Отчёт по заданию практикума по курсу БММО «Байесовская смесь распределений Бернулли»

Мурат Апишев
great-mel@yandex.ru, MelLain@github.com
МГУ имени М. В. Ломоносова, Москва
15 ноября 2014 г.

Содержание

1 Постановка задачи

1.1 Описание вероятностной модели

Пусть дана выборка $\boldsymbol{X}$ объёма N, $\boldsymbol{x} \in \boldsymbol{X}$, где $\boldsymbol{x} = (x_1, \ldots, x_D)^T$ —набор из D случайных величин $\boldsymbol{x}_i$, каждая из которых имеет распределение Бернулли с параметром μ_i. Рассмотрим смесь из K таких распределений с весами $\boldsymbol{\pi}_k$:

$$p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\pi}) = \sum_{k=1}^{K} \boldsymbol{\pi}_k p(\boldsymbol{x} \mid \boldsymbol{\mu}_k) \tag{1}$$

где $\boldsymbol{\mu} = \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K)\}$, $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_K\}$

Для каждого объекта $\boldsymbol{x}$ введём скрытую переменную $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K)^T$ — бинарный вектор, у которого только одна компонента равна 1, а все остальные нулевые. Тогда можно записать следующее условное распределение:

$$p(\boldsymbol{x} \mid \boldsymbol{Z}, \boldsymbol{\mu}) = \prod_{k=1}^{K} p(\boldsymbol{x} \mid \boldsymbol{\mu}_k)^{z_k}, \quad p(\boldsymbol{x} \mid \boldsymbol{\mu}) = \prod_{i=1}^{D} \mu_i^{x_i} (1 - \mu_i)^{1 - x_i} \tag{2}$$

Введём распределение на $\boldsymbol{z}$:

$$p(\boldsymbol{Z} \mid \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_k} \tag{3}$$

Дополнительно введём априорные распределения $\boldsymbol{\mu}$ и $\boldsymbol{\pi}$:

$$p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \boldsymbol{\pi}_k^{\alpha_k - 1} \tag{4}$$

$$p(\boldsymbol{\mu}_k \mid a,\, b) = \prod_{i=1}^{D} \mathrm{Beta}(\boldsymbol{\mu}_{ki} \mid a,\, b) = \frac{\boldsymbol{\mu}_{ki}^{a-1}(1 - \boldsymbol{\mu}_{ki})^{b-1}}{\mathrm{B}(a,\, b)} \tag{5}$$

Будем рассматривать симметричное распределение Дирихле, т.е. $\boldsymbol{\alpha} = (\alpha,\, \ldots,\, \alpha)^{\mathrm{T}}$. При таких условиях получаем следующее совместное распределение модели:

$$p(\boldsymbol{X},\, \boldsymbol{Z},\, \boldsymbol{\mu},\, \boldsymbol{\pi} \mid \boldsymbol{\alpha},\, a,\, b) = p(\boldsymbol{X} \mid \boldsymbol{Z},\, \boldsymbol{\mu}) p(\boldsymbol{Z} \mid \boldsymbol{\pi}) p(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) \prod_{k=1}^{K} p(\boldsymbol{\mu}_k \mid a,\, b) \tag{6}$$

1.2 Формулировка задания

Требовалось решить следующую задачу

$$p(\boldsymbol{X},\, \boldsymbol{\pi} \mid \alpha, a, b) \to \max_{\boldsymbol{\pi}} \tag{7}$$

Для этого предлагалось воспользоваться вариационным EM-алгоритмом, на E-шаге которого считается вариационное приближение:

$$p(\boldsymbol{Z},\, \boldsymbol{\mu} \mid \boldsymbol{X},\, \boldsymbol{\pi},\, \alpha,\, a,\, b) \approx q(\boldsymbol{Z}) q(\boldsymbol{\mu}), \tag{8}$$

а на M-шаге считается точечная оценка на $\boldsymbol{\pi}$:

$$\mathbb{E}_{q(\boldsymbol{Z})} q(\boldsymbol{\mu}) \ln p(\boldsymbol{X},\, \boldsymbol{Z},\, \boldsymbol{\mu},\, \boldsymbol{\pi} \mid \boldsymbol{\alpha},\, a,\, b) \to \max_{\boldsymbol{\pi}} \tag{9}$$

В рамках выполнения задания было необходимо решить следующие подзадачи:

