

Применение тематической модели классификации в информационном анализе электрокардиосигналов

Соболева Д. М.

Научный руководитель: Воронцов К. В.

Московский государственный университет им. М.В. Ломоносова

Факультет ВМК

Кафедра Математических Методов Прогнозирования

1 мая 2017 г.

Тематическая модель классификации

Дано: W^c – словарь терминов «метки классов»

$C = |W^c|$ – число различных классов

W^{gram3} – словарь терминов «триграммы»

$W = W^{gram3} \cup W^c$ – общий словарь терминов

D — коллекция текстовых документов

Найти: модель $p(c|d) = \sum_{t \in T} \varphi_{ct} \theta_{td}$, $c \in W^c$

$\varphi_t = p(c|t)$ – распределение классов в теме t

$\theta_{td} = p(t|d)$ – распределение тем в документе d

Тематическая модель классификации

Критерий оптимизации: максимум логарифма правдоподобия:

$$L(\varphi, \theta) = \sum_{d \in D} \sum_{w \in W^{gram3}} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \tau \sum_{d \in D} \sum_{w \in W^c} n_{dw} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + R(\varphi_{gram3t}, \theta)$$

$R(\varphi_{gram3t}, \theta)$ – регуляризатор разреживания матрицы φ_{gram3t}

τ – вес модальности «метки классов»

n_{dw} – частота термина w в документе d

$$\begin{cases} L(\varphi, \theta) \rightarrow \max_{\varphi, \theta} \\ \sum_{w \in W} \varphi_{wt} = 1, & \varphi_{wt} \geq 0 \\ \sum_{t \in T} \theta_{td} = 1, & \theta_{td} \geq 0 \end{cases}$$

Наивный байесовский классификатор

Найти: модель $p(c|d) = \sum_{w \in W} p(c|w)p_{dw}$

$$p_{dw} = \frac{n_{dw}}{n_d}$$

$$p(c|w) = \frac{n_{cw}}{n_w}$$

n_{dw} – частота термина w в документе d

n_d – длина документа d

n_{cw} – частота термина w в классе c

n_w – частота термина w во всей коллекции документов

Синдромный алгоритм

Найти: модель¹ $p(c|d) = \sum_{w \in W} \gamma_{cw} p_{dw}$

Формула весов признаков:

$$\Gamma^5: \gamma_{cw} = \ln \frac{B_w(X_m, \theta)(1 - B_w(X, \theta))}{B_w(X, \theta)(1 - B_w(X_m, \theta))}$$

$$\Gamma^3: \gamma_{cw} = B_w(X_m, \theta) - B_w(X, \theta)$$

Критерий отбора K признаков с наибольшими значениями:

$$B_w(X_m, \theta) = \frac{1}{|X_m|+2} (\sum_{s \in X_m} [p_w(s) \geq \theta] + 1)$$

$$p_{dw} = \frac{n_{dw}}{n_d}$$

n_{dw} – частота термина w в документе d

n_d – длина документа d

¹Целых В.Р. Статистические обоснования информационного анализа электрокардиосигналов для диагностики заболеваний внутренних органов.

Метрики качества (1)

1 Мера AUC

$$AUC = \frac{1}{|X_0||X_m|} \sum_{d \in X_0} \sum_{d' \in X_m} [p(c|d) > p(c|d')]$$

2 Мера LogLoss

$$LogLoss = -[c = +1] \log p(c|d) - [c = -1] \log(1 - p(c|d))$$

Метрики качества.

Тематическая модель классификации

- 1 Разреженность матрицы φ по каждой отдельной модальности

$$\varphi = p(w|t), \quad w \in W^c, W^{gram3}$$

- 2 Разреженность матрицы $p(t|c)$, $c \in W^c$

$$p(t|c) = \frac{p(c|t)p(t)}{p(c)}, \quad c \in W^c, \quad t \in T$$

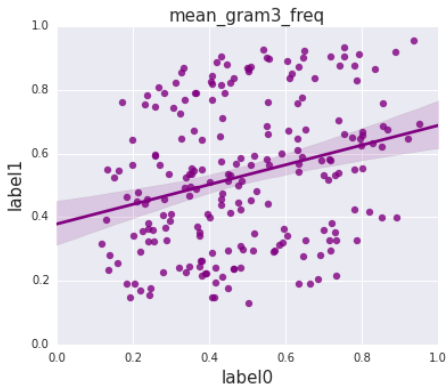
$$p(t) = \sum_{d \in X} p(t|d)p(d) \quad p(d) = \frac{1}{|D|} \quad p(c = +1) = \frac{|D_0|}{|D|}$$

Данные

Эталонная болезнь — хронический холецистит (ХХЭ).

