Московский государственный университет им. Ломоносова
Факультет Вычислительной математики и кибернетики
Кафедра Математических Методов Прогнозирования

Соболева Дарья Михайловна

Применение тематической модели классификации в информационном анализе электрокардиосигналов

Научный руководитель:
д. ф.-м. н. Воронцов Константин Вячеславович

Москва
2016

Содержание

1 Эксперимент №1

1.1 Описание эксперимента

Введем обозначения:

W^c – словарь терминов «метки классов».

$C = |W^c|$ – число различных классов документов.

W^{gram3} – словарь терминов «триграммы».

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

D – коллекция текстовых документов (кардиограмм).

Тематическая модель классификации:

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

Используемые метрики качества:

1. Мера AUC – площадь под рок-кривой в координатах чувствительность-специфичность

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

2. Мера LogLoss. Оценка уверенности классификатора

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

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.

4. Разреженность матрицы φ по каждой отдельной модальности

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

5. Разреженность матрицы $p(t|c)$

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

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

Эксперименты проводятся на эталонной болезни «Хронический холецистит» (ХХЭ).

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

Во множество исследуемых параметров классификатора входят:

- Число тем $|T|$
- Вес модальности «метки классов» τ

Для получения несмещенных оценок применяется процедура $t \times k$-кратной кросс-валидации. Объединенная выборка больных и здоровых делится случайным образом на k блоков одинаковой (с точностью до округления) длины. Каждый блок по очереди используется в качестве контрольного для обученной модели по выборке из остальных k – 1 блоков. Данная процедура повторяется t раз, что позволяет более устойчиво оценить средние ошибки и их доверительные интервалы.

1.2 Цель

Построение конкурентноспособной тематической модели классификации, подбор её параметров и стратегии регуляризации для достижения максимально возможной разреженности распределений $p(gram3|t), p(c|t)$.

1.3 Проведение эксперимента

1.3.1 Поиск оптимального числа тем ($|T|$)

1. $|T| = 2, \tau \in range(1, 700)$.

