Введение в теорию информации. Сжатия тестов.

Дискретный анализ 2012/13

Андрей Калинин, Татьяна Романова

13 апреля 2013 г.

Раздел

Базовые понятия теории вероятностей

Наивный байесовский классификатор

Теория информации

Сжатие текстов

- Модели данных
- Адаптивные модели

Вероятность

Задано пространство элементарных событий (конечное или бесконечное):

$$\Omega = \{\omega_1, \omega_2, \ldots\}$$

и функция $p(\omega_i) \in [0, 1]$, удовлетворяющая условию нормировки:

$$\sum p(\omega_i) = 1.$$

- Значения $p(\omega_i)$ — вероятности элементарных событий ω_i.
- Множества $A \subset \omega$ — события и их вероятность:

$$P(A) = \sum_{\omega_i \in A} p(\omega_i).$$

На практике

- Возможность проведения экспериментов, исходом которых является наступление ω_i (событие A наступает, если наступает $\omega_i \in A$)
- Связь Ω с субэлементарным уровнем, например, если событие n-кратное бросание монеты, то можно выделить связь с однократным бросанием.
- Возможность проведения серии опытов, в результате чего:

$$\frac{N(A)}{N} \to P(A) \qquad \text{при } N \to \infty.$$

Объединение и пересечение событий

$$P(A + B) = P(A) + P(B) - P(AB)$$

Если события не пересекаются, то

$$P(A + B) = P(A) + P(B)$$

Условная вероятность

- $P(B|A)$ — вероятность наступления B при условии наступления в то же время события A, или условная вероятность.
- При этом, из всех $\omega_i \in A$ рассматриваются только $\omega_i \in B$ и $P(B|A)$ равнялось бы $P(AB)$, если бы $P(A) = 1$. Нормируя, получаем:

$$P(B|A) = \frac{P(AB)}{P(A)}.$$

- Или, формула умножения вероятностей:

$$P(AB) = P(A)P(B|A).$$

Формула Байеса

Разбиение Ω на полную группу несовместимых событий $A_1, \ldots, A_n$ позволяет любое событие B записать в виде

$$B = BA_1 + \cdots + BA_n,$$

откуда $P(B) = P(BA_1) + \cdots + P(BA_n)$, и получается формула полной вероятности:

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_n)P(A_n).$$

Пусть $P(A), P(B) > 0$. Тогда из

$$P(AB) = P(A|B)P(B) = P(B|A)P(A)$$

вытекает

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)},$$

что приводит к формуле Байеса:

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_k P(B|A_k)P(A_k)}.$$

Примеры

- Имеются три карточки. На одной — с обоих сторон A, на другой — B. На третьей картонке с одной стороны A, с другой — B. Одна из картонок выбирается наугад и кладётся на стол. Предположим, на видимой стороне картонки оказывается буква A. Какова вероятность того, что на другой стороне — тоже A?

Примеры

- Имеются три карточки. На одной — с обоих сторон A, на другой — B. На третьей картонке с одной стороны A, с другой — B. Одна из картонок выбирается наугад и кладётся на стол. Предположим, на видимой стороне картонки оказывается буква A. Какова вероятность того, что на другой стороне — тоже A?
- Некоторая болезнь имеется у 1 человека на 100. Имеется тест, определяющий наличие болезни с 95%-ой точностью (т.е., в 5% случаев срабатывания теста его результат неверен). Допустим, для тестирования случайно отбирают людей и тест «сработал» на каком-то человеке. Какова вероятность того, что он болен этой болезнью?

Парадокс Монти Холла

- Представьте, что вы стали участником игры, в которой вам нужно выбрать одну из трех дверей. За одной из дверей находится автомобиль, за двумя другими дверями — козы. Вы выбираете одну из дверей, например, номер 1, после этого ведущий, который знает, где находится автомобиль, а где — козы, открывает одну из оставшихся дверей, например, номер 3, за которой находится коза. После этого он спрашивает вас, не желаете ли вы изменить свой выбор и выбрать дверь номер 2. Увеличатся ли ваши шансы выиграть автомобиль, если вы примете предложение ведущего и измените свой выбор?

