Байесовские методы в машинном обучении
Практическое задание № 2: EM алгоритм для детектива
ВМК, 417 группа, Штыков Павел

Штыков Павел

20 ноября 2022 г.

1 Теория

1.1 E шаг

$$\begin{aligned}
p(d_k|X_k, \theta, A) = & \\
&= \frac{p(d_k, X_k, \theta, A)}{p(X_k, \theta, A)} \\
&= \frac{p(d_k, X_k|\theta, A)\ p(\theta, A)}{p(X_k|\theta, A)\ p(\theta, A)} \\
&= \frac{p(d_k, X_k|\theta, A)}{\sum_{d_k} p(d_k, X_k|\theta, A)} \\
&= \frac{p(X_k|d_k, \theta)\ p(d_k|A)}{\sum_{d_k} p(X_k|d_k, \theta)\ p(d_k|A)}
\end{aligned}$$

1.2 M шаг

Перепишем вид матожидания:

$$M = \mathbb{E}_{q(d)}\ \log p(X, d|\theta, A) = \sum_K \sum_{d_k} (\log p(X_k|d_k, \theta) + \log p(d_k|A)) q_k(d_k)$$

1.2.1 Без MAP

- $\arg\max_{A_{ij}} M$:

  Так как у нас есть ограничение на матрицу $A : \sum_{ij} A = 1$, то оптимизация условная, следовательно выпишем лагранжиан:

$$L = \sum_K \sum_{d_k} (\log p(X_k|d_k, \theta) + \log p(d_k|A))q_k(d_k) + \lambda(1 - \sum_{ij} A).$$

  Тогда:

$$\begin{aligned}
\frac{\partial L}{\partial A_{ij}} &= \\
&= \frac{\partial}{\partial A_{ij}} \left(\sum_K \sum_{d_k} (\log p(X_k|d_k, \theta) + \log p(d_k|A))q_k(d_k) + \lambda(1 - \sum_{ij} A) \right) \\
&= \frac{\partial}{\partial A_{ij}} \sum_K \sum_{d_k} \log p(d_k|A))q_k(d_k) - \lambda \\
&= \frac{\sum_K q_k(d_k = (i, j))}{A_{ij}} - \lambda
\end{aligned}$$

  Следовательно максимум L достигается при $A_{ij} = \dfrac{\sum_K q_k(d_k = (i, j))}{\lambda}$. Подставим найденный A_{ij} в L: $\lambda - \sum_{ij} \sum_K q_k(d_k = (i, j)) = 0$, следовательно $\lambda = K$.

  И
$$\arg\max_{A_{ij}} M = \arg\max_{A_{ij}} L = \frac{\sum_K q_k(d_k = (i, j))}{K}$$

- $\arg\max_{F_{ij}} M$:

$$
\begin{aligned}
\frac{\partial M}{\partial F_{ij}} = \\
&= \frac{\partial}{\partial F_{ij}} \sum_K \sum_{d_k} (\log p(X_k|d_k,\theta) + \log p(d_k|A)) q_k(d_k) \\
&= \frac{\partial}{\partial F_{ij}} \sum_K \sum_{d_k} q_k(d_k) \log p(X_k|d_k,\theta) \\
&= \frac{\partial}{\partial F_{ij}} \sum_K \sum_{d_k} q_k(d_k) \log \underbrace{C \exp\left(\frac{(X_k(d_k^1+i, d_k^2+j) - F_{ij})^2}{s^2}\right)}_{\text{остался только } ij \text{ компонент}} \\
&= \text{ сокращаем всевозможные константы и дифференцируем} \\
&= \sum_K \sum_{d_k} q_k(d_k)(X_k(d_k^1+i, d_k^2+j) - F_{ij})
\end{aligned}
$$

Следовательно, приравнивая производную к нулю и выражая F_{ij} находим (учитывая суммы):

$$
\arg\max_{F_{ij}} M = \frac{\sum_K \sum_{d_k} q_k(d_k) X_k(d_k^1+i, d_k^2+j)}{\sum_K \underbrace{\sum_{d_k} q_k(d_k)}_{=1}} = \frac{\sum_K \sum_{d_k} q_k(d_k) X_k(d_k^1+i, d_k^2+j)}{K}
$$

- $\arg\max_{B_{ij}} M$:

  Поиск $\arg\max_{B_{ij}} M$ аналогичен $\arg\max_{F_{ij}} M$ с той лишь разницей, что суммирование происходит не по всем d_k, а только по тем, что лежат вне области F: $\hat{d}_k = \{(i,j)|(i,j) \in B \setminus F\}$.