1. Выписать все необходимые формулы EM-алгоритма.
2. Вывести формулу для подсчёта функционала $\mathcal{L}(q)$ (вариационной нижней границы).
3. Реализовать вариационный EM-алгоритм на Matlab или Python [1] .
4. Реализовать в EM-алгоритме поиск из нескольких случайных начальных приближений, с выбором лучшего по значению $\mathcal{L}(q)$.
5. Протестировать EM-алгоритм на модельных данных (на сгенерированных данных с известными параметрами распределений).
6. Протестировать алгоритм на коллекции MNIST и сделать выводы.
7. Исследовать зависимость логарифма правдоподобия на обучающей и контрольной выборках от кластеризации. Правдоподобие вычислялось по формуле

$$p(\boldsymbol{X}) = \prod_{n=1}^{N} p(\boldsymbol{x}_n \mid \mathbb{E}_{q(\boldsymbol{\mu})} \boldsymbol{\mu},\, \boldsymbol{\pi}_{ML}), \tag{10}$$

   где $p(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\pi})$ задано формулой (1), $\boldsymbol{\pi}_{ML}$ — точечная оценка на параметры $\boldsymbol{\pi}$, полученная в результате работа EM-алгоритма.
8. Рассмотреть величины $q(z_{nk} = 1)$ в качестве признаков $n-$го объекта. Обучить любой классификатор на базе MNIST. Исследовать, как ведёт себя матрица точности на контрольной выборке в зависимости от кластеризации.

[1] Был выбран первый вариант.

2 Вывод формул для EM-алгоритма

Прежде, чем приступить непосредственно к выводу формул, распишем логарифм совместного распределения:

$$\ln p(\boldsymbol{X},\, \boldsymbol{Z},\, \boldsymbol{\mu},\, \boldsymbol{\pi} \mid \boldsymbol{\alpha},\, a,\, b) = \ln p(\boldsymbol{X} \mid \boldsymbol{Z},\, \boldsymbol{\mu}) + \ln p(\boldsymbol{Z} \mid \boldsymbol{\pi}) + \ln p(\boldsymbol{\pi},\, \boldsymbol{\alpha}) + \sum_{k=1}^{K} \ln p(\boldsymbol{\mu}_k \mid a, b) =$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} \ln p(\boldsymbol{x}_n \mid \boldsymbol{\mu}_k)^{z_{nk}} + \sum_{n=1}^{N}\sum_{k=1}^{K} \ln \pi_k^{z_{nk}} + \ln \mathrm{Dir}(\boldsymbol{\pi} \mid \boldsymbol{\alpha}) + \sum_{k=1}^{K}\sum_{i=1}^{D} \ln \mathrm{Beta}(\mu_{ki} \mid a,\, b) =$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \left(\sum_{i=1}^{D} \Big[\ln(\mu_{ki}^{x_{ni}}) + \ln(1-\mu_{ki})^{1-x_{ni}} \Big] \right) + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \pi_k +$$

$$+ \ln \left(\frac{\Gamma(\sum_{k=1}^{K} \alpha_k)}{\prod_{k=1}^{K}(\alpha_k)} \prod_{k=1}^{K} \boldsymbol{\pi}_k^{\alpha_k - 1} \right) + \sum_{k=1}^{K}\sum_{i=1}^{D} \ln \left(\frac{\mu_{ki}^{a-1}(1-\mu_{ki})^{b-1}}{\mathrm{B}(a,\, b)} \right) = \Big\{ \alpha_k = \alpha,\ \forall k = \overline{1, K} \Big\} =$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \left(\sum_{i=1}^{D} \Big[x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1-\mu_{ki}) \Big] \right) + \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \ln \pi_k + \ln \Gamma(K\alpha) -$$

$$- \; K \ln \Gamma(\alpha) + \sum_{k=1}^{K} (\alpha - 1) \ln \pi_k + \sum_{k=1}^{K}\sum_{i=1}^{D} \Big[(a-1) \ln \mu_{ki} + (b-1) \ln(1-\mu_{ki}) - \ln \mathrm{B}(a,\, b) \Big] =$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \left(\sum_{i=1}^{D} \Big[x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1-\mu_{ki}) \Big] + \ln \pi_k \right) +$$

$$+ \ln \Gamma(K\alpha) - K \ln \Gamma(\alpha) + \sum_{k=1}^{K} (\alpha - 1) \ln \pi_k - \sum_{k=1}^{K}\sum_{i=1}^{D} \ln \mathrm{B}(a,\, b) +$$

$$+ \sum_{k=1}^{K}\sum_{i=1}^{D} \Big[(a-1) \ln \mu_{ki} + (b-1) \ln(1-\mu_{ki}) \Big]$$

2.1 Формулы E-шага

Выпишем формулу для $\ln q(\boldsymbol{Z})$:

$$\ln q(\boldsymbol{Z}) = \mathbb{E}_{q(\boldsymbol{\mu})} \ln p(\boldsymbol{X},\, \boldsymbol{Z},\, \boldsymbol{\mu},\, \boldsymbol{\pi} \mid \boldsymbol{\alpha},\, a,\, b) =$$

$$= \mathbb{E}_{q(\boldsymbol{\mu})} \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \left(\sum_{i=1}^{D} \Big[x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1-\mu_{ki}) \Big] + \ln \pi_k \right) + \mathrm{const} =$$

$$= \sum_{n=1}^{N}\sum_{k=1}^{K} z_{nk} \left(\sum_{i=1}^{D} \Big[x_{ni} \mathbb{E}_{q(\boldsymbol{\mu})} \ln \mu_{ki} + (1 - x_{ni}) \mathbb{E}_{q(\boldsymbol{\mu})} \ln(1-\mu_{ki}) \Big] + \ln \pi_k \right) + \mathrm{const}$$