Независимость

События A и B называют независимыми, если $P(B|A) = P(B)$, т.е.

$$P(AB) = P(A)P(B).$$

Случайные величины

- Числовая функция $X(\omega)$, заданная на Ω представляет собой случайную величину (например, 1 при выпадении герба, 0 — решки).
- Среднее значение $m_x = \mathrm{E}(X)$,

$$\mathrm{E}(X) = \sum_{\omega \in \Omega} X(\omega) P(\omega),$$

  называют математическим ожиданием $X(\omega)$.
- Математическое ожидание линейно:

$$\mathrm{E}(\alpha X + \beta Y) = \alpha \mathrm{E}(X) + \beta \mathrm{E}(Y).$$

Парадокс ожидания серии

- Какая комбинация, 00 или 01, появится раньше в случайной «01»-последовательности?
- Очевидно, равновероятно, т.к. после первого нуля появление 0 или 1 — вероятность 1/2.

Парадокс ожидания серии

- Какая комбинация, 00 или 01, появится раньше в случайной «01»-последовательности?
- Очевидно, равновероятно, т.к. после первого нуля появление 0 или 1 — вероятность 1/2.
- Однако, число шагов (среднее время ожидания) m_{00} и m_{01} — не одинаково:
  - Пусть m_0 — среднее число шагов до 01 при условии, что первая цифра последовательности была 0, а m_1 — то же самое, но для 1. Тогда:

$$m_0 = 1 + \frac{1}{2} + \frac{1}{2}m_0, \qquad m_1 = 1 + \frac{1}{2}m_0 + \frac{1}{2}m_1,$$

    откуда $m_0 = 3$, $m_1 = 5$ и $m_{01} = .5(m_0 + m_1) = 4$.
  - Если же m_0^* и m_1^* аналоги m_0 и m_1 в для 00, то:

$$m_0^* = 1 + \frac{1}{2} + \frac{1}{2}m_1^*, \qquad m_1^* = 1 + \frac{1}{2}m_0^* + \frac{1}{2}m_1^*,$$

    и в конечном итоге это даёт $m_{00} = .5(m_0^* + m_1^*) = 6$.

Петербургский парадокс

- Если герб при неоднократном бросании монеты выпадает первый раз в n-й попытке, участнику игры выплачивается 2^n рублей.
- Математическое ожидание выигрыша,

$$2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + \cdots + 2^n \cdot \frac{1}{2^n} + \cdots = 1 + 1 + \cdots ,$$

  бесконечно.
- Сколько можно заплатить за участие в игре?

Петербургский парадокс

- Если герб при неоднократном бросании монеты выпадает первый раз в n-й попытке, участнику игры выплачивается 2^n рублей.
- Математическое ожидание выигрыша,
$$2 \cdot \frac{1}{2} + 4 \cdot \frac{1}{4} + \cdots + 2^n \cdot \frac{1}{2^n} + \cdots = 1 + 1 + \cdots \ ,$$
бесконечно.
- Сколько можно заплатить за участие в игре?
- Кажется, что сколько угодно — казино проиграет.
- Но это «в среднем», а единичная игра скорее не принесёт большого выигрыша и платить большие деньги не имеет смысла.

Раздел

Задача классификации

- C — некоторый класс, к которому можно отнести наблюдаемый объект. Любой объект относится либо к C, либо к $\neg C$.
- $F_1, \ldots, F_n$ — факторы, некоторые наблюдаемые характеристики объекта.
- Задача — при неких полученных значениях факторов определить принадлежность события к классу, т.е. вычислить:
$$P(C|F_1, \ldots, F_n).$$
- Количесто факторов может быть велико.
- Примеры факторов: слова в тексте, результаты измерений, телеметрические данные и т.п.

