EM-алгоритм: ещё ММП

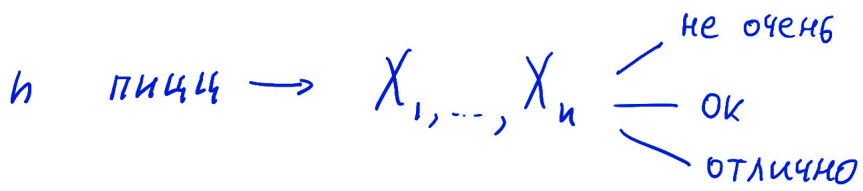
На этой неделе мы рассмотрим первую вероятностную модель, которую нельзя оценить методом максимального правдоподобия в обычном смысле в силу того, что не хватает нужных данных. Тем не менее, можно на основе принципа ММП построить итерационный алгоритм, который эту задачу поможет хорошо решить.

3.1 Скрытые переменные

В попытке построить вероятностную модель выборки X часто оказывается полезным учесть какой-то второй фактор, которого в данных нет. Конкретнее, предполагается, что есть ещё некоторые величины Y_i которые как-то влияют на распределение X_i и учёт подобного влияния позволяет построить, с одной стороны, технически более логичную модель из простых блоков, а с другой – лучше объяснить происхождение данных с позиции опыта и здравого смысла. Рассмотрим такой пример.

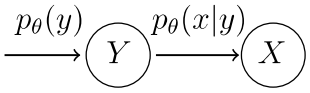
Пример 3.1. *В конце года после экзамена есть хорошая традиция собраться вместе и пообщаться за пиццей и настолками. В такой ситуации всегда надо заказывать достаточно много пицц. Конечно, коробки разных пиццерий стараются делать разными, но иногда это не очень выходит или нужно смотреть на какие-то совсем мелкие детали, вертеть в руках коробку, чтобы отличить одну пиццерию от другой.*

Представим себе ситуацию. Было перепробовано n похожих пицц из двух пиццерий, про каждую разные группы людей сошлись во мнении: не очень (0), ок(1) или отлично(2). Можем ли мы понять, не заглядывая в чеки, какова была доля пицц из пиццерии Гого-Пицца и БотманПицца? Но ещё, можно ли понять как по качеству каждая пиццерия готовит разные пиццы (распределение по оценкам не очень, ок и отлично)?



В данном примере формально мы наблюдаем только независимую выборку из дискретных оценок пицц $X_1, .., X_n$. Что мы можем по ней сказать? Мы можем в целом понять, какая доля пицц какой оценке соответствует; доля каждой в выборке – оценка максимального правдоподобия для распределения с конечным числом значений. Можем пробовать проверять гипотезы с помощью тестов максимального правдоподобия (Вальд, LR, LM). При этом учитывая простоту модели, оценки ММП получатся удовлетворительные даже при относительно небольших выборках (несколько десятков). На этом список наших возможностей заканчивается.

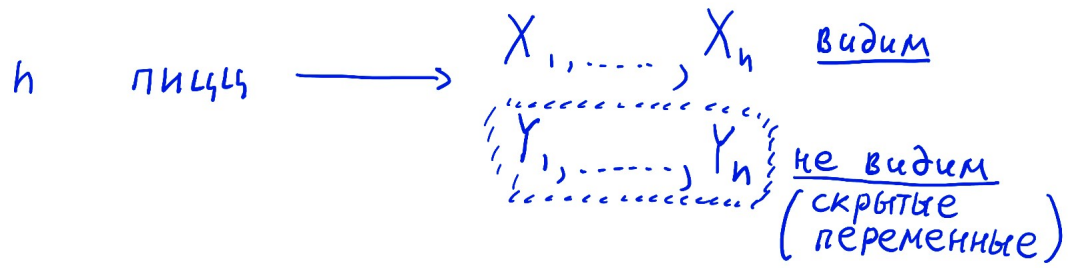
На вопрос из примера так мы не сможем ответить; для этого нужно пересмотреть вероятностную модель. По сути у нас есть две переменных, про которые хочется что-то узнать: название пиццерии Y