Отсюда получаем формулу для вычисления $q(\boldsymbol{Z})$:

$$\boxed{ q(\boldsymbol{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \rho_{nk}^{z_{nk}},\ \rho_{nk} = \frac{\pi_k \exp\Big\{ \sum\limits_{i=1}^{D} \big[x_{ni} \mathbb{E}_{q(\boldsymbol{\mu})} \ln \mu_{ki} + (1 - x_{ni}) \mathbb{E}_{q(\boldsymbol{\mu})} \ln(1-\mu_{ki}) \big] \Big\}}{\sum\limits_{k=1}^{K} \pi_k \exp\Big\{ \sum\limits_{i=1}^{D} \big[x_{ni} \mathbb{E}_{q(\boldsymbol{\mu})} \ln \mu_{ki} + (1 - x_{ni}) \mathbb{E}_{q(\boldsymbol{\mu})} \ln(1-\mu_{ki}) \big] \Big\}} } \tag{11}$$

[2]

Выпишем формулу для $q(\boldsymbol{\mu})$:

$$\begin{aligned}
\ln q(\boldsymbol{\mu}) = \mathbb{E}_{q(\boldsymbol{Z})} \ln p(\boldsymbol{X},\, \boldsymbol{Z},\, \boldsymbol{\mu},\, \boldsymbol{\pi} \mid \boldsymbol{\alpha},\, a,\, b) = \\
= \mathbb{E}_{q(\boldsymbol{\mu})} \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\sum_{i=1}^{D} \Big[x_{ni} \ln \mu_{ki} + (1 - x_{ni}) \ln(1 - \mu_{ki}) \Big] \right) + \\
+ \sum_{k=1}^{K} \sum_{i=1}^{D} \Big[(a-1) \ln \mu_{ki} + (b-1) \ln(1 - \mu_{ki}) \Big] + \text{const} = \\
= \sum_{k=1}^{K} \sum_{i=1}^{D} \left(\sum_{n=1}^{N} p(z_{nk} = 1) x_{ni} \ln \mu_{ki} + \sum_{n=1}^{N} p(z_{nk} = 1)(1 - x_{ni}) \ln(1 - \mu_{ki}) + \right. \\
\left. + (a-1) \ln \mu_{ki} + (b-1) \ln(1 - \mu_{ki}) \right) + \text{const} = \\
= \sum_{k=1}^{K} \sum_{i=1}^{D} \left(\Big[\sum_{n=1}^{N} p(z_{nk} = 1) x_{ni} + a - 1 \Big] \ln \mu_{ki} + \Big[\sum_{n=1}^{N} p(z_{nk} = 1)(1 - x_{ni}) + b - 1 \Big] \ln(1 - \mu_{ki}) \right) + \\
+ \text{const}
\end{aligned}$$

Отсюда получаем формулу для вычисления $q(\boldsymbol{\mu})$:

$$\boxed{q(\boldsymbol{\mu}) = \prod_{k=1}^{K} \prod_{i=1}^{D} \text{Beta}(\mu_{ki} \mid \hat{a},\, \hat{b}),\; \hat{a} = a + \sum_{n=1}^{N} \rho_{nk} x_{ni},\; \hat{b} = b + \sum_{n=1}^{N} \rho_{nk}(1 - x_{ni})} \tag{12}$$

2.2 Формула M-шага

В общем виде формула M-шага имеет вид 9. Для решения задачи максимизации запишем лагранжиан, учтя ограничение на веса смеси ($\sum_{k=1}^{K} \pi_k = 1$):

$$\begin{aligned}
\mathcal{L} = \sum_{n=1}^{N} \sum_{k=1}^{K} p(z_{nk} = 1) \left(\sum_{i=1}^{D} \Big[x_{ni} \mathbb{E}_{q(\boldsymbol{\mu})} \ln \mu_{ki} + (1 - x_{ni}) \mathbb{E}_{q(\boldsymbol{\mu})} \ln(1 - \mu_{ki}) \Big] + \ln \pi_k \right) + \\
+ \ln \Gamma(K\alpha) - K \ln \Gamma(\alpha) + \sum_{k=1}^{K} (\alpha - 1) \ln \pi_k - \sum_{k=1}^{K} \sum_{i=1}^{D} \ln \text{B}(a,\, b) + \\
+ \sum_{k=1}^{K} \sum_{i=1}^{D} \Big[(a-1) \mathbb{E}_{q(\boldsymbol{\mu})} \ln \mu_{ki} + (b-1) \mathbb{E}_{q(\boldsymbol{\mu})} \ln(1 - \mu_{ki}) \Big] + \lambda (1 - \sum_{k=1}^{K} \pi_k)
\end{aligned}$$

