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

6 декабря 2016 г.

Тематическая модель двухклассовой 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 $p(c|d) = \sum_{t \in T} \varphi_{ct}\theta_{td}, c \in W^c$
$\varphi_{wt} = p(w|t)$ — вероятности терминов w в каждой теме t
$\theta_{td} = p(t|d)$ — вероятности тем t в каждом документе d

Тематическая модель двухклассовой классификации

Критерий оптимизации: максимум логарифма правдоподобия:

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

τ — вес модальности «метки классов»

n_{dw} — сколько раз термин w встретился в документе d

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

Метрики качества (1)

1. Мера AUC

$$AUC = \frac{1}{C}\sum_{c\in C}\frac{1}{|D_c||D'_c|}\sum_{d\in D_c}\sum_{d'\in D'_c}[p(c|d) > p(c|d')]$$

2. Мера LogLoss

$$-\ln p(y_{true}|y_{pred}) = -(y_{true}\ln y_{pred} + (1 - y_{true})\ln(1 - y_{pred}))$$

3. Перплексия по каждой отдельной модальности

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

$$P = exp(-\frac{1}{n}L)$$

n - длина коллекции в словах.

Метрики качества (2)

1. Разреженность матрицы φ по каждой отдельной модальности
   $\varphi = p(w|t), \quad w \in W^c, W^{gram3}$
2. Разреженность матрицы $p(t|c)$

$$p(t|c) = \frac{p(c|t)p(t)}{p(c)}$$

$$p(t) = \sum_{d \in D} p(t|d)p(d) \qquad p(d) = \frac{1}{n_d} \qquad p(c) = \frac{1}{n_c}$$

Цель экспериментов

Построение конкурентноспособной тематической модели классификации, подбор её параметров и стратегии регуляризации для достижения максимально возможной разреженности распределений $p(w|t), p(c|t), p(t|d)$.

Описание экспериментов

Эталонная болезнь — хронический холецистит (XX).

X — кардиограммы ($|X| = 372$)
X_m — кардиограммы больных($|X_m| = 224$)

Метод оценки моделей: LOOCV.
Метод оптимизации: покоординатный спуск.

Рассматриваемые диапазоны изменения параметров:

1. $|T| \in range(C, 6C, 1)$
   $\tau \in range(1, 1e5, 10)$.

Начальная инициализация матриц φ и θ — случайные числа.

AUC. Последняя итерация

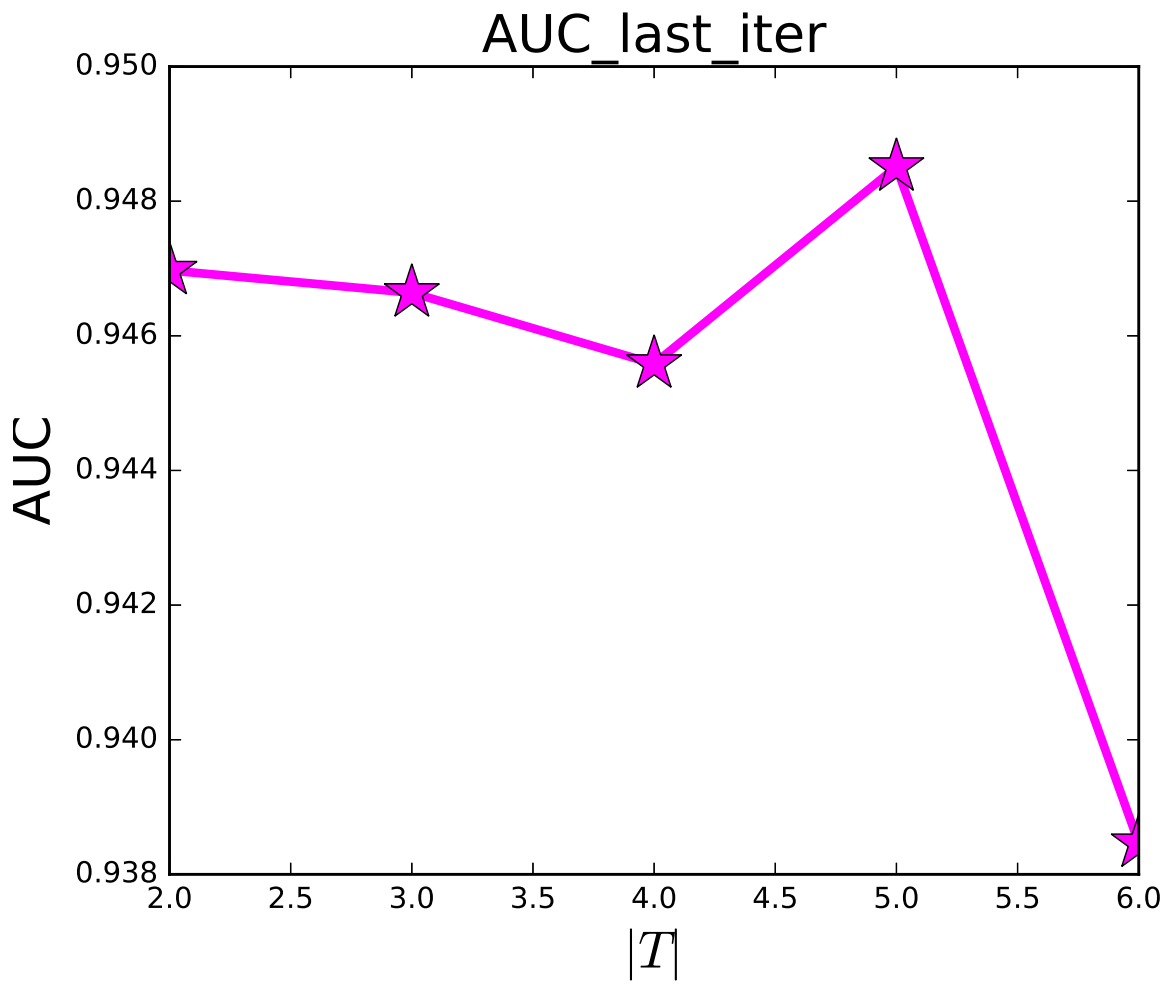

Рис.: AUC, последняя итерация