  Следовательно:
$$
\arg\max_{B_{ij}} M = \frac{\sum_K \sum_{\hat{d}_k} q_k(d_k) X_k(i,j)}{\sum_K \underbrace{\sum_{\hat{d}_k} q_k(d_k)}_{\text{уже не равно 1}}}
$$

- $\arg\max_{s^2} M$:

$$
\begin{aligned}
\frac{\partial M}{\partial s^2} = & \\
&= \frac{\partial}{\partial s^2} \sum_K \sum_{d_k} (\log p(X_k|d_k, \theta) + \log p(d_k|A)) q_k(d_k) \\
&= \frac{\partial}{\partial s^2} \sum_K \sum_{d_k} q_k(d_k) \sum_{ij} \log \frac{1}{\sqrt{2\pi} s} \exp\left(\frac{(X_k(i,j) - [(i,j) \in F] F_{ij} - [(i,j) \in B \setminus F] B_{ij})^2}{2s^2} \right) \\
&= \frac{\partial}{\partial s^2} \sum_K \sum_{d_k} q_k(d_k) \sum_{ij} \log \frac{1}{\sqrt{2\pi s^2}} \\
&\quad + \frac{\partial}{\partial s^2} \sum_K \sum_{d_k} q_k(d_k) \sum_{ij} \frac{(X_k(i,j) - [(i,j) \in F] F_{ij} - [(i,j) \in B \setminus F] B_{ij})^2}{2s^2} \\
&= -\frac{HWK}{s^2} + \sum_K \sum_{d_k} q_k(d_k) \sum_{ij} \frac{(X_k(i,j) - [(i,j) \in F] F_{ij} - [(i,j) \in B \setminus F] B_{ij})^2}{s^4}
\end{aligned}
$$

Приравниваем производную к 0 и находим необходимую s^2:

$$
\arg\max_{s^2} M = \frac{1}{HWK} \sum_K \sum_{d_k} q_k(d_k) \sum_{ij} (X_k(i,j) - [(i,j) \in F] F_{ij} - [(i,j) \in B \setminus F] B_{ij})^2
$$

1.2.2 C MAP

Пусть i_k^{MAP}, j_k^{MAP} координаты MAP для k-го изображения. Тогда:

- $\arg\max_{A_{ij}} M$:

$$
\arg\max_{A_{ij}} M = \frac{\sum_K [(i_k^{MAP}, j_k^{MAP}) = (i,j)]}{K}
$$

- $\arg\max_{F_{ij}} M$:

$$
\arg\max_{F_{ij}} M = \frac{\sum_K X_k(i_k^{MAP} + i, j_k^{MAP} + j)}{K}
$$

- $\arg\max_{B_{ij}} M$:

$$
\arg\max_{B_{ij}} M = \frac{\sum_K X_k(i_k^{MAP}, j_k^{MAP})}{\sum_K [(i_k^{MAP}, j_k^{MAP}) \in \hat{d}_k]}
$$

- $\arg\max_{s^2} M$:

$$\arg\max_{s^2} M = \frac{1}{HWK}\sum_K (X_k(i_k^{MAP}, j_k^{MAP})$$
$$- [(i_k^{MAP}, j_k^{MAP}) \in F] F_{i_k^{MAP}, j_k^{MAP}} - [(i_k^{MAP} j_k^{MAP}) \in B \setminus F] B_{i_k^{MAP} j_k^{MAP}})^2$$

1.3 Нижняя оценка логарифма неполного правдоподобия

$$L(q,\theta,A) = \sum_k \sum_{d_k} q(d_k)\left(\log p(X_k, d_k|\theta, A) - \log q(d_k)\right)$$

2 Эксперименты

Сгенерируем данные:

Рис. 1: Пример сгенерированных данных: "лицо фон и все вместе

2.1 Влияние начального приближения

Исследуем влияние начального приближения на работу алгоритма. Генерировать параметры будем из следующих распределений: $F, B \sim Uniform(0, 255)$, $s \sim Uniform(0, 511)$, $\sum_{ij} A = 1$. Картинка зашумлена с $s = 100$.