Продифференцируем его и приравняем производную нулю:

$$\frac{d\mathcal{L}}{d\pi_k} = \sum_{n=1}^{N} p(z_{nk} = 1) \frac{1}{\pi_k} + (\alpha - 1) \frac{1}{\pi_k} - \lambda = 0$$

Найдём λ:

$$\lambda = \sum_{k=1}^{K} \lambda \pi_k = \sum_{k=1}^{K} \left(\sum_{n=1}^{N} p(z_{nk} = 1) + (\alpha - 1) \right) = \sum_{k=1}^{K} \sum_{n=1}^{N} p(z_{nk} = 1) + \sum_{k=1}^{K} (\alpha - 1) = N + K(\alpha - 1)$$

[2] $\rho_{nk} = p(z_{nk} = 1)$

Получаем окончательную формулу для M-шага:

$$\boxed{\pi_k = \frac{\sum_{n=1}^{N} \rho_{nk} + \alpha - 1}{N + K(\alpha - 1)}} \tag{13}$$

2.3 Вид оптимизируемого функционала

Запишем общую формулу для функционала $\mathcal{L}(q)$ в вариационном выводе:

$$\mathcal{L}(q) = \int \ln \frac{p(X,\, T)}{\prod_{j=1}^{J} q_j(T_j)} \prod_{j=1}^{J} q_j(T_j) dT_j = \mathbb{E}_{q_1(\mathrm{T}_1),\ldots,q_J(\mathrm{T}_\mathrm{J})} \ln \frac{p(X,\, T)}{\prod_{j=1}^{J} q_j(T_j)} \to \max_{q_1,\ldots,q_J}$$

В случае нашей задачи функционал примет следующий вид:

$$\begin{aligned}
\mathcal{L}(q) = &\sum_{n=1}^{N}\sum_{k=1}^{K} \rho_{nk} \left(\sum_{i=1}^{D} \Big[x_{ni} \mathbb{E}_{q(\boldsymbol{\mu})} \ln \mu_{ki} + (1 - x_{ni}) \mathbb{E}_{q(\boldsymbol{\mu})} \ln(1 - \mu_{ki}) \Big] + \ln \pi_k \right) + \\
&+ \ln \Gamma(K\alpha) - K \ln \Gamma(\alpha) + \sum_{k=1}^{K} (\alpha - 1) \ln \pi_k - \sum_{k=1}^{K}\sum_{i=1}^{D} \ln \mathrm{B}(a,\, b) + \\
&+ \sum_{k=1}^{K}\sum_{i=1}^{D} \Big[(a-1) \mathbb{E}_{q(\boldsymbol{\mu})} \ln \mu_{ki} + (b-1) \mathbb{E}_{q(\boldsymbol{\mu})} \ln(1 - \mu_{ki}) \Big] - \\
&- \sum_{n=1}^{N}\sum_{k=1}^{K} \rho_{nk} \ln \rho_{nk} - \sum_{k=1}^{K}\sum_{i=1}^{D} \underbrace{\int \mathrm{Beta}(\mu_{ki} | \hat{a}, \hat{b}) \ln \mathrm{Beta}(\mu_{ki} | \hat{a}, \hat{b}) d\mu_{ki}}_{\ln B(\hat{a},\hat{b}) - (\hat{a}-1)\psi(\hat{a}) - (\hat{b}-1)\psi(\hat{b}) + (\hat{a}+\hat{b}-2)\psi(\hat{a}+\hat{b})}
\end{aligned} \tag{14}$$

3 Тестирование и эксперименты

3.1 Тестирование EM-алгоритма на модельных данных

Корректность EM-алгоритма можно проверить с помощью восстановления параметров смеси сгенерированных данных. Произведём генерацию данных со следующими параметрами:

- $K = 20$
- $\alpha = 0.01$
- $a = b = 2$
- $D = 50$
- $N = 1000$

А EM-алгоритм запуститм с набором гиперпараметров следующего вида:

- $K = 50$
- $\alpha = 0.1$
- $a = b = 1$

В результате работы EM-алгоритма параметры генерации $\boldsymbol{\pi}$ и $\boldsymbol{\mu}$ были восстановлены достаточно точно. Сравним векторы весов компонент смеси (рассмотрим только ненулевые):

Как видно из таблицы выше, веса компонент, несмотря на неверные априорные гиперпараметры, восстановлены достаточно точно. Поскольку более 98% массы смеси сосредоточено в

$\boldsymbol{\pi}$ (при генерации)	$\boldsymbol{\pi}$ (после восстановления)
0.00000	0.00010
0.00000	0.00009
0.00067	0.00107
0.00948	0.00201
0.00118	0.00211
0.98865	0.99454
0.00000	0.00009

Таблица 1: Сравнение априорных и апостериорных весов компонент смеси.