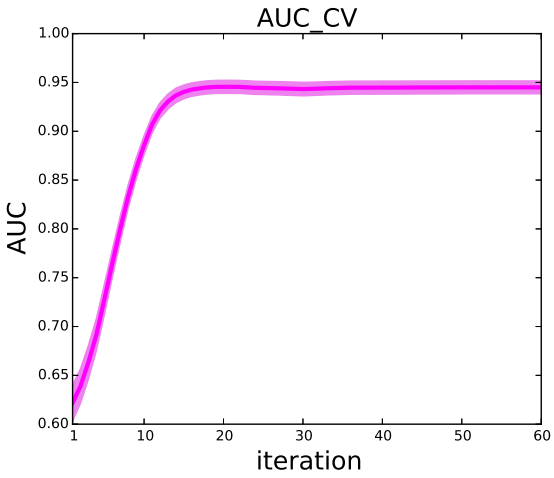

Рис. 1: AUC

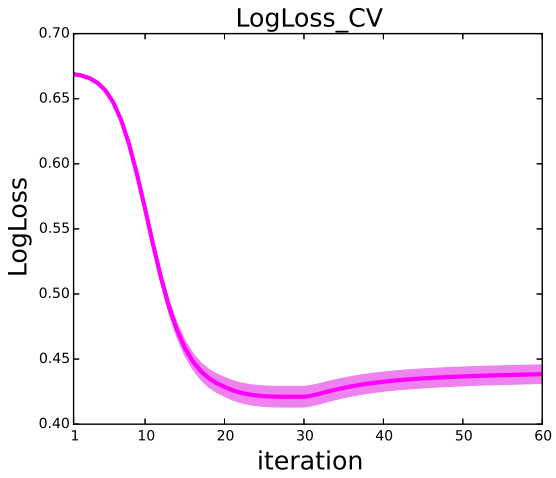

Рис. 2: LogLoss

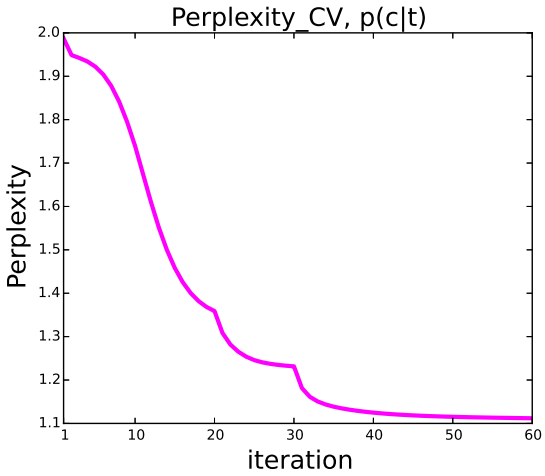

Рис. 3: Перплексия, триграммы

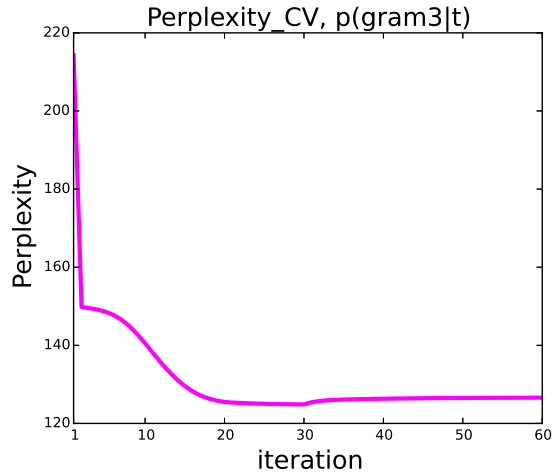

Рис. 4: Перплексия, метки классов

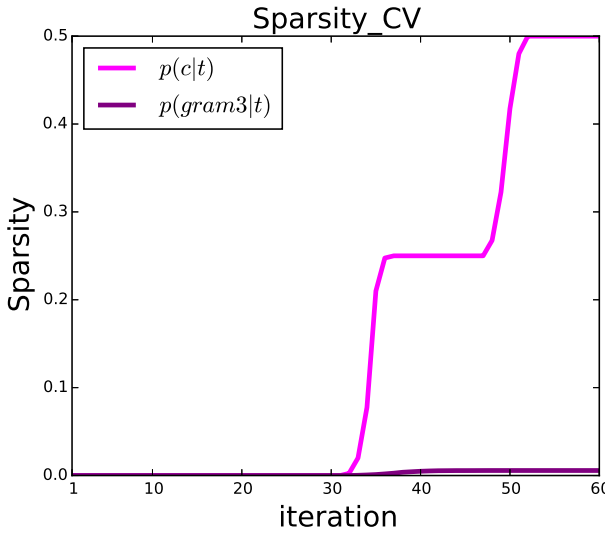

Рис. 5: Разреженность, $p(w|t)$

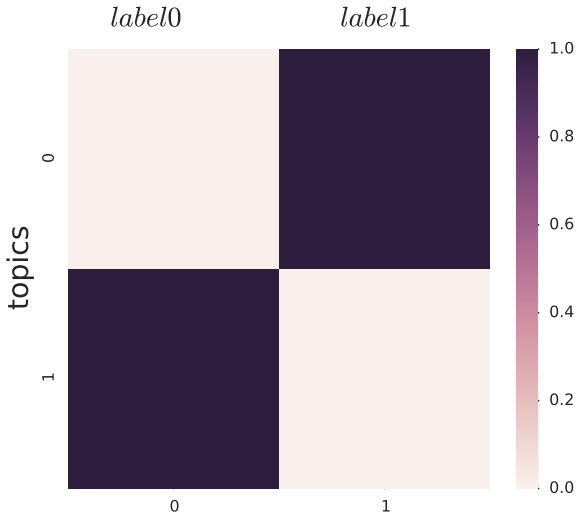

Рис. 6: $p(t|c)$

Краткие выводы: увеличение параметра τ способствует интенсивному разреживанию матрицы $p(c|t)$. Достигнута максимальная степень разреженности. Модель показывает высокие результаты (оценка AUC, LogLoss).