На Рис. 2 зависимость $L(q,\theta,A)$ от итерации для разных запусков:

На Рис. 3 пример лиц и фонов при разных запусках:

Мы видим, что хоть в абсолютных числах $L(q,\theta,A)$ отличается от запуска к запуску, но на получаемых картинках это отражается не сильно. Однако в последующих экспериментах мы будем использовать мультистарт там, где это не сильно затратно по времени, для

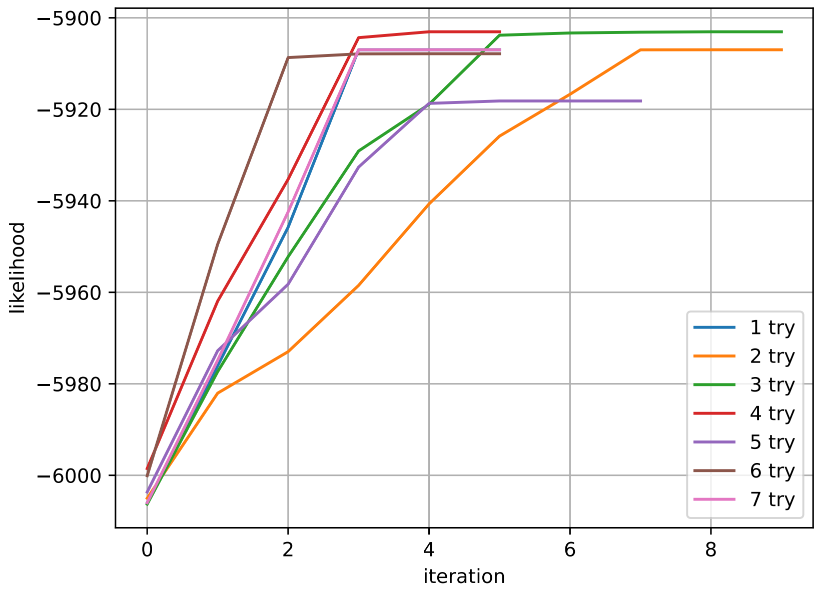

Рис. 2: Зависимость $L(q, \theta, A)$ от итерации для разных запусков

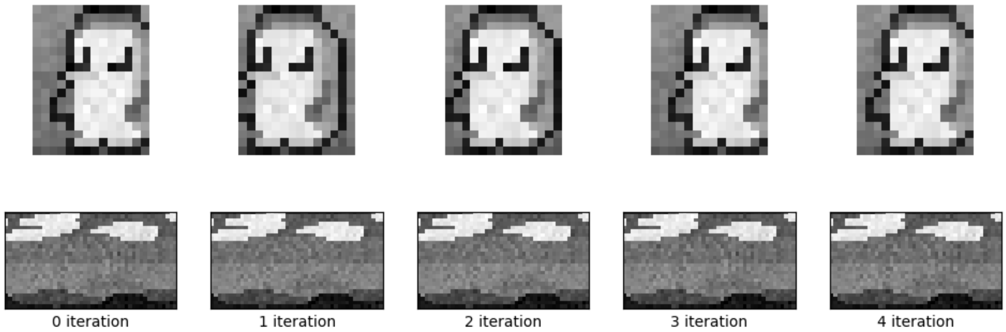

Рис. 3: Пример лиц и фонов при разных запусках

большей достоверности.

2.2 Влияние размеров выборки и зашумленности

На Рис 4 изображена тепловая карта значения $L(q, \theta, A)$ в зависимости от разного количества картинок в выборке и уровня шума.

На Рис 5 приведены полученные лица и фоны при разном количестве картинок и одинаковом шуме.

На Рис 6 приведены полученные лица и фоны при разном шуме и одинаковом количестве картинок.

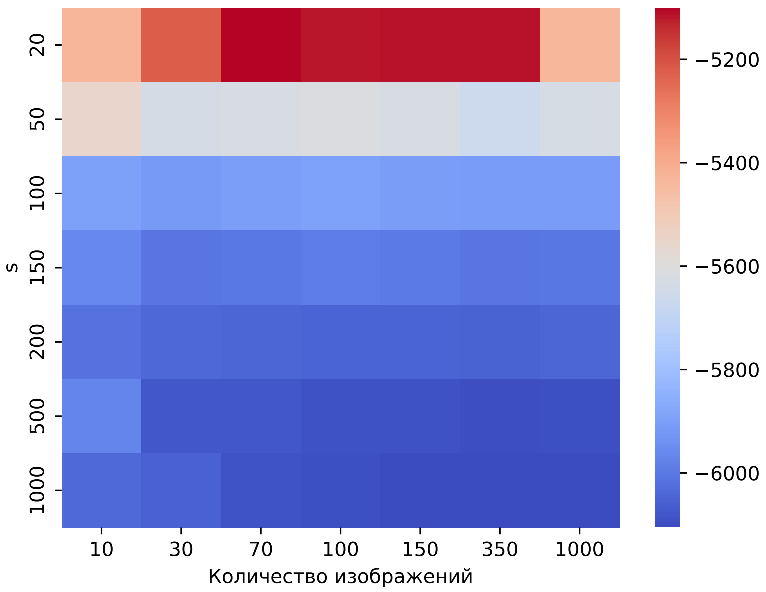

Рис. 4: Тепловая карта значения $L(q, \theta, A)$ в зависимости от разного количества картинок в выборке и уровня шума

