Санкт-Петербургский государственный политехнический университет

Лабораторная работа

по курсу «Стохастические модели»

«Проверка гипотезы о присутствии закодированного сообщения в наборе кодов»

Студент: Руцкий В. В.
Группа: 5057/2
Преподаватель: Иванков А. А.

Санкт-Петербург 2011

1 Постановка задачи

Дан упорядоченный набор кодов:

$$M' = (s'_0, \ldots, s'_n), \quad s'_i \in \Sigma', \quad \Sigma' = \{c'_0, \ldots, c'_k\},$$

где c'_i — коды из алфавита Σ'. Необходимо проверить гипотезу о том, что в сообщении закодировано сообщение на английском языке

$$M = (s_0, \ldots, s_n), \quad s_i \in \Sigma, \quad \Sigma = \{c_0, \ldots, c_k\},$$

где c_i — буквы английского алфавита и разделители, при условии, что кодирование было произведено переобозначением с помощью некоторой биекции $g\colon \Sigma \to \Sigma'$ исходных символов кодами:

$$M' = g(M) \stackrel{\text{def}}{=} (g(s_0), \ldots, g(s_n)).$$

2 Решение

Основная идея решения поставленной задачи состоит в следующем:

1. переберём все биекции
$$g_j \in G = \{g\colon \Sigma \to \Sigma'\},$$
2. для каждой g_j «декодируем» сообщение:
$$M^*_j = g^{-1}(M') = (s^*_{j0}, \ldots, s^*_{jn}),$$
3. проверим насколько полученное сообщение M^*_j «похоже» на английский текст.

Если найдутся биекции для которых декодированное сообщение похоже на английский язык, то ответ на поставленный вопрос будет положительным, иначе — отрицательным.

Проверка соответствия декодированного сообщения английскому языку проводится исходя из двух грубых моделей текстов на английском языке.

2.1 Модель независимых символов

Текст на английском языке из n символов $M = (s_0, \ldots, s_n)$ можно представить как n наблюдений некоторой дискретной случайной величины ξ_{eng}, принимающая значения из Σ с некоторыми вероятностями $\bar{p}^\circ = (p^\circ_1, \ldots, p^\circ_k), \quad p^\circ_1 + \ldots + p^\circ_k = 1$.

Предположим, что в M' закодировано сообщение на английском языке. Тогда для выбранной биекции g_j каждый символ декодированного сообщения $M^*_j = (s^*_{j0}, \ldots, s^*_{jn})$ — это независимое наблюдение некоторой дискретной случайной величины ξ, принимающей значения из Σ с вероятностями $\bar{p} = (p_1, \ldots, p_k), \quad p_1 + \ldots + p_k = 1$, а проверяемая гипотеза H_0 будет состоять в том, что $\xi = \xi_{\text{eng}}$, т. е. $\bar{p} = \bar{p}^\circ$.

Рассмотрим случайную величину

$$\nu_i = \sum_{t=1}^{n} I(s^*_{jt} = c_i), \quad i = 1, \ldots, k,$$

— частоты различных исходов наблюдений ($\nu_1 + \ldots + \nu_k = n$). Вектор $\bar{\nu} = (\nu_1, \ldots, \nu_k)$ имеет полиномиальное распределение $M(n; p_1, \ldots, p_k)$. Эффективной оценкой $\bar{p}$ по методу максимального правдоподобия является $\bar{\nu}/n$.[1] Из центральной предельной теоремы для полиномиального распределения[2] следует, что при $n \to \infty$ оценка $p_i = \nu_i/n$, как случайная величина, асимптотически обладает нормальным распределением:

$$\mathcal{L}\left(\frac{\nu_i}{n}\right) \stackrel{n\to\infty}{=} \mathcal{N}(p_i, \sqrt{np^\circ_i}), \quad i = 1, \ldots, k.$$

[1] § 3.5, пример 12 в [1].
[2] Упр. 33 к главе 1 в [1].

Тогда статистика

$$X_n^2 = \sum_{i=1}^{k} \frac{(\nu_i - np_i^\circ)^2}{np_i^\circ} = \sum_{i=1}^{k} \frac{\nu_i^2}{np_i^\circ} - n$$

асимптотически при $n \to \infty$ будет обладать распределением χ^2. Применим к полученной статистике критерий χ^2 Пирсона:

$$H_0 \text{ rejected} \Leftrightarrow \left\{ X_n^2 > \chi^2_{1-\alpha,n-1} \right\}.$$

В зависимости от того, будет ли отвергнута нулевая гипотеза или нет, будем соответственно считать, что декодированное сообщение «не похоже» на английский текст или «похоже».

2.2 Модель цепи Маркова первого порядка

Представим текст на английском языке как траекторию марковского процесса первого порядка $X(t)$:

$$\mathbf{P}(x_{t+1} = y_{t+1} | x_\tau = y_\tau, \tau \leqslant t) = \mathbf{P}(x_{t+1} = y_{t+1} | x_t = y_t), \quad y_t \in \Sigma,$$

— вероятность появления символа в позиции $t+1$ зависит только от того, какой символ был в позиции t. Такой марковский процесс задаётся матрицей вероятностей появления символа j сразу после символа i: $S = (p_{ij})_{i,j=1}^k$.

Рассмотрим в траектории процесса $X(t)$ все позиции $T \subset \mathbb{N}$, где встречается символ $c_i \in \Sigma$, для фиксированного i. Вероятности появления конкретных символов на следующих позициях $\{t+1 | t \in T\}$ описываются i-й строкой матрицы S. Таким образом появление следующего за c_i символа описывается некоторой случайной величиной ξ_{eng_i}, принимающей значения из Σ с вероятностями $(p_{i1}, \ldots, p_{ik})$.

Для проверки схожести декодированного текста с текстом на английском языке применим критерий χ^2 Пирсона, описанный в пункте 2.1, для случайной величины ξ_{eng_i}.

Список литературы

[1] Г.И. Ивченко and Ю.И. Медведев. *Введение в математическую статистику.* М: Издательство ЛКИ, 2010.