2. $|T| = 3, \tau \in range(1, 10000)$

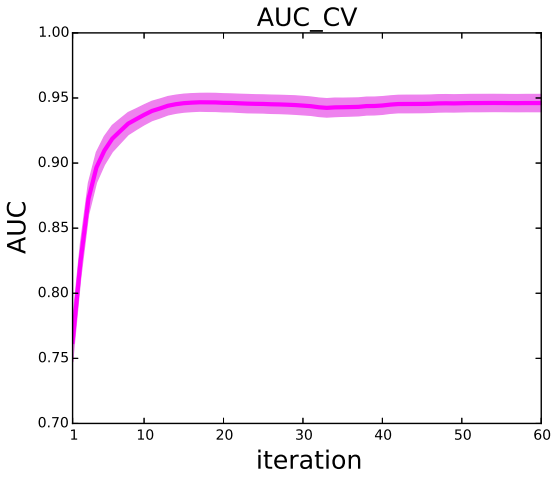

Рис. 7: AUC

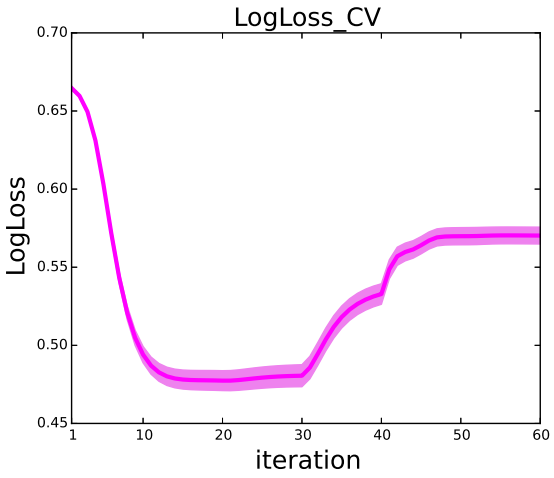

Рис. 8: LogLoss

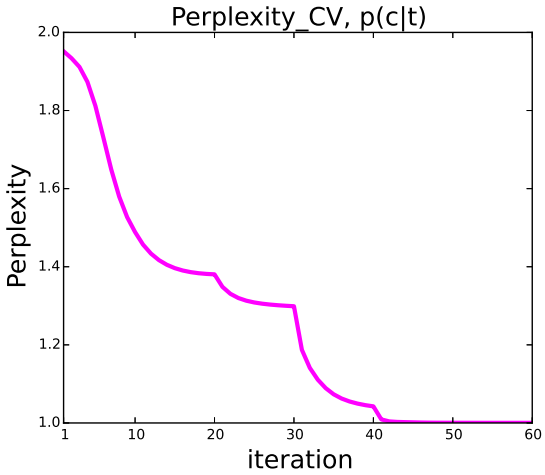

Рис. 9: Перплексия, триграммы

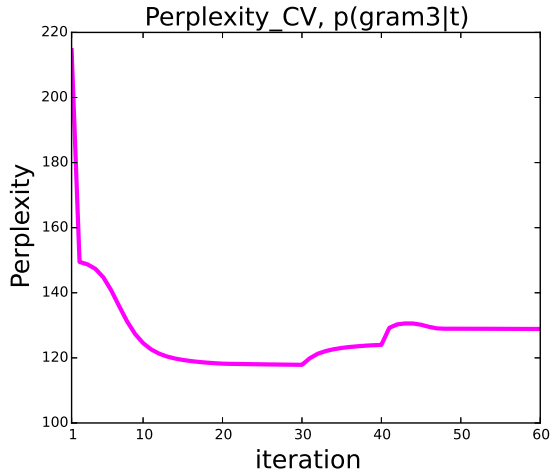

Рис. 10: Перплексия, метки классов

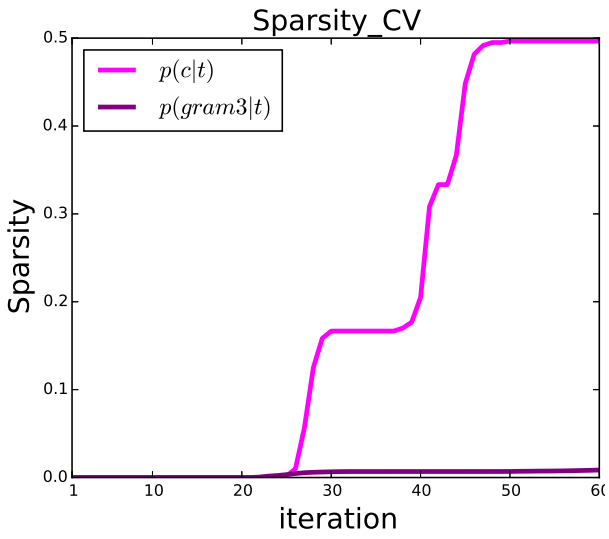

Рис. 11: Разреженность, $p(w|t)$

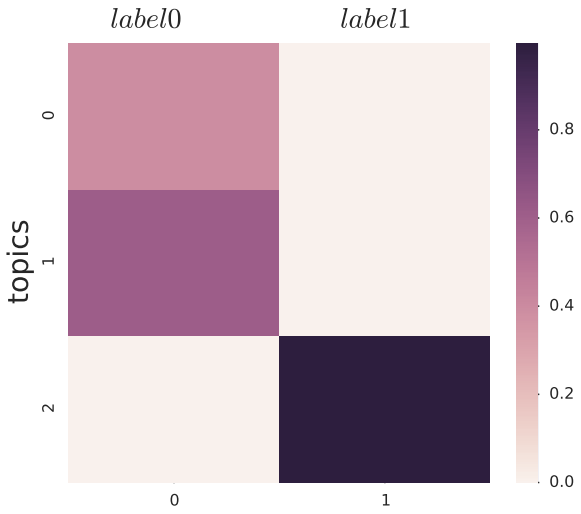

Рис. 12: $p(t|c)$

Краткие выводы: для достижения максимальной разреженности матрицы $p(c|t)$ требуется значительное увеличение параметра τ. Сильное разреживание отрицательно влияет на уверенность классификатора (оценка LogLoss). Качество на начальных итерациях выше, чем у предыдущей модели. Выделились диагностические эталоны, описывающие класс больных (label1). Класс здоровых (label0) описывается хуже.

