Наивный баесовский классификатор

Анна Алтынова, 331

Май 2021

1 Постановка задачи

Дано:

- Множество объектов X, обладающих конечным набором признаков $\{E_i\}$.
- Каждый признак E_i принимает некоторое численное значение, например $k \in \{0, 1, ...n\}$.
- Определен конечный набор классов объектов $\{F_j\}$, на которые делится X
- Для подмножества $X_t \subset X$ знаем, к какому классу принадлежат его объекты, также знаем значения признаков объектов.

Найти:
Научиться определять, к какому классу относится данный объект $x \in X, x \notin X_t$, зная значения его признаков $\{E_i\}$.

1.1 Пример: Классификация спама

Рассмотрим набор писем, в которых выделен набор ключевых слов $\{E_i\}$, являющихся маркерами спама. Тогда E_i может принимать значения $\in \{0, 1\}$ - входит или не входит слово в документ, F_0, F_1 - классы спама и не-спама. Необходимо по тексту письма определить, является ли оно спамом.

2 Модель

Построим функции $g_i \ \forall i$, которые по данному $x \in X$ определят вероятность вхождения x в класс F_i. Тогда для ответа останется выбрать F_i т.ч.$max_j\{g_j(x)\}$ достигается при i(принцип максимального правдоподобия).
Пусть x обладает набором характеристик $E_x := \{e_i\}$. Нужно посчитать
$g_i(x) = P(F_j|E_x) =^{Bayes'Th.}= \frac{P(E_x|F_j)*P(F_j)}{P(E_x)}$,
если предположить, что $\{E_i\}$ независимы("наивность"), т.е.
$P(F_j|E_i, E_k) = P(F_j|E_i) * P(F_j|E_k) \ \forall i, k, j$, то
$P(F_j|E_x) = P(F_j|e_1) * (P(F_j|e_2) * ... * P(F_j|e_n) = \frac{P(e_1|F_j)*...*P(e_n|F_j)*P(F_j)}{P(e_1)*...*P(e_n)}$
Причем распределение вероятностей
$P(E_i = k|F_j)$, т.е. вероятность, с которой признак E_i принимает значение k в классе F_j, а также $P(E_i = k), P(F_i)$ мы можем вычислить непосредственным подсчетом с помощью тренировочного множества.

2.1 Классификация спама

Возвращаясь к примеру писем, $P(F_0)$ это отношение количества спама $x \in X_t$ к общему количеству писем $x \in X_t$,
$P(F_1) = 1 - P(F_0)$,
$P(E_i = 1)$ это отношение количества писем $x \in X_t$, в которых есть маркер-слово E_i к общему количеству писем $x \in X_t$
$P(E_i = 0) = 1 - P(E_i = 1)$
$P(E_i = 1|F_0)$ это отношение количества спама, в котором есть маркер-слово E_i к общему количеству спама,
$P(E_i = 0|F_0) = 1 - P(E_i = 1|F_0)$
$P(E_i = 1|F_1), P(E_i = 0|F_1)$ - аналогично для не-спама,
тогда $\forall x \notin X_t$ обладающего набором маркеров $E_x = \{e_i\}$
$P(F_0|E_x) = \frac{P(e_1|F_0)*...*P(e_n|F_0)*P(F_j)}{P(e_1)*...*P(e_n)}$ - можем вычислить.
Чтобы определить, является ли письмо спамом, остается установить границу α, такую что
$P(F_0|E_x) >= \alpha \Rightarrow x \in F_0$, иначе $x \in F_1$

3 Замечания

- На практике может произойти ситуация, в которой $P(E_i = k|F_j) = 0$, однако тестовый объект, в котором $E_i = k$ относится к F_j, но согласно формуле вероятность принадлежности x к F_j будет нулевой. Чтобы этого не происходило, нулевые $P(E_i = k|F_j)$ можно заменить на небольшое число $\omega > 0$

- Для замены умножения на сложение можно использовать логарифм, т.е все $P(A)$ заменить на $ln(P(A))$