Модель

- Из формулы условной вероятности:

$$P(C|F_1,\dots,F_n) = \frac{P(C)P(F_1,\dots,F_n|C)}{P(F_1,\dots,F_n)}$$

- Для оценки вероятности принадлежности к классу C интересено только $P(C)P(F_1,\dots,F_n|C)$, т.к. $P(F_1,\dots,F_n)$ не зависит от класса.
- При этом $P(C)P(F_1,\dots,F_n|C) = P(C,F_1,\dots,F_n)$.
- Тогда:

$$\begin{aligned}
P(C,F_1,\dots,F_n) &= P(C)P(F_1,\dots,F_n|C) = \\
&= P(C)P(F_1,C)P(F_2,\dots,F_n,|C,F_1) = \\
&= P(C)P(F_1,C)P(F_2|C,F_1)\times \\
&\qquad \times P(F_3,\dots,F_n,|C,F_1,F_2) = \\
&= P(C)P(F_1,C)P(F_2|C,F_1)P(F_3|C,F_1,F_2)\times\cdots \\
&\qquad \times P(F_n|C,F_1,F_2,F_3,\dots,F_{n-1}).
\end{aligned}$$

Наивная модель

- Наивное предпололжение: предполагаем независимость факторов друг от друга.
- Т.к. факторы независимы, то для $i \neq j$:

$$P(F_i|C, F_j) = P(F_i|C)$$

- Тогда:

$$\begin{aligned} P(C, F_1, \ldots, F_n) &= P(C)P(F_1|C)P(F_2|C)P(F_3|C)\cdots P(F_n|C) \\ &= P(C)\prod_{i=1}^{n} P(F_i|C). \end{aligned}$$

- И вся модель:

$$P(C|F_1, \ldots, F_n) = \frac{1}{Z} P(C)\prod_{i=1}^{n} P(F_i|C)$$

где Z зависит только от $F_1, \ldots, F_n$, т.е. константа при заданных значениях факторов.

Классификация

- Предварительно нужно оценить вероятности $P(F_i|C)$.
- После чего:

$$\mathrm{classify}(f_1, \ldots, f_n) = \arg\max_c P(C = c) \prod_{i=1}^{n} P(F_i = f_i | C = c).$$

Классификация текстовых документов

- Характеристика — наличие слова в тексте, т.е. нужно предварительно оценить
$$P(w_i|C).$$
- Предполагаем, что слова никак не связаны друг с другом, их вес не зависит от длины документа и т.п. Тогда:
$$P(D|C) = \prod_i P(w_i|C)$$
- Для условной вероятности:
$$P(C|D) = \frac{P(C)}{P(D)} P(D|C).$$

Два класса

- Допустим, есть только два класса, S и $\neg S$.
- Тогда можно оценить принадлежность документа к каждому из классов:

$$P(D|S) = \prod_i P(w_i|S)$$

$$P(D|\neg S) = \prod_i P(w_i|\neg S).$$

- Или:

$$P(S|D) = \frac{P(S)}{P(D)} \prod_i P(w_i|S)$$

$$P(\neg S|D) = \frac{P(\neg S)}{P(D)} \prod_i P(w_i|\neg S)$$

- Взяв отношение двух величин:

$$\frac{P(S|D)}{P(\neg S|D)} = \frac{P(S)}{P(\neg S)} \prod_i \frac{P(w_i|S)}{P(w_i|\neg S)}$$

- Для удобства расчётов можно перейти к логарифмам:

$$\ln \frac{P(S|D)}{P(\neg S|D)} = \ln \frac{P(S)}{P(\neg S)} + \sum_i \ln \frac{P(w_i|S)}{P(w_i|\neg S)}$$

- Документ относится к категории S если $P(S|D) > P(\neg S|D)$, т.е.

$$\ln \frac{P(S|D)}{P(\neg S|D)} > 0.$$

Раздел

Интуитивное понятние информации

- Сколько информации содержится в строке текста? На сколько ее можно сжать без потери смысла сообщения?
- Есть 10 букв и алфавит из 33 символов, сколько различных строк можно представить?
- Допустим, мы можем передать Σ^N различных сообщений. Только n из них будут иметь смысл. для того, чтобы указать, какое из n сообщений мы передаем, достаточно $I = \log_2 n$ бит.
- С удвоением строки, количество возможных сообщений растет квадратично.
- Самое лучшее соотношение количества информации к длине сообщения у случайных строк. Для передачи 2^N случайных равновероятных сообщения требуется ровно N бит, меньше — невозможно, больше — избыточно.