3. $|T| = 4, \tau \in range(1, 700)$

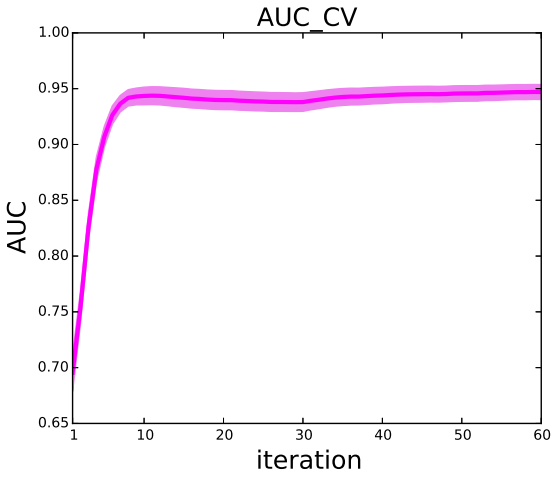

Рис. 13: AUC

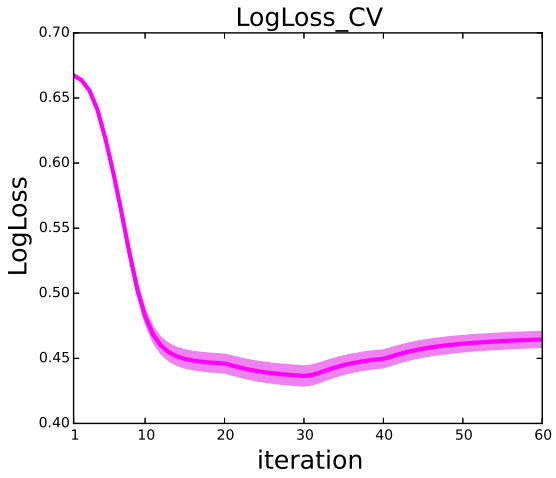

Рис. 14: LogLoss

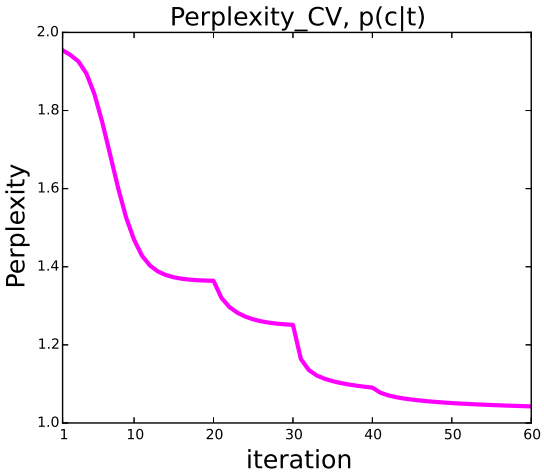

Рис. 15: Перплексия, триграммы

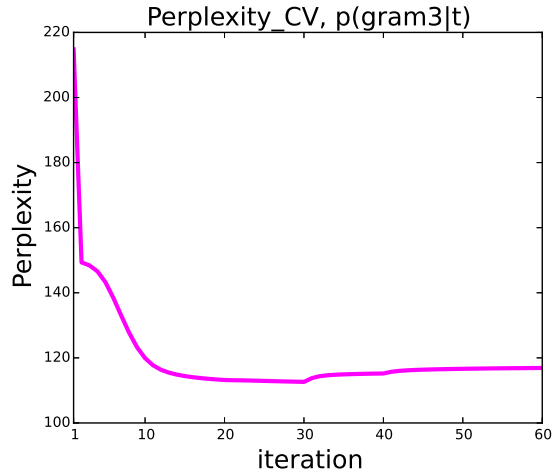

Рис. 16: Перплексия, метки классов

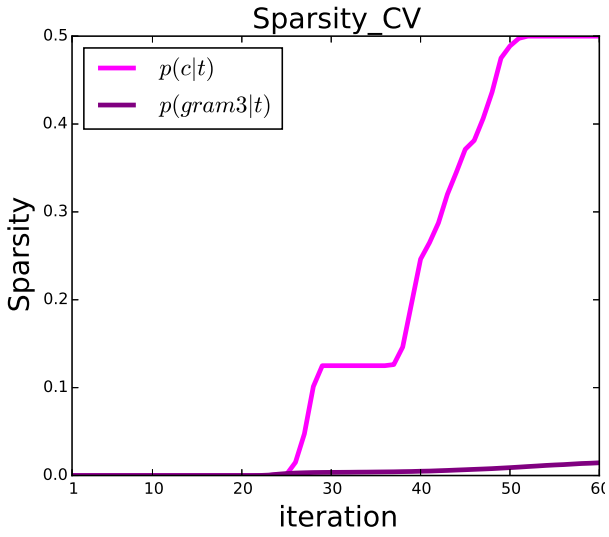

Рис. 17: Разреженность, $p(w|t)$

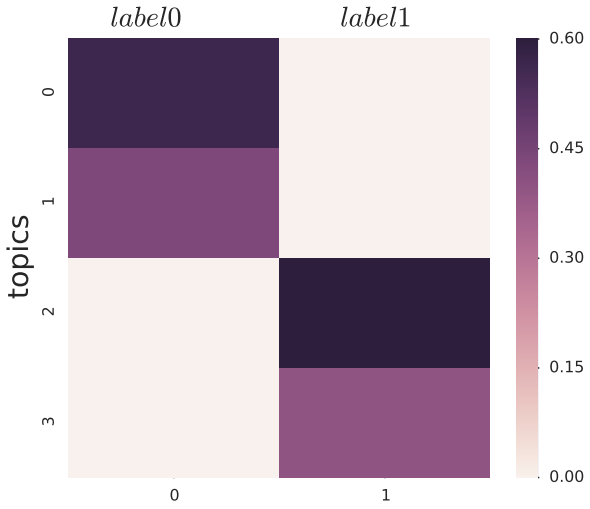

Рис. 18: $p(t|c)$

Краткие выводы: схожа с моделью $|T| = 2$. Качество в смысле меры LogLoss хуже, как и выделение диагностических эталонов в обоих классах.