$\boldsymbol{\mu}_k$ (при генерации)	0.20	0.03	0.14	0.78	0.67	0.53	0.70	0.49
$\boldsymbol{\mu}_k$ (после восстановления)	0.20	0.03	0.15	0.76	0.68	0.50	0.70	0.50

Таблица 2: Сравнение априорных и апостериорных центров превалирующего кластера.

одной компоненте, качество восстановления $\boldsymbol{\mu}$ можно оценивать по строке, соответствующей этой компоненте смеси:

Хорошее качество восстановления очевидно. Для доказательства корректности созданного EM-алгоритма осталось лишь убедиться в монотонном возрастании функционала качества 14. Это видно из следующего графика:

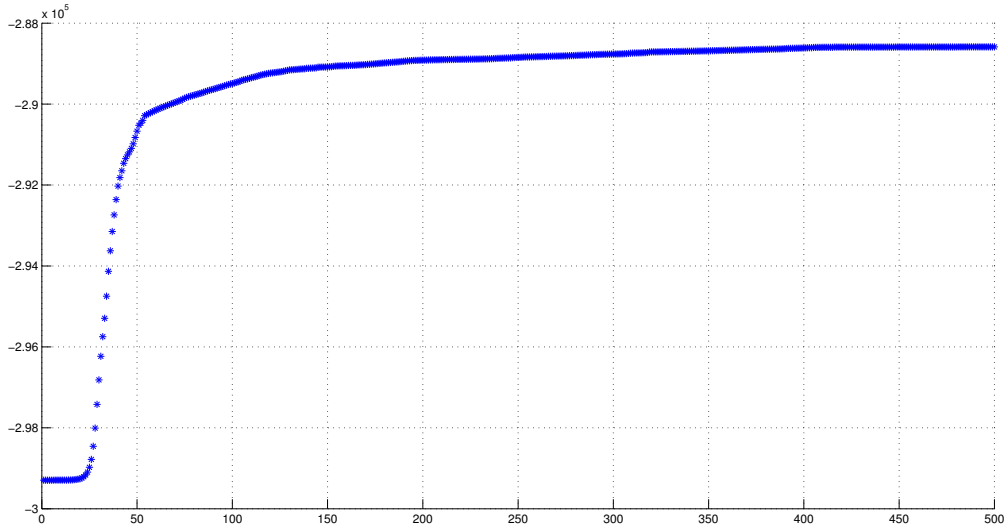

Рис. 1: График изменения значения функционала качества

3.2 Тестирование EM-алгоритма на коллекции MNIST

Рассмотрим работу реализованного вариационного EM-алгоритма на коллекции MNIST. Выборка представлена 60000 объектами [5] , каждый из которых является бинарным вектором длины 784. Основным предметом данного исследования является оценка качества кластеризации объектов коллекции при различных значениях параметров гиперпараметров α, a, b и K.