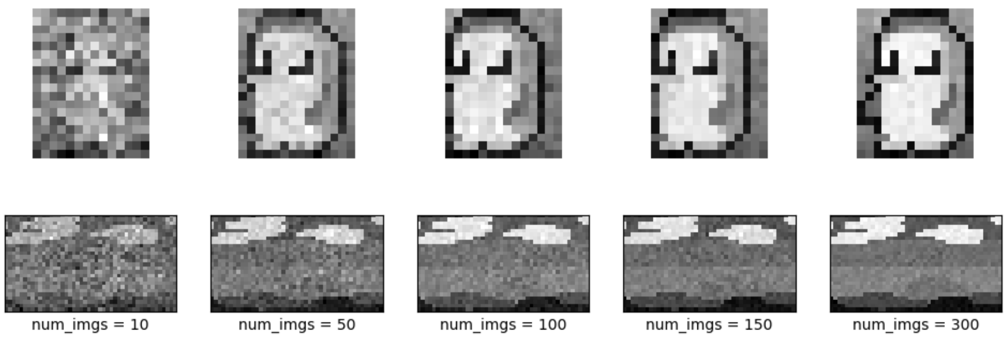

Рис. 5: Результаты при разном количестве картинок и одинаковом шуме $s = 130$.

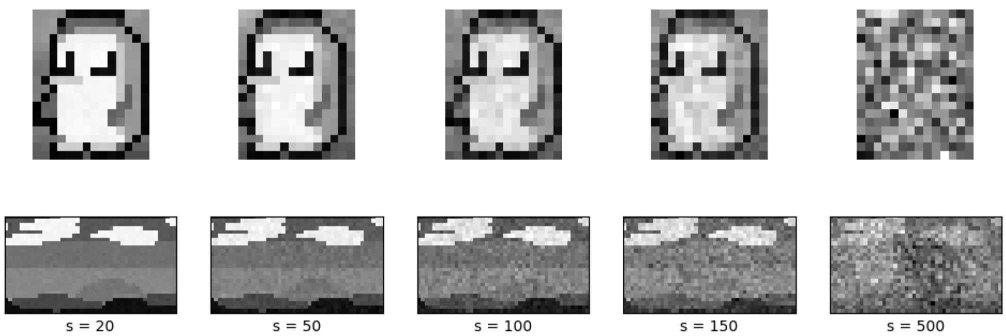

Рис. 6: Результаты разном шуме и одинаковом количестве картинок $num_imgs = 70$.

Мы видим ожидаемый результат: чем меньше шум и больше картинок, тем лучше результат и наоборот. При этом начиная с шума $s = 500$ значение $L(q, \theta, A)$ после обучения остается около -6000, что означает (если вспомнить график 2), что модель не обучается совсем. Это можно наглядно увидеть на рис 7.

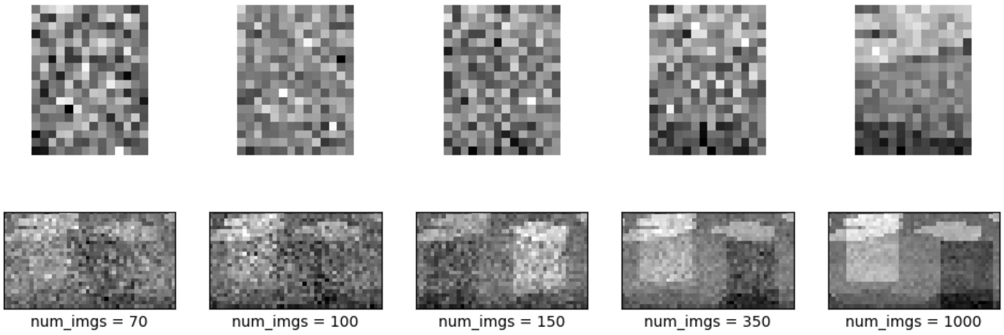

Рис. 7: Результаты при разном количестве картинок и одинаковом шуме $s = 500$.

2.3 Сравнение EM и Hard-EM

На рис 8 и 9 приведены результаты работы EM и hard-EM алгоритмов при разном уровне шума ($s = 100$ и $s = 150$ соответственно) на выборке с 50-ю картинками.

