Задание №3.3. Модель гауссовой смеси. Максимизация ожидания

Максимизация ожидания (Expectation Maximization, EM) является классическим алгоритмом обучения без учителя (т.е. обучения при наличии скрытых или латентных переменных). В этом задании мы рассмотрим один из подходов, позволяющих адаптировать EM-алгоритм к гибридной ситуации, когда в выборке среди неразмеченных примеров встречаются также примеры с метками (значениями скрытых переменных).

В стандартной постановке задачи обучения без учителя у нас есть $m \in \mathbb{N}$ неразмеченных примеров $\{x^{(1)}, \dots, x^{(m)}\}$. Мы хотим восстановить параметры распределения $p(x, z; \theta)$ по данным, но значения $z^{(i)}$ нам неизвестны. Классический EM-алгоритм как раз и предназначен для этих целей, когда мы максимизируем недоступное нам напрямую распределение $p(x; \theta)$ косвенным образом, выполняя в цикле сначала E-шаг, а потом M-шаг, во время которых максимизируется доступная нам нижняя граница распределения $p(x; \theta)$. Наша целевая функция может быть выписана следующим образом:

$$\ell_{unsup}(\theta) = \sum_{i=1}^{m} \log p\left(x^{(i)}; \theta\right) = \sum_{i=1}^{m} \log \sum_{z^{(i)}} p\left(x^{(i)}, z^{(i)}; \theta\right).$$

Попробуем построить расширенную версию EM-алгоритма, который можно будет применять в гибридной ситуации. Пусть теперь у нас есть дополнительные $\tilde{m} \in \mathbb{N}$ размеченных примеров $\left\{\left(\tilde{\mathrm{x}}^{(1)}, \tilde{\mathrm{z}}^{(1)}\right), \dots, \left(\tilde{\mathrm{x}}^{(\tilde{m})}, \tilde{\mathrm{z}}^{(\tilde{m})}\right)\right\}$, в которых как x, так и z нам известны. Мы хотим одновременно максимизировать маргинальное правдоподобие параметров, используя неразмеченные примеры, а также полное правдоподобие параметров, используя размеченные примеры из выборки. Для этого мы объединим обе функции правдоподобия в одну взвешенную сумму (используя дополнительный гиперпараметр α). Конкретно, наша гибридная целевая функция $\ell_{semi-sup}(\theta)$ может быть записана следующим образом:

$$\ell_{sup}(\theta) = \sum_{i=1}^{\tilde{m}} \log p\left(\tilde{\mathrm{x}}^{(i)}, \tilde{\mathrm{z}}^{(i)}; \theta\right), \qquad \ell_{semi-sup}(\theta) = \ell_{unsup}(\theta) + \alpha \ell_{sup}(\theta).$$

Мы можем вывести шаги гибридного EM-алгоритма тем же самым способом, который мы использовали для его стандартной реализации. Настоятельно рекомендуем вам сделать это самостоятельно. Ниже мы приводим уже готовые результаты.

E-шаг (гибридный алгоритм)

Для каждого $i \in \{1, \dots, m\}$ установить:

$$Q_i^{(t)}\left(\mathrm{z}^{(i)}\right) \coloneqq p\left(z^{(i)} \middle| x^{(i)}; \theta^{(t)}\right).$$

M-шаг (гибридный алгоритм)

$$\theta^{(t+1)} \coloneqq \underset{\theta}{\operatorname{argmax}} \left[\sum_{i=1}^{m} \left(\sum_{z^{(i)}} Q_i^{(t)}\left(\mathrm{z}^{(i)}\right) \log \frac{p\left(x^{(i)}, z^{(i)}; \theta\right)}{Q_i^{(t)}\left(\mathrm{z}^{(i)}\right)} \right) + \alpha \left(\sum_{i=1}^{\tilde{m}} \log p\left(\tilde{\mathrm{x}}^{(i)}, \tilde{\mathrm{z}}^{(i)}; \theta\right) \right) \right].$$

Вопрос №1. Сходимость [2 балла]

Для начала мы покажем, что этот алгоритм сходится. Для того чтобы это доказать, достаточно показать, что наша гибридная цель $\ell_{semi-sup}(\theta)$ монотонно возрастает с каждой итерацией E и M шагов. Более конкретно, пусть $\theta^{(t)}$ – параметры, полученные после выполнения t EM-шагов. Покажите, что $\ell_{semi-sup}\left(\theta^{(t+1)}\right) \geq \ell_{semi-sup}\left(\theta^{(t)}\right)$.

