Ансамблевый классификатор на основе коллаборативной фильтрации

Ю. Кашницкий, В. Сегодин, А. Лоптев, Д.И. Игнатов

1 Постановка задачи

- Есть входная таблица вида "объекты–признаки", где один из признаков — целевой (то, что будем предсказывать), и набор классификаторов $cl_1, ..., cl_N$;
- Построена таблица "объекты–классификаторы" с оценками качества классификации (как вариант, "зазор" классификации (classification margin) — вероятность правильной классификации минус максимальная вероятность неправильной классификации) в процессе кросс-валидации;
- Есть тестовые объекты, для которых такие оценки получить не можем, так как, естественно, не знаем целевого класса;
- Задача — применить рекомендательную систему (коллаборативную фильтрацию) и каждому тестовому объекту «рекомендовать» свой классифкатор. Таким образом составить ансамблевый классификатор.

2 Игрушечный пример

4 обучающих объекта (g_1 — g_4), один – тестовый (g_5), 2 признака объектов (m_1, m_2), 4 классиффикатора (cl_1 — cl_4). Таблицы "объекты–признаки" и "объекты–классификаторы" приведены ниже. Оценка 0.3 для объекта g_2 и классификатора cl_2, например, — это classification margin для объекта g_2 и классификатора cl_2, обученного на всех объектах, кроме g_2 (в случае Leave-One-Out кросс-валидации), то есть на g_1, g_3 и g_4.

	m_1	m_2
g_1	1	2
g_2	3	1
g_3	2	4
g_4	4	2
g_5	5	1

	cl_1	cl_2	cl_3	cl_4
g_1	0.5	0.2	0.6	0.8
g_2	0.4	0.3	0.7	0.2
g_3	0.3	0.5	0.6	0.2
g_4	0.7	0.8	0.5	0.3
g_5	?	?	?	?

Задача – путем оптимизации предсказать для g_5 "зазоры классифкации" и выбрать наиболее подходящий классификатор для g_5.

Обозначения [1]:

- $n_{clf} = 4$ — число классификаторов (аналогия с n_u – числом пользователей рекомендательной системы фильмов)
- $n_{obj} = 4$ — число объектов (аналогия с фильмами)
- X — признаки объектов обучающей выборки (фильмов). Размер: $(n_{feat}+1) \times n_{obj})$. (3×4). Первая строка– из единиц (intercept term);
- Θ — признаки классификаторов (будут найдены в процессе оптимизации). Размер: $(n_{feat} + 1) \times n_{clf}$. (3×4);
- Y — матрица оценок качества классификаций на объектах. Размер: $n_{obj} \times n_{clf}$. (4×4);
- μ — вектор средних значений по строкам матрицы Y (нужен для нормализации оценок);
- J — функция потерь, которую надо минимизировать;
- λ — коэффициент регуляризации.

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 2 & 4 \\ 2 & 1 & 4 & 2 \end{bmatrix}, \Theta = \begin{bmatrix} \theta_0^{(1)} & \theta_0^{(2)} & \theta_0^{(3)} & \theta_0^{(4)} \\ \theta_1^{(1)} & \theta_1^{(2)} & \theta_1^{(3)} & \theta_1^{(4)} \\ \theta_2^{(1)} & \theta_2^{(2)} & \theta_2^{(3)} & \theta_2^{(4)} \end{bmatrix}, Y = \begin{bmatrix} 0.5 & 0.2 & 0.6 & 0.8 \\ 0.4 & 0.3 & 0.7 & 0.2 \\ 0.3 & 0.5 & 0.6 & 0.2 \\ 0.7 & 0.8 & 0.5 & 0.3 \end{bmatrix}$$

Этапы:

1. Нормализация: $Y \to Y - \mu = Y - \begin{bmatrix} 0.525 \\ 0.4 \\ 0.4 \\ 0.575 \end{bmatrix} = \begin{bmatrix} -0.025 & -0.325 & 0.075 & 0.375 \\ 0 & -0.1 & 0.3 & -0.2 \\ -0.1 & 0.1 & 0.2 & -0.2 \\ 0.125 & 0.225 & -0.075 & -0.275 \end{bmatrix}$;
2. Инициализация $\theta_i^{(j)}$ малыми сучайными величинами;
3. Минимизация

$$J(\theta^{(1)}, \theta^{(2)}, \theta^{(3)}, \theta^{(4)}) = \sum_{i=1}^{n_{obj}} \sum_{j=1}^{n_{clf}} (\Theta^{(j)^T} X^{(i)} - y_j^{(i)})^2 + \lambda \sum_{i=1}^{n_{obj}} \sum_{k=1}^{n_{clf}} \left(\theta_k^{(i)}\right)^2$$

.

4. После оптимизации получим матрицу Θ^*. Новому объекту g_5 с признаками x_5 предсказываем вектор оценок классификации $\Theta^{*T} x_5$ и выбираем классификатор с максимальной оценкой.

[1] Согласно видео-лекциям профессора Andrew NG про рекомендательные системы `http://www.youtube.com/watch?v=saXRzxgFN0o&list=PL_npY1DYXHPT-3dorG7Em6d18P4JRFDvH`