Статистическая оценка качества рубрикатора, основанная на разделимости распределений в пространстве эмбеддингов.

Федоров Артем Максимович
418 группа ВМК МГУ им. М.В. Ломоносова

Московский Государственный Университет им. М. В. Ломоносова

11 октября 2024 г.

Введение

Генезис задачи классификации: X, K, $\varphi : X \to K$

$\Rightarrow$ приближение задачи для решения: $X_{train}, [X_{test}], Inference \longleftarrow \mathcal{R}$

Оценка целесообразности $\mathcal{R}$: предлагается определить разделимости разметчиками меток рубрикатора и проанализировать результаты с экспертами.

- Система меток отвечала поставленной перед экспертами изначальной задаче
- Наблюдатель (разметчик) может здраво оценить целесообразность проставления метки объекту (метки R разделимы)

Постановка задачи: для заданной эмбеддерной модели $E : X \to \mathbb{R}^n$ восстановить распределение образа размеченных данных $E(X_{train})$ для получения индексов попарной разделимости распределений классов для задачи типа Signal Detection

Восстановление распределений:

Выводится EM алгоритм для смесей распределений **von-Mises-Fisher**, $n >> 1$

$$\begin{cases} p(x \mid \theta) = \sum_{i=0}^{k_i} c_n(k_i) e^{k_i \langle \mu_i, x \rangle} \\ c_n(k_i) = \dfrac{k_i^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(k_i)} \end{cases}$$

Вывод индексов раделимости:

Предлагается альтернатива индексу d'. Выводится явный вид для индексов KL и χ^2 дивергенций.

$$i_{KL}(S_1, S_2),\ i_{\chi^2}(S_1, S_2)$$