Энтропия

- Пусть у нас есть язык, алфавит которого состоит из k символов: $a_1, a_2, ...a_k$, и для каждого символа известна вероятность его появления в тексте $p_1, p_2, ..., p_k$.
- Мы хотим определить количество информации в сообщении из данного языка.
- Если сообщение имеет длину N, то симолов a_1 в нем будет примерно Np_1, a_2 — Np_2 и т. п. Посчитаем, сколько различных сообщений такого вида (с таким распределением вероятностей) может быть:
$$\frac{N!}{(Np_1)!(Np_2)!...(Np_k)!}$$
- Чтобы узнать, сколько бит нужно, для передачи одного из таких сообщений, нужно взять логарифм по основанию 2. Это же число можно назвать средним количеством информации в сообщении или энтропией:
$$I(s) = H(s) = N\sum_{i=1}^{k} -p_i * log_2(p_i)$$

Энтропия

- Чтобы узнать, сколько бит нужно, для передачи одного из таких сообщений, нужно взять логарифм по основанию 2. Это же число можно назвать средним количеством информации в сообщении или энтропией:

$$I(s) = H(s) = N \sum_{i=1}^{k} -p_i * log_2(p_i)$$

- Часто энтропией называют среднее количество информации на символ. Информационная двоичная энтропия для независимых случайных событий x с k возможными состояниями рассчитывается по формуле:

$$H(x) = -\sum_{i=1}^{k} p_i * log_2 p_i$$

- Условная энтропия n-го порядка — энтропия для алфавита, в котором известны вероятности появления буквы после n других. Условная энтропия всегда меньше безусловно, например, для русского языка

Раздел

Базовые понятия теории вероятностей

Наивный байесовский классификатор

Теория информации

Сжатие текстов

- Модели данных
- Адаптивные модели

Введение в проблему

- Компрессия текстов использовалась до компьютеров: азбука Морзе, коды Брайля.
- Нужна для уменьшения объёмов хранения и передачи данных.
- Часть более общей задачи компрессии данных: отличие в возможности точного восстановления информации (для некоторых данных это не всегда возможно и не всегда требуется).

Развитие

- **1950-е** коды Хаффмана.
- **1970-е** Зив-Лемпель (словарные методы) и арифметическое кодирование (символьные методы): адаптивное сжатие.
- **1980-е** PPM (prediction by partial matching, предсказание по частичному совпадению).
- **1990-е** преобразование Барроуза-Уиллера, сжатие блоков данных; улучшение вычислительных характеристик алгоритмов.
  - На текущий момент: два бита на символ для среднего английского текста.
  - Предположительно один бит на символ недостижим.

Классификация методов компрессии текстов

- Символьные (статистические): кодирование одного символа как можно меньшей последовательностью бит.
  - Коды Хаффмана.
  - Арифметическое кодирование.

  Основное отличие между методами в том, как вычисляются вероятности появления следующих символов — построение модели текста.
- Словарные: несколько часто встречающихся символов кодируются меньшей последовательностью бит, указывающих на соответствующий вход в словарь. В основном это варианты алгоритма Зива-Лемпеля.

Раздел

Модель данных

- Задача модели — предсказывать следующие символы в потоке данных.
- Кодер и декодер используют одинаковую модель для упаковки и распаковки текста.

Информация

- Количество бит для символа s — информация, в нём содержащаяся. $I(s) = -\log P(s)$.
- Средняя информация по всему алфавиту:

$$H = \sum_s P(s)I(s) = \sum_s -P(s)\log P(s)$$

- Энтропия — нижний предел для алгоритмов компрессии данных, измеряемый в битах на символ.
- Коды Хаффмана часто оказываются близкими к этому пределу, однако в некоторых ситуациях очень неэффективны (в случаях $P(s) \to 1$ и $I(s) \to 0$).
- Арифметический кодер свободен от этих недостатков, но более медленный.