Гибридный GMM

Теперь мы вернемся к модели гауссовской смеси (Gaussian Mixture Model, GMM) и применим к ней наш гибридный вариант EM-алгоритма. Рассмотрим сценарий, при котором наши данные порождаются $k \in \mathbb{N}$ распределениями Гаусса с неизвестными средними $\mu_j \in \mathbb{R}^d$ и ковариациями $\Sigma_j \in \mathbb{S}_+^d$, где $j \in \{1, \dots, k\}$. У нас есть m точек данных $x^{(i)} \in \mathbb{R}^d, i \in \{1, \dots, m\}$ и каждая из них имеет соответствующую скрытую (латентную) переменную $z^{(i)} \in \{1, \dots, k\}$, указывающую, к какому распределению принадлежит $x^{(i)}$. Более конкретно, $z^{(i)} \sim Multinomial(\phi)$, так что $\sum_{j=1}^{k} \phi_j = 1$ и $\phi_j \geq 0$ для всех j, а $x^{(i)}|z^{(i)} \sim \mathcal{N}(\mu_{z^{(i)}}, \Sigma_{z^{(i)}})$ независимые одинаково распределенные. Таким образом получаем параметры нашей модели: μ, Σ и ϕ.

У нас также есть дополнительные $\tilde{m}$ точек данных $\tilde{x}^{(i)} \in \mathbb{R}^d$, $i \in \{1, \dots, \tilde{m}\}$ и связанные с ними *известные* нам значения $\tilde{z}^{(i)} \in \{1, \dots, k\}$, указывающие, к какому распределению принадлежат $\tilde{x}^{(i)}$. Заметьте, что $\tilde{z}^{(i)}$ – это известные константы, в то время как $z^{(i)}$ – это неизвестные случайные величины. Как и ранее, мы предполагаем, что $\tilde{x}^{(i)}|\tilde{z}^{(i)} \sim \mathcal{N}(\mu_{\tilde{z}^{(i)}}, \Sigma_{\tilde{z}^{(i)}})$ независимые одинаково распределенные.

В итоге мы имеем суммарно $m + \tilde{m}$ примеров в выборке, из которых m являются неразмеченными, то есть со скрытыми z, и $\tilde{m}$ являются размеченными, для которых нам известны значения $\tilde{z}$. Классический EM-алгоритм разработан таким образом, что он берет только m неразмеченных примеров на вход и подгоняет параметры модели μ, Σ и ϕ.

Наша задача – применить гибридную версию EM-алгоритма к модели GMM, для того чтобы воспользоваться той дополнительной информацией, которую нам дают $\tilde{m}$ размеченных точек данных.

Вопрос №2. E-шаг (гибридный алгоритм) [2 балла]

Явно укажите все латентные переменные, которые необходимо рассчитывать на E-шаге. Выведите E-шаг для расчета всех указанных переменных. Ваше итоговое выражение может содержать только x, z, μ, Σ, ϕ и константы.

Вопрос №3. M-шаг (гибридный алгоритм) [3 балла]

Явно укажите все параметры, которые необходимо рассчитывать на M-шаге. Выведите M-шаг для расчета всех указанных параметров. Более конкретно, выведите выражения в закрытой форме для обновления значений параметров $\mu^{(t+1)}, \Sigma^{(t+1)}, \phi^{(t+1)}$, используя гибридную целевую функцию.

Вопрос №4. Классический EM-алгоритм [2 балла]

В этом вопросе мы будем работать только с m неразмеченными примерами. Следуйте инструкциям в gmm.py по реализации обычного классического EM-алгоритма и запустите его на неразмеченном датасете, чтобы он работал, пока не сойдется.

Сделайте три запуска алгоритма и с помощью предоставленной функции отрисуйте точечные графики, показывающие распределение примеров по кластерам (по одному графику на каждый запуск). Точки данных на графике должны быть раскрашены в цвета тех кластеров, к которым они были приписаны (т.е. выбирается тот кластер, для которого на заключительном E-шаге была найдена наибольшая вероятность).

Включите три графика в ваш отчет.

Вопрос №5. Гибридный EM-алгоритм [2 балла]

Теперь рассмотрим все примеры: и неразмеченные, и размеченные (по пять примеров на кластер). Мы предоставили стартовый код, разделяющий выборку на матрицы x и x_tilde, соответствующие неразмеченным и размеченным данным. Добавьте свой код в gmm.py для реализации гибридного EM-алгоритма и запустите его на датасете, чтобы он работал, пока не сойдется.

Сделайте три графика, как в предыдущем подзадании, и включите их в ваш отчет.

Вопрос №6. Сравнение классического и гибридного EM-алгоритмов [1 балл]

Кратко опишите, какие различия вы видите между двумя версиями алгоритма по следующим вопросам:

- количество итераций, требующееся для схождения алгоритма,
- стабильность (как сильно меняются кластера при случайной инициализации алгоритма),
- общее качество кластеризации.

Замечание. Выборка была сгенерирована смесью из трех маловариативных и одного сильновариативного нормальных распределений. Этот факт может помочь вам в определении качества работы двух алгоритмов.