4. $|T| = 5, \tau \in range(1, 700)$

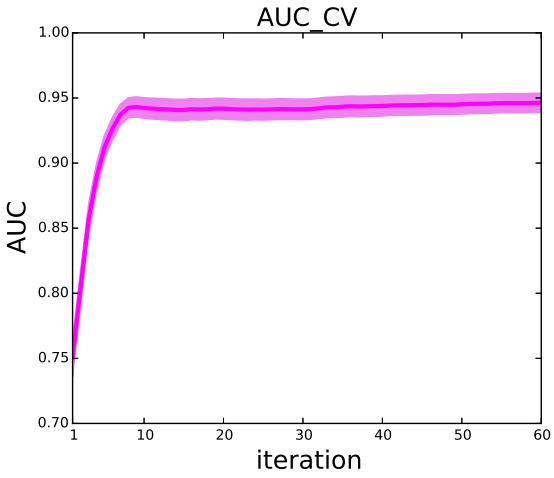

Рис. 19: AUC

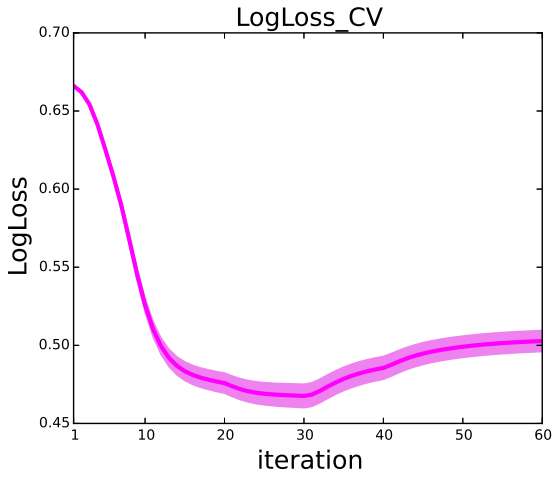

Рис. 20: LogLoss

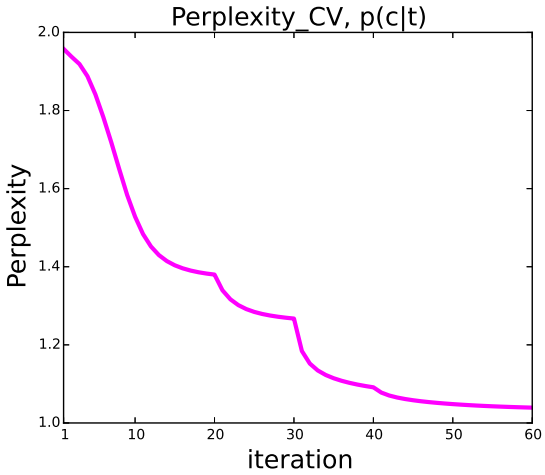

Рис. 21: Перплексия, триграммы

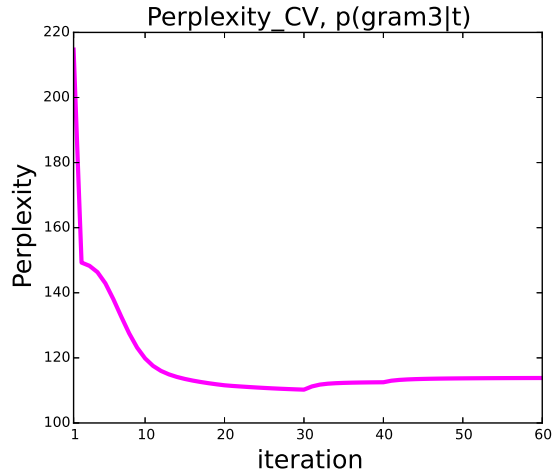

Рис. 22: Перплексия, метки классов

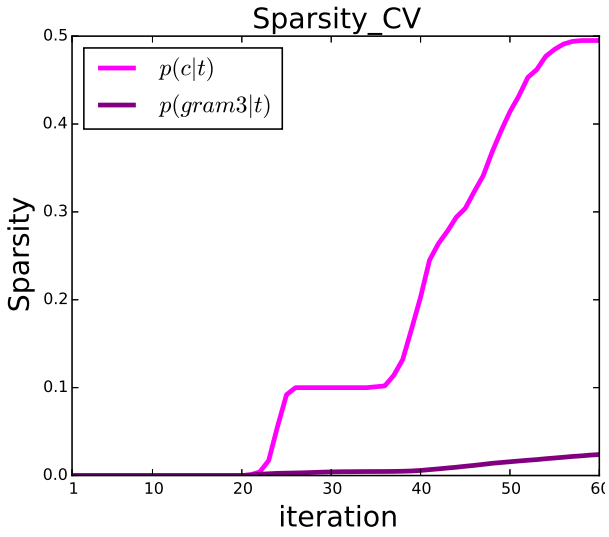

Рис. 23: Разреженность, $p(w|t)$

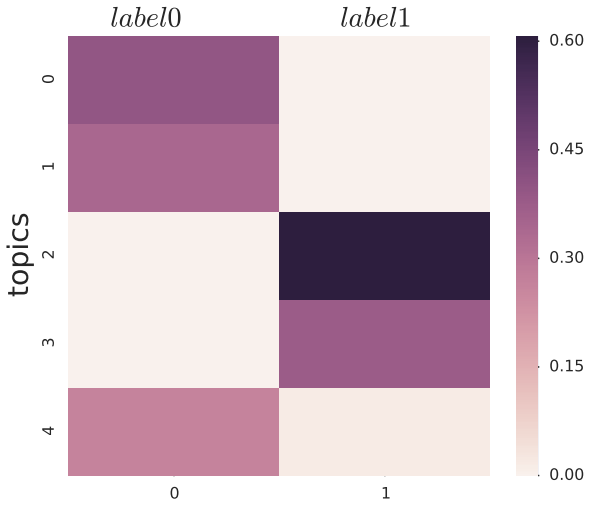

Рис. 24: $p(t|c)$

Краткие выводы: адекватнее модели с $|T| = 3$. Однако не является конкурентом моделям с четными значениями выше рассмотренного параметра $|T|$ в смысле меры LogLoss. Диагностические эталоны также выделяются сравнительно хуже.