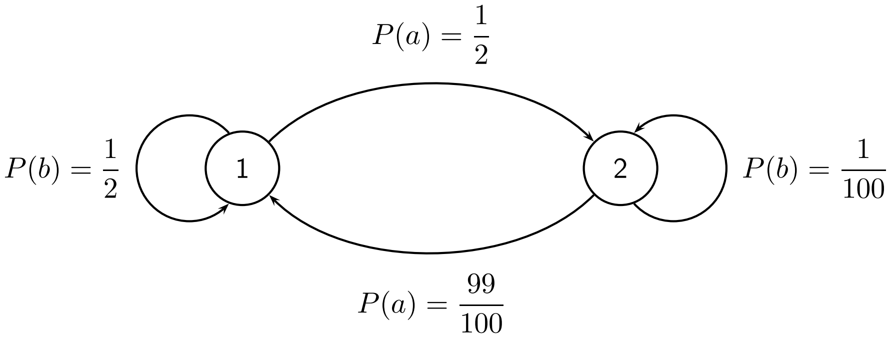
Использование моделей

- Кодер и декодер должны использовать одну и ту же модель.
- Кодер не может «подсматривать» вперёд для оценки вероятностей (потому что этого не может сделать декодер).
- Символы с $P(s) = 0$ не могут быть закодированы ($I(s) = \infty$).
- Символы с $P(s) = 1$ не нуждаются в сжатии ($I(s) = 0$).
- Лучшая модель та, которая предсказывает настоящие символы из потока с максимальной вероятностью.

Контекстно ограниченная модель

- Вероятность появления английской буквы u в английском тексте 2%, т.е. $I(u) \simeq 5{,}6$ битов.
- Однако, после появления буквы q вероятность того, что следующий символ будет u 95%, т.е. $P(u|q) \simeq .95$ и $I(u|q) \simeq 0{,}074$ бит. ($quality$, $quite$, $quiet$, …).
- В этих случаях кодирование исключений будет требовать больше бит (слова $Iraq$, $Qantas$), но в среднем объём сжатых текстов уменьшится.
- Использование предыдущих m символов для предсказания — контекстно ограниченная модель порядка m.

Модель конечного автомата



- В состояниях содержится информация о распределении символов.
- Такая модель непредставима контекстно ограниченной моделью (отслеживает чётное или нечётное количество символов).

Вероятностное распределение символов

- Кодер должен закодировать обрабатываемый символ и только после этого изменить распределение.
- Декодер раскодирует символ в соответствии с текущим распределением, после чего делает аналогичное изменение.
- Предполагается безошибочная передача данных между кодером и декодером.
- Если известна формальная грамматика текста (C, Java), то её можно использовать для моделирования.

Раздел

Оценка вероятностей

- Статическое моделирование: во время разработки модели назначается распределение и используется для всех входных текстов.
- Полустатическое моделирование: создание распределения вероятностей на основе входного файла. Требуется два прохода по файлу (сбор статистики и сжатие) а так же явная передача данных модели от кодера к декодеру.
- Адаптивное моделирование: измененение распределения вероятностей во время сжатия входных данных.

Пример

> *"I never heerd a silful old married feller of twenty years' standing pipe "my wife" in a more used note than 'a did,", said Jacob Smallbury. It migh*

Для модели нулевого порядка:

- $P(t) = 49{,}983/768{,}078 = 6.5\%$.
- $P(e) = 9.4\%$.
- $P(x) = 0.11\%$.

Так как следующая буква t, декодер закодирует её $-\log 0.065 = 3.94$ битами.

Символы с нулевой частотой

Необходимо избегать ситуаций, при которых некоторые символы оцениваются с нулевой вероятностью.

- В примере ни разу не встретилась Z, так что простая оценка $P(Z) = 0$ и эта буква не сможет быть закодирована. Строка MighZ довольно необычна, но не невероятна.

