i=1}^{N} w_i f_i(c', x))}$$

$$= \arg\max_{c \in C} (\exp \sum_{i=1}^{N} w_i f_i(c, x))$$

$$= \arg\max_{c \in C} \sum_{i=1}^{N} w_i f_i(c, x)$$

Обучение логистической регрессии

- подбираем веса, которые делают разметку классами в обучающем множестве более вероятной
- условный метод максимального правдоподобия:

$$\hat{w} = \arg\max_{w} \left(\log P(y^{(j)}|x^{(j)}) \right)$$

где $(y^{(j)}|x^{(j)}) \in$ обучающему множеству
Для всего обучающего множества:

$$\hat{w} = \arg\max_{w} \sum_{j} \log P(y^{(j)}|x^{(j)})$$

Целевая функция

$$L(w) = \sum_j \log P(y^{(j)}|x^{(j)})$$

$$= \log \sum_j \frac{\exp(\sum_{i=1}^{N} w_i f_i(y^{(j)}, x^{(j)}))}{\sum_{y' \in Y} \exp(\sum_{i=1}^{N} w_i f_i(y'^{(j)}, x^{(j)}))}$$

$$= \log \sum_j \exp(\sum_{i=1}^{N} w_i f_i(y^{(j)}, x^{(j)})) - \log \sum_j \sum_{y' \in Y} \exp(\sum_{i=1}^{N} w_i f_i(y'^{(j)}, x^{(j)}))$$

$\rightarrow$ задача оптимизации выпуклых функций

Градиентный спуск

$$L'(w) = \sum_j f_k(y^{(j)}, x^{(j)}) - \sum_j \sum_{y' \in Y} P(y'|x^{(j)}) f_k(y'^{(j)}, x^{(j)})$$

Интерпретация:

$$L'(w) = \sum_j \text{наблюдаемая частота } (f_k) \text{ - ожидаемая частота } (f_k)$$

Проблема переобучения

- Если в обучающем множестве есть шумы $\Rightarrow$ веса будут подогнаны с их учетом
- По воле случая может оказаться, что неинформативный признак будет сочетаться только с одним классом.

Выводим коэффициент регуляризации $R(w)$:

$$\hat{w} = \operatorname*{argmax}_{w} \sum_{j} \log P(y^{(j)}|x^{(j)}) - \alpha R(w)$$

его предназначение - штрафовать за слишком большое значение весов

L1 и L2

L1 (евклидово расстояние):

$$R(W) = ||W||_2^2 = \sum_{j=1}^{N} w_j^2$$

L2 (расстояние городских кварталов):

$$R(W) = ||W||_1 = \sum_{i=1}^{N} |w_i|$$

Отбор признаков

- Многие классификаторы не предполагают внедрения коэфициента регуляризации
- Вместо этого применяется процедура отбора признаков: вычислить меру "качества"каждого признака и оставить только лучшие
- Отбор происходит на отладочном множестве

Пример - взаимная информация:

$$MI(w) = -\sum_{i=1}^{C} P(c_i) \log P(c_i)$$
$$+P(w) \sum_{i=1}^{C} P(c_i|w) \log P(c_i|w)$$
$$+P(\bar{w}) \sum_{i=1}^{C} P(c_i|\bar{w}) \log P(c_i|\bar{w})$$

Задание 5

- Скачать корпус https://drive.google.com/open?id=0B4bl7YMqDnViYVo0V2FubFVh
- Для каждой записи оставить только первый тег в качестве класса
- Реализовать классификатор по тегам с использованием модели Логистической регрессии
- Добавить признак кол-во частей речи (существительное, прилагательное, глагол, наречие)
- Извлечь наиболее информативные признаки http://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.ht
- Получить оценки эффективности классификатора на тестовом множестве