5. $|T| = 6, \tau \in range(1, 10000)$

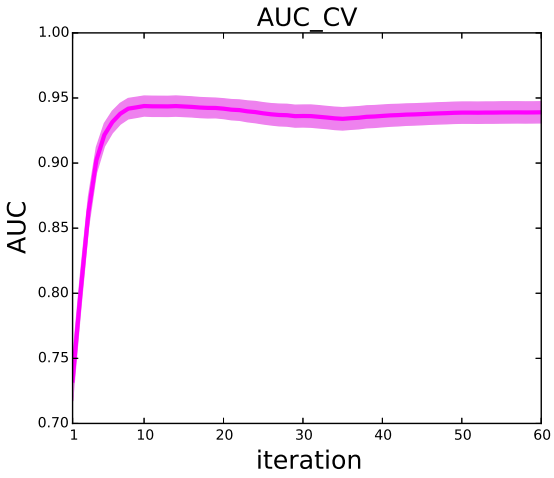

Рис. 25: AUC

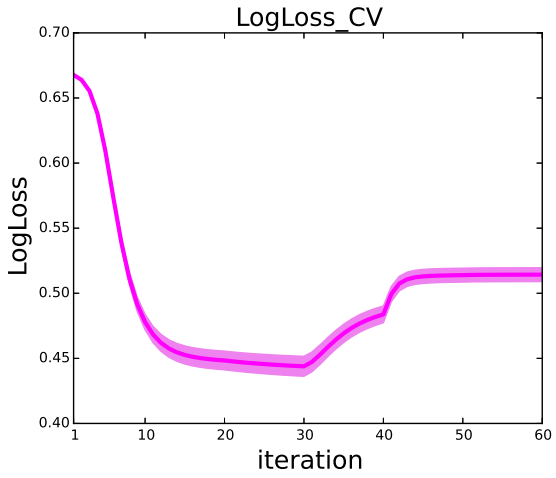

Рис. 26: LogLoss

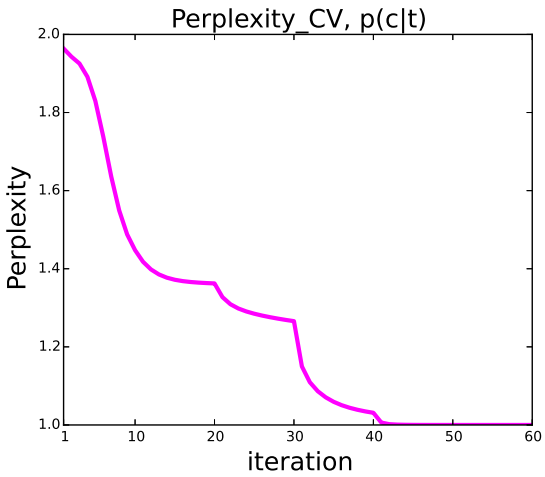

Рис. 27: Перплексия, триграммы

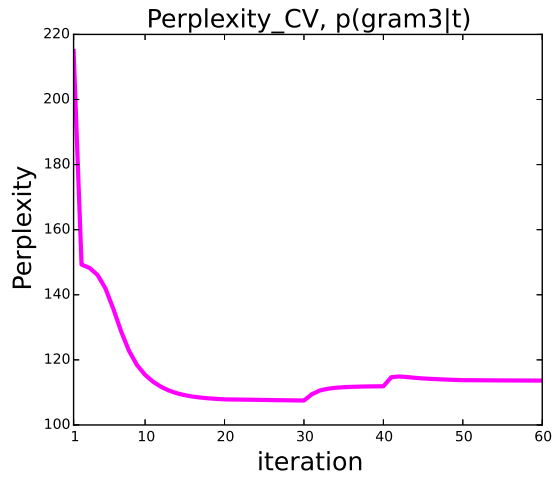

Рис. 28: Перплексия, метки классов

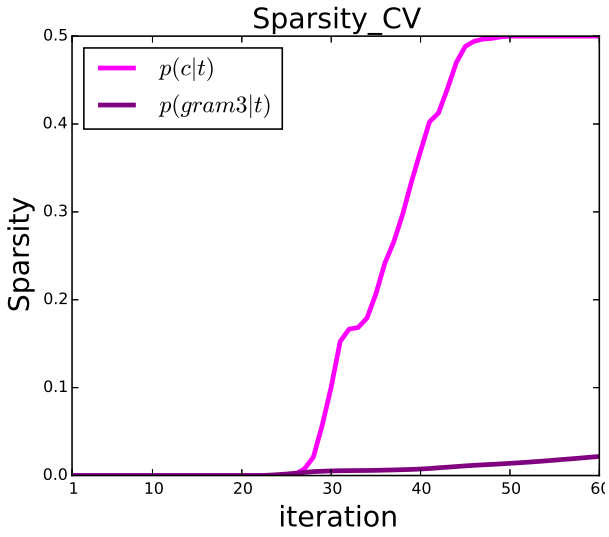