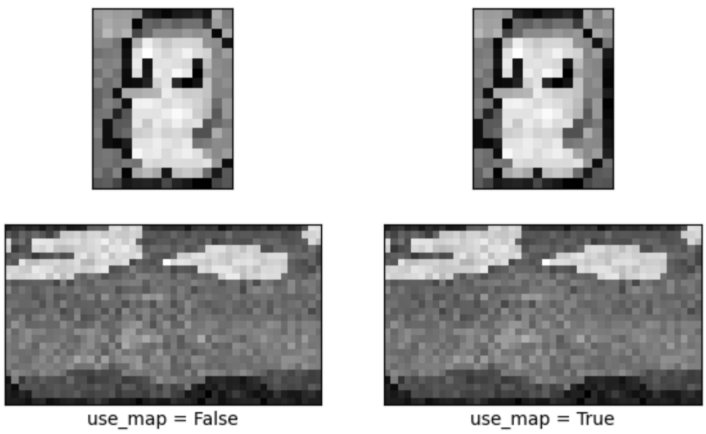

Рис. 8: Результаты работы EM и hard-EM $s = 100, num_imgs = 50$

Мы видим, что если при $s = 100$ влияние MAP не сильно заметно, то при $s = 150$ потери в качестве при аппроксимации явно видны. Так происходит, так как при большой дисперсии аппроксимация нормального шума дельта функцией является слишком неточной. При этом же hard-EM работает примерно в 10 раз быстрее. Соответственно hard-EM можно использовать на слабо зашумленных данных для ускорения расчетов. Однако для

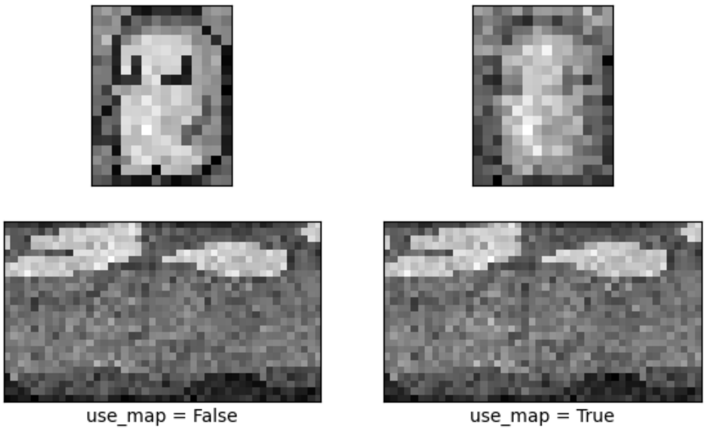

Рис. 9: Результаты работы EM и hard-EM $s = 150, num_imgs = 50$

поиска преступника мы будем использовать полноценный EM алгоритм (для получения наилучшего качества!).

2.4 Вычисление преступника

Применим EM алгоритм к зашумленным фото преступника! Результат приведет на рис. 10.

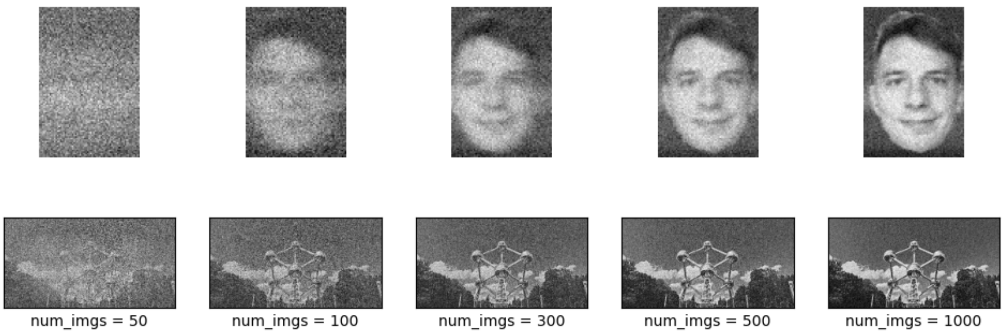

Рис. 10: Результаты работы EM на снимках преступника

Мы видим, что фон становится узнаваем уже со 100-а картинок в выборке. Лицо же становится различимым только с 500-1000 картинок в выборке. Однако найти такого человека среди членов байесгруппы не удалось...

2.5 Возможные модификации алгоритма

Идеи по улучшению алгортма:

- Для мультистарт версии алгоритма мы можем для каждого последующего запуска семплировать параметры не из равномерного распределения, а из некоторого нормального распределения с модой в точке равной значениям финальных параметров с прошлого запускаю

- Можно сделать нечто среднее между EM и hard-EM: Будем хранить не MAP, а top-k координат наибольших апостериорных вероятностей.