Символы с нулевой частотой

Необходимо избегать ситуаций, при которых некоторые символы оцениваются с нулевой вероятностью.

- В примере ни разу не встретилась Z, так что простая оценка $P(Z) = 0$ и эта буква не сможет быть закодирована. Строка MighZ довольно необычна, но не невероятна.
- Один из способов решения проблемы — назначение всем символам, которые до сих пор не встречались, суммарной единичной частоты:
  1. В примере 82 символа встретились, 46 ASCII-символов не встречались не разу. Тем самым, они суммарно должны разделить между собой $1/768{,}079$.
  2. Т.е., $P(Z) = 1/(46 \times 768{,}079) = 1/35{,}331{,}634$, что соответствует 25.07 битам.

Символы с нулевой частотой

Необходимо избегать ситуаций, при которых некоторые символы оцениваются с нулевой вероятностью.

- В примере ни разу не встретилась Z, так что простая оценка $P(Z) = 0$ и эта буква не сможет быть закодирована. Строка MighZ довольно необычна, но не невероятна.
- Один из способов решения проблемы — назначение всем символам, которые до сих пор не встречались, суммарной единичной частоты:
    1. В примере 82 символа встретились, 46 ASCII-символов не встречались не разу. Тем самым, они суммарно должны разделить между собой $1/768{,}079$.
    2. Т.е., $P(Z) = 1/(46 \times 768{,}079) = 1/35{,}331{,}634$, что соответствует 25.07 битам.
- Другой споосб — назначение всем символам изначальной единичной частоты. Тогда $P(Z) = 1/768{,}206$, соответствует 19.6 битам.

Символы с нулевой частотой

Необходимо избегать ситуаций, при которых некоторые символы оцениваются с нулевой вероятностью.

- В примере ни разу не встретилась Z, так что простая оценка $P(Z) = 0$ и эта буква не сможет быть закодирована. Строка MighZ довольно необычна, но не невероятна.
- Один из способов решения проблемы — назначение всем символам, которые до сих пор не встречались, суммарной единичной частоты:
  1. В примере 82 символа встретились, 46 ASCII-символов не встречались не разу. Тем самым, они суммарно должны разделить между собой $1/768{,}079$.
  2. Т.е., $P(Z) = 1/(46 \times 768{,}079) = 1/35{,}331{,}634$, что соответствует 25.07 битам.
- Другой споосб — назначение всем символам изначальной единичной частоты. Тогда $P(Z) = 1/768{,}206$, соответствует 19.6 битам.
- Выбор способа борьбы с проблемой существенно влияет на эффективность сжатия коротких текстов и сглаживается для больших.

Адаптивные модели высоких порядков

"I never heerd a silful old married feller of twenty years' standing pipe "my wife" in a more used note than 'a did,", said Jacob Smallbury. It migh

- Модель первого порядка:
  1. $c[h] = 37{,}525$, $c[t|h] = 1{,}133$.
  2. Опуская проблему символов с нулевой частотой, $P(t|h) = 1{,}133/37{,}525 = 3.02\%$,
  3. Будет закодировано 5.05 битами (хуже, чем модель нулевого порядка для этого случая, т.к. за h чаще следует e).
- Модель второго порядка:
  1. $c[gh] = 1{,}754$, $c[t|gh] = 1{,}129$.
  2. $P(t|gh) = 64.4\%$.
  3. Будет использован код длиною 0.636 бит.

Изменение структуры модели

- В адаптивной модели может меняться не только распределение частот, но и структура модели.
- Например, в модели конечного автомата могут добавляться новые состояния и генерироваться новые переходы в эти состояния.
- Обычно структура менятся в тех местах модели, которые часто используются.

Выводы

- Адаптивная модель позволяет добиться хорошего качества сжатия.
- Однако, с её использованием нельзя получить доступ к случайному месту внутри сжатого файла: нужно пройти весь путь от начала сжатого файла до этого места, чтобы получить нужную модель.