X — кардиограммы ($|X| = 784$)

X_m — кардиограммы больных ($|X_m| = 343$)



Цель экспериментов

Построение конкурентноспособной тематической модели классификации, подбор её параметров и стратегии регуляризации для достижения максимально возможной разреженности распределений:

$$p(w|t), w \in W^c, W^{gram3}$$

Описание экспериментов

Методы оценки моделей:

- 1 10×10 -кратная кросс-валидация.
- 2 hold-out

Начальная инициализация матрицы φ :

- 1 Случайные числа
- 2 Наивный байес совместно с:
 - KMeans
 - Наивным KMeans
 - Растопыrkой

Матрица θ всегда инициализируется равномерным распределением.

Начальная инициализация.

Тематическая модель.

$|T| = 2$ а.к.а наивный байесовский классификатор.

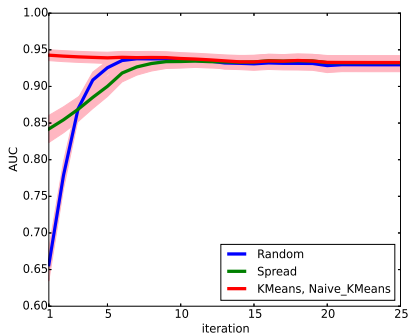


Рис.: AUC. Начальная инициализация

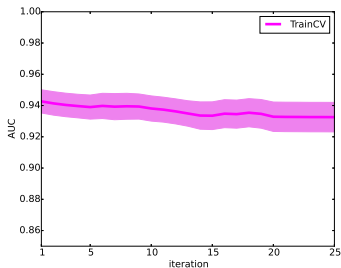
$|T| = 2$ а.к.а наивный байесовский классификатор

Рис.: AUC

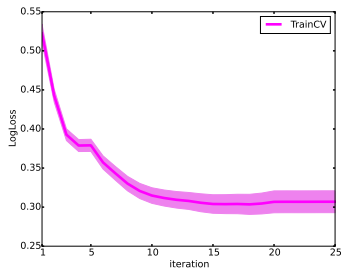


Рис.: LogLoss

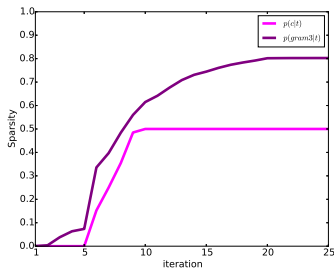
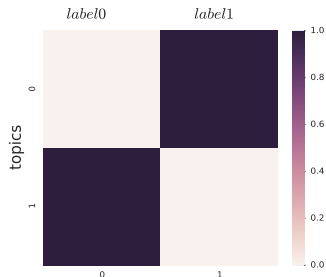
$|T| = 2$ а.к.а наивный байесовский классификатор

Рис.: Разреженность

Рис.: $p(t|c)$

Начальная инициализация.

Тематическая модель.

$|T| = 8$ а.к.а метод В.М. Успенского.

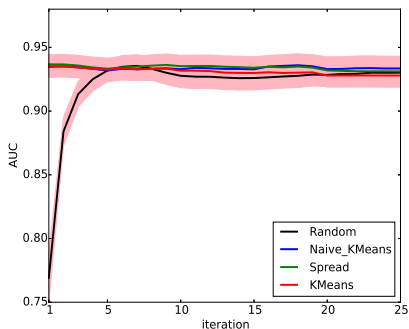


Рис.: AUC. Начальная инициализация

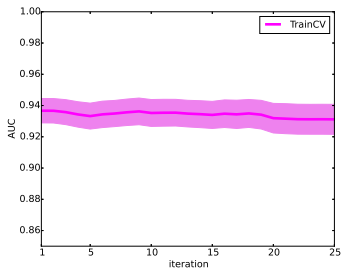
$|T| = 8$ а.к.а метод В.М. Успенского

Рис.: Auc

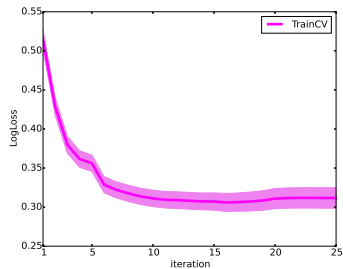


Рис.: LogLoss

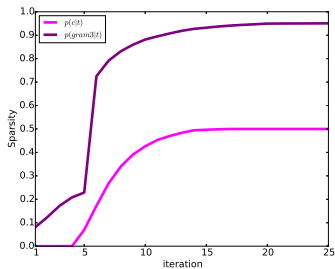
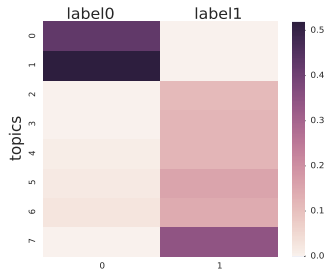
$|T| = 8$ а.к.а метод В.М. Успенского

Рис.: Разреженность

Рис.: $p(t|c)$

Сравнение моделей по всем болезням

Результаты доступны по ссылкам:

[AUC. LogLoss.](#)