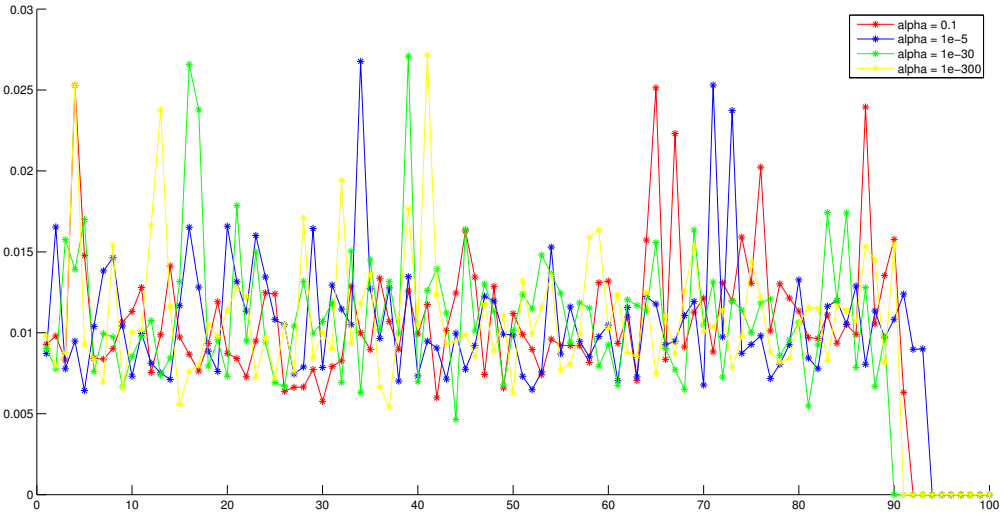

Рис. 2: Сравнение результатов кластеризации в зависимости от гиперпараметра α

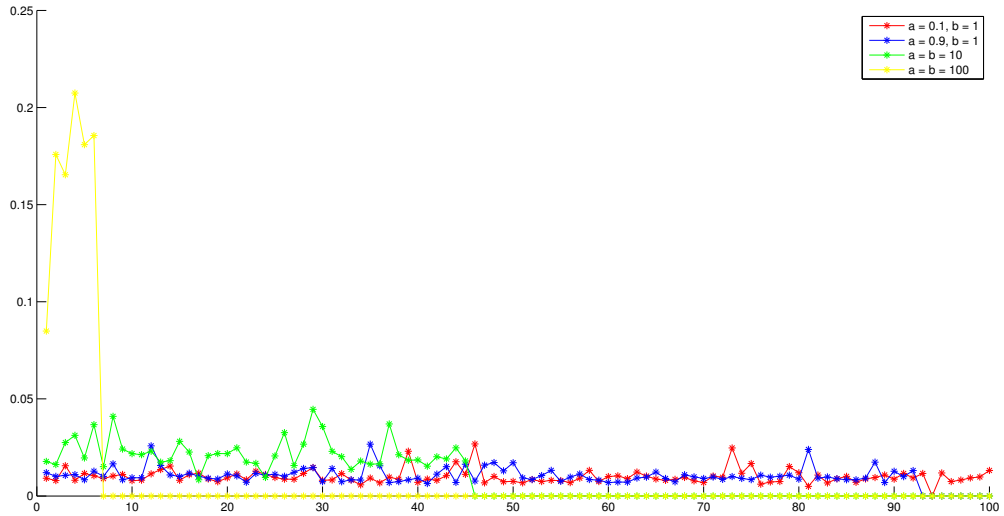

Рис. 3: Сравнение результатов кластеризации в зависимости от гиперпараметров a и b

Для начала рассмотрим зависимость числа кластеров и распределений весов в зависимости от α при фиксации прочих величин.

[5] изображениями рукописных цифр размером 28×28 пикселей

Из графика 2 видно, что параметр α не оказывает никакого существенного влияния на распределение весов компонент смеси, обнуления значительной части весов добиться не удалось. Вероятно, это связано с большой размерностью признакового пространтсва выборки, поскольку в аналогичных экспериментах с синтетическими и немногомерными данными варьирование α немедленно сказывалось на кластеризации.

Рассмотрим теперь подробнее параметры a и b. Как показано на рис. 3, эти две величины напрямую влияют на результаты кластеризации. В случае, когда $a \in (0, 1), b = 1$, количество кластеров несколько уменьшается при увеличении a. При выборе $a = b \in (0, +\infty)$ ситуация аналогичная и ещё более выраженная — увеличение значений гиперпараметров приводит к всё более редким большим кластерам. Визуализация полученных центров кластеров приведена на рис. 4 и 5. Она полностью соответствует полученным выводам — кластеров либо много, и они описывают небольшое количество объектов (т.е. картинки достаточно чёткие), либо мало, и они очень разымыты (кластер пытается описать несколько начертаний одной цифры или даже несколько цифр).

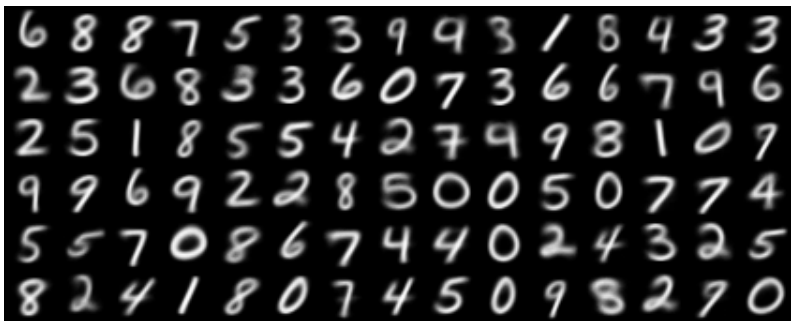

Рис. 4: Визуализация центров кластеров при $a = b = 1$

Рис. 5: Визуализация центров кластеров при $a = b = 100$

С точки зрения специфики коллекции, изображённые на рис. 5 кластеры логичны: цифра «5» сливается с «6», «4» — с «9», «3» — с «8». «0» и «1» мало на кого похожи, поэтому выделились каждый в отдельный кластер.

Заданное значение K	100	200	500
Выявлено кластеров	45	49	43

Таблица 3: Априорное число кластеров и количество выявленных EM-алгоритмом ($a = b = 10$).