Рис. 29: Разреженность, $p(w|t)$

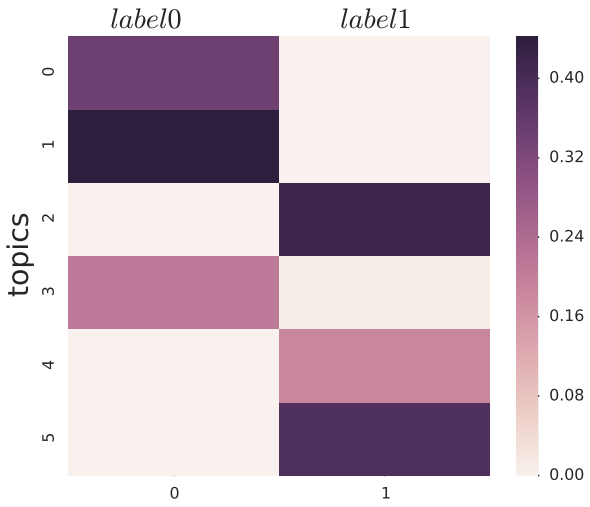

Рис. 30: $p(t|c)$

Краткие выводы: максимальная разреженность матрицы $p(c|t)$ достигается при достаточно большом значении параметра τ, что отрицательно влияет на качество в смысле меры LogLoss, а также AUC. Кривая AUC от итерации проседает при резком увеличении τ. Диагностические эталоны выделяются хуже.

