Обзор статьи Spectral Hashing

Святокум Полина

НИУ ВШЭ

02.10.2017

Семантическое кодирование

Представление данных в сжатом бинарном виде с "сохранением расстояния". Преимущества:

- Сжатие данных
- Возможность быстро найти похожий объект

Требования:

- Легко может быть вычислен для нового объекта выборки
- Небольшая битность
- Расстояние Хемминга между кодами похожих объектов небольшое

Требования

- Каждый бит каждого кода имеет равную вероятность быть 0 или 1
- Биты независимы
- Расстояние Хемминга между кодами похожих объектов минимальное возможное

Задача минимизации

Пусть $y_i{}_{i=1}^{n}$ – коды объектов, $W_{n \times n}$ – матрица близости
В качестве расстояния используется $W_{i,j} = \exp(-\|x_i - x_j\|^2/\epsilon^2)$

$$\text{minimize: } \sum_{i,j} W_{ij} \|y_i - y_j\|^2$$
$$\text{subject to: } y_i \in \{-1, 1\}^k$$
$$\sum_i y_i = 0$$
$$\frac{1}{n} \sum_i y_i y_i^T = I$$

Решение для $k = 1$

Бинарное кодирование длины 1 разбивает граф W на две равные доли.
В таком случае функция минимизации $\sum\limits_{i,j} W_{ij} \|y_i - y_j\|^2$ описывает стоимость разреза.
Такая задача является NP-трудная.

Лапласиан графа

$$W = (w_{ij})_{i,j=1}^{n}, \;\; d_i = \sum_{j=1}^{n} w_{ij}, \;\; D = diag(d_i)$$

$L = D - W$ – лапласиан.

$$L \succeq 0$$

$$L \cdot \bar{1} = 0$$

Спектральная релаксация

Пусть $Y = (y_1|y_2|\dots|y_k)^T$

$$\begin{gathered}\text{minimize: } tr(Y^T L Y)\\ \text{subject to: } Y(i,j) \in \{-1, 1\}\\ Y^T \overline{1} = 0\\ Y^T Y = I\end{gathered}$$

Спектральная релаксация

$$\text{minimize: } tr(Y^T L Y)$$
$$\text{subject to: } Y^T \bar{1} = 0$$
$$Y^T Y = I$$

Решения – собственные вектора L, соответствующие минимальным собственным значениям (не считая тривиальное).
Для нахождения кода объекта не из выборки используют метод Нистрома. Но в таком случае кодирование одного элемента по сложности не выигрывает у поиска ближайшего соседа примитивным способом.

Обобщение для объектов не из выборки

Предположим все объекты выборки порождаются $p(\cdot)$

$$\text{minimize: } \iint \|y(x_1) - y(x_2)\|^2 W(x_1, x_2) p(x_1) p(x_2) dx_1 dx_2$$
$$\text{subject to: } y(x) \in \{-1, 1\}^k$$
$$\int y(x) p(x) dx = 0$$
$$\int y(x) y(x)^T p(x) dx = I$$
$$W(x_1, x_2) = e^{-\|x_1 - x_2\|^2/\epsilon^2}$$

Обобщение для объектов не из выборки

$$\text{minimize: } \iint \|y(x_1) - y(x_2)\|^2 W(x_1, x_2) p(x_1) p(x_2) dx_1 dx_2$$
$$\text{subject to: } \int y(x) p(x) dx = 0$$
$$\int y(x) y(x)^T p(x) dx = I$$
$$W(x_1, x_2) = e^{-\|x_1 - x_2\|^2 / \epsilon^2}$$

Решение

Взвешенный оператор Лапласа — Бельтрами
$g = L_p f \leftrightarrow \frac{g(x)}{p(x)} = D(x)f(x)p(x) - \int_s W(s,x)f(s)p(s)ds$, где
$D(x) = \int_s W(x,s)$
Решение задачи – собственные функции L_p ($L_p f = \lambda f$), соответствующие минимальным собственным значениям (не считая тривиальное).

Собственные функции оператора Лапласа — Бельтрами

Для равномерного распределения $U[a, b]$ собственная функция $\Phi_k(x)$ и собственное значение λ_k равны

$\Phi_k(x) = \sin\left(\frac{\pi}{2} + \frac{k\pi}{b-a}x\right)$

$\lambda_k = 1 - e^{-\frac{k^2}{2}|\frac{k\pi}{b-a}|^2}$

Для многомерного распределения, представимого в виде $p(x) = \prod_i u_i(x_i)$, собственная функция, соответствующая собственному значению $\lambda_{i_1}\lambda_{i_2}\cdots\lambda_{i_d}$, равна $\Phi_{i_1}(x_1)\Phi_{i_2}(x_2)\cdots\Phi_{i_d}(x_d)$.

Заметим, что $sign(\Phi(x_1)\Phi(x_2)) = sign(\Phi(x_1))sign(\Phi(x_2))$, потому биты кодировки не будут независимыми.

Алгоритм

- Найти главные компоненты с помощью PCA. Будем считать, что данные равномерно распределены по многомерному прямоугольнику, образованному главными компонентами.
- Вычисляем k наименьших собственных функций L_p. Для этого вычисляем k наименьших собственных значений для каждого направления, и из полученных dk чисел выбираем k наименьших.
- Применяем найденные функции и округляем полученные значения (по знаку).

Результаты

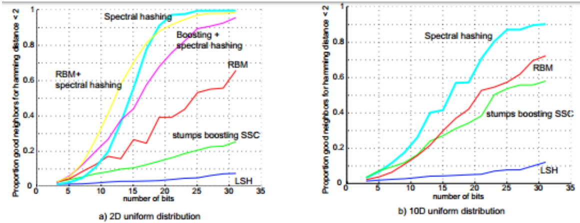

Figure 4: **left:** results on 2D rectangles with different methods. Even though spectral hashing is the simplest, it gives the best performance. **right:** Similar pattern of results for a 10 dimensional distribution.

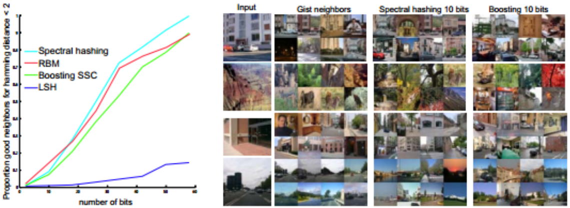

Figure 5: Performance of different binary codes on the LabelMe dataset described in [3]. The data is certainly not uniformly distributed, and yet spectral hashing gives better retrieval performance than boosting and LSH.

Критика

Понравилось:

- Попробовали совместить свой метод с известными методами МО.

Не понравилось:

- Не приведено сравнение с методом Нистрома.
- Нет ссылок на решения задач минимизации.
- Нет сравнения по времени работы.
- Мало наборов данных

Естественные предположения

- В оптимальном бинарном кодировании все коды не коррелируют.
- Данные поступают из единого распределения

Неестественные предположения

- В оптимальном бинарном кодировании вероятность появления 0 или 1 одинакова
- Данные поступают из многомерного равномерного распределения