Рассмотрим теперь роль гиперпараметра K. Она, как оказалось, является достаточно незначительной — даже в случае слишком большого априорного числа компонент смеси EM-алгоритм всё равно сходится примерно к одному и тому же числу кластеров. Единственным существенным ограничением на K является его ограниченность снизу — задаваемое значение должно быть не меньше разумного возможного реального числа кластеров в данных. Эти выводы подтверждаются таблицей 3. Из неё же видно, что при кластеризации MNIST параметр K должен быть равен хотя бы 50.

3.3 Исследование зависимости логарифма правдоподобия от кластеризации

Поскольку ранее было показано, что кластеризация главным образом зависит от параметров a и b, для начала все исследования в этом разделе будем проводить именно с ними [6] .

Как было сказано ранее, правдоподобие считается по формуле 10. На рис. 6 и 7 приведены графики правдоподобия на обучающей и тестовой выборках, для случаев $a \in (0, 1), b = 1$ и $a = b \in (0, +\infty)$ соответственно. Объем обучающей выборки — 15000 объектов, объём теста — 5000 [7] . Важно учесть, что правдоподобие было нормировано на объём выборки, по которой вычислялось, для сопоставимости результатов.

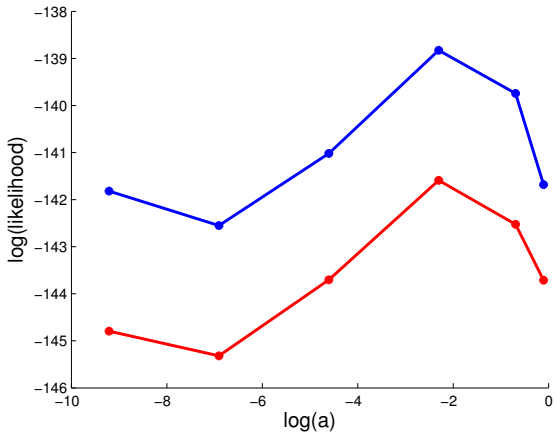

Рис. 6: Значения логарифма правдоподобия на обучающей (синий) и тестовой (красный) выборках, $a \in (0, 1), b = 1$

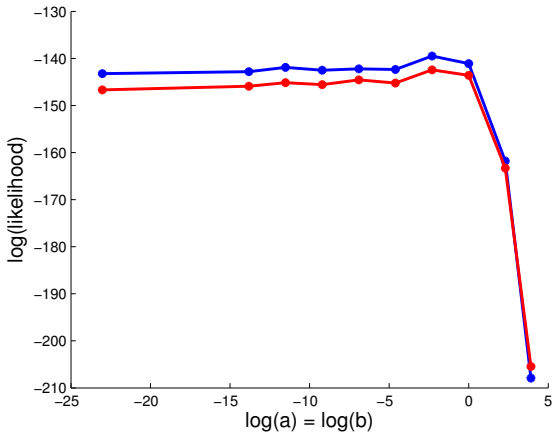

Рис. 7: Значения логарифма правдоподобия на обучающей (синий) и тестовой (красный) выборках, $a = b \in (0, +\infty)$

[6] Все прочие величины имеют значения по-умолчанию.

[7] Объём выборки был уменьшен для ускорения вычислений по сетке значений.

Изображённое на этих графиках позволяет сделать вывод, что никакого существенного изменения различия между значением логарфима правдоподобия на обучающей и тестовой выборках в зависимости от кластеризации не возникает. Однако этот вывод верен только потому, что в EM-алгоритме было взято максимальное количество кластеров $k = 50$. В аналогичном эксперименте, в котором $K = 500$, при очень близких к нулю значениях $a = b$ наблюдается процесс переобучения, связанный с тем, что маленькие значения гиперпараметров провоцируют появление очень большого числа небольших кластеров.

3.4 Обучение классификатора на основе результатов кластеризации

В данном разделе описываются эксперименты с классификацией MNIST при помощи классификатора SVM из библиотеки LibSVM [8] , которому в качестве выборки предоставлялись значения ρ_{nk}, посчитанные на последней итерации работы EM-алгоритма. Обучающая выборка — 50000 объектов, тестовая — 10000. Требовалось оценить изменения, происходящие в матрице точности, в зависимости от кластеризации. Как и ранее, изменяемыми параметрами являются a и b. Важно, что K во всех экспериментах равен 50, что означает сильное уменьшение размерности признакового пространства и, соответственно, размера выборки. В таблицах 4, 5 и 6 приведены матрицы точности [9] для наборов параметров $a = b = 1$, $a = b = 10$ и $a = b = 100$ [10] соответственно.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	95	0	1	0	0	1	1	0	1	0
1	0	96	0	2	0	0	1	0	1	0
2	1	0	92	2	0	0	0	1	5	0
3	0	0	1	82	0	9	0	0	7	1
4	0	1	2	0	58	1	1	4	0	33
5	0	0	1	9	1	81	3	0	5	1
6	1	0	0	0	0	1	98	0	0	0
7	0	1	1	1	2	0	0	85	1	9
8	1	1	1	6	1	4	0	1	87	0
9	1	0	0	2	10	0	0	9	1	77

Таблица 4: Матрица точности для параметров ($a = b = 1$), общая точность $\approx 85\%$.