6. $|T| = 7, \tau \in range(1, 900)$

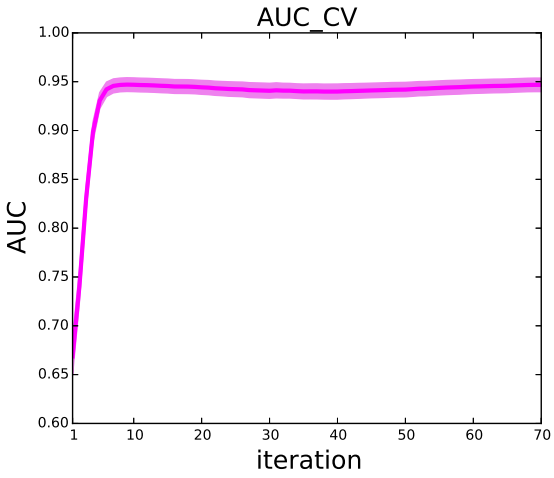

Рис. 31: AUC

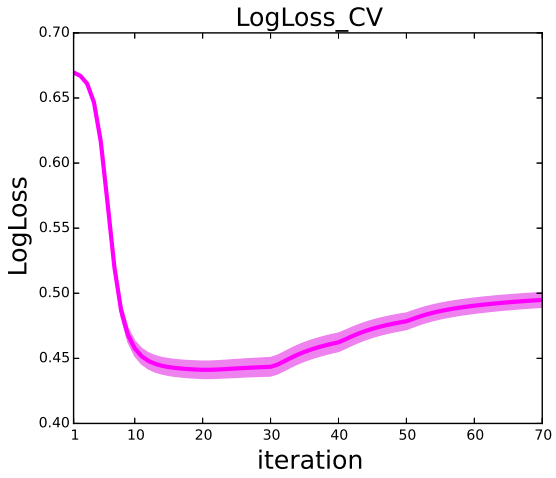

Рис. 32: LogLoss

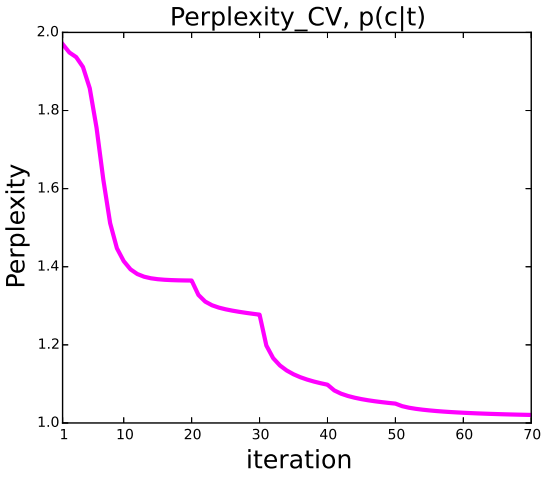

Рис. 33: Перплексия, триграммы

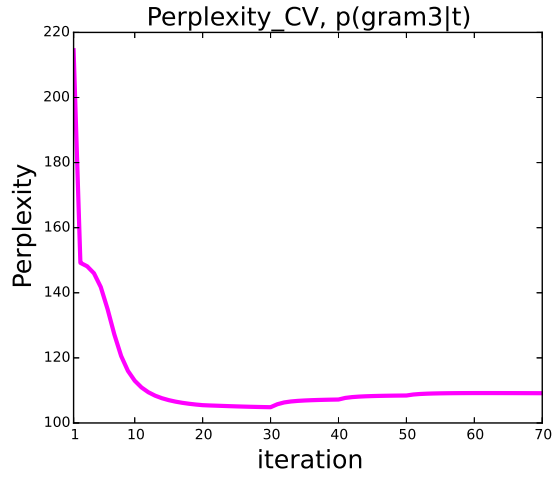

Рис. 34: Перплексия, метки классов

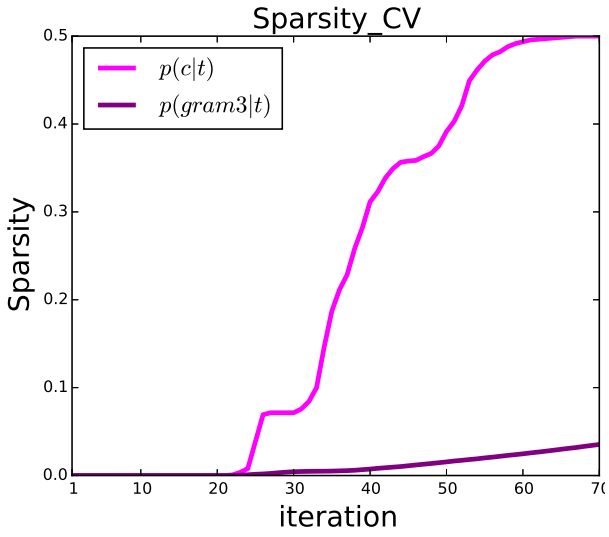

Рис. 35: Разреженность, $p(w|t)$

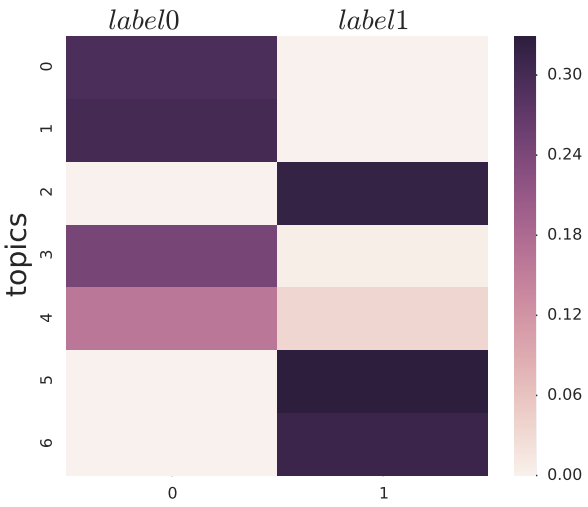

Рис. 36: $p(t|c)$

Краткие выводы: модель потребовала больше итераций для сходимости. Поведение в пределах рассмотренных моделей.