В матрице 4 хорошо заметно, что относительно «мягкая» кластеризация позволила создать

[8] Поскольку в данной работе не ставится задача максимизации качества классификации за счёт классификатора, все его параметры имеют значения по-умолчанию во всех экспериментах.

[9] Матрицы точности в данном случае представляют собой процент объектов каждого класса, отнесённых к каждому классу. Поскольку проценты были округлены до ближайших целых, незначительное количество данных было потеряно, однако это никак не сказалось на общей картине. В самих иллюстрациях матриц зелёным отмечены ячейки с процентом правильных ответов для объектов каждого класса, белым — проценты небольших ошибок, красным — больших.

[10] Эксперименты были проведены на сетке значений (1, 2, 4, 7, 10, 20, 40, 60, 100), представленные здесь наиболее хорошо характеризуют различные значения параметров. Лучший результат дали значения $a = b = 1$, оптимизация параметров классификатора позволила бы улучшить его ещё сильнее.

признаки, с помощью которых SVM довольно неплохо справился с задачей классификации. Единственной серьёзной проблемой осталось плохое различение «4» и «9».

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0	95	0	0	1	0	1	1	0	2	0
1	0	96	0	0	0	1	0	0	2	0
2	1	0	89	2	1	1	0	1	5	0
3	0	0	1	83	0	1	0	0	14	1
4	0	1	1	0	67	2	1	1	1	27
5	1	0	0	21	2	66	3	0	5	1
6	1	0	0	0	0	1	97	0	0	0
7	0	1	1	0	3	0	0	82	1	10
8	1	1	0	15	1	3	0	1	77	1
9	1	1	0	2	38	0	0	8	2	50

Таблица 5: Матрица точности для параметров ($a = b = 10$), общая точность $\approx 80\%$.

Рассмотрим теперь матрицу 5. Увеличение параметров a и b привело к уменьшению числа кластеров. Это привело к усугублению проблемы с «4» и «9», а также к тому, что некоторые начертания объектов из «5» «8» были ошибочно отнесены к классу «3».

Матрица 6, построенная для случая $a = b = 100$, очень наглядно подтверждает полученные ранее выводы. Количество кластеров явно меньше количества цифр, объекты классов «2», «5» и «9» почти полностью были классифицированы неверно. Объекты остальных классов довольно сильно смешались друг с другом (кроме «1», единица мало похожа на остальные цифры). Очевидно, что столь жёсткое «привязывание» объектов к небольшому числу «размытых» кластеров плохо сказывается на дальнейшей их классификации. Это означает, что используется пространство признков размерности ≤ 10, что явно мало, ибо в исходных 784 признаках значимых было явно больше.

4 Заключение и выводы

В ходе выполнения данной работы был реализован, а затем протестирован, вариационный EM-алгоритм, исследована зависимость кластеризации данных от значений гиперпараметров априорных распределений, а также качество классификации выборки в зависимости от кластеризации, порождающей признаковое описание.

Основные выводы:

1. EM-алгоритм действительно достаточно точно восстанавливает параметры порождающей модели.
2. Гиперпараметры априорных распределений a и b оказались существенными для результатов кластеризации, а K и α — напротив, малозначительными.
3. Зависимость разницы между логарифмами правдоподобия обучающей и тестовой выборок от кластеризации оказалась сильной в том случае, если количество кластеров получилось очень большим.

	0	1	2	3	4	5	6	7	8	9
0	74	0	0	5	0	0	2	0	19	0
1	0	96	0	1	0	0	0	0	3	0
2	1	6	1	12	2	0	65	1	12	0
3	0	6	0	68	3	0	2	1	19	0
4	0	4	0	0	47	0	2	41	5	0
5	1	5	0	34	6	0	2	5	48	0
6	1	6	1	1	0	0	85	0	7	0
7	0	6	0	0	25	0	0	67	1	0
8	0	13	0	28	3	0	1	6	49	0
9	1	3	0	2	46	0	0	47	2	0

Таблица 6: Матрица точности для параметров ($a = b = 100$), общая точность $\approx 49\%$.

4. Кластеризация в большой степени влияет на классификацию объектов, если признаковое описание этих объектов получается в результате этой кластеризации.

Список литературы

[1] Ветров Д.П., Кропотов Д.А. — Байесовские методы машинного обучения, учебное пособие по спецкурсу, 2007

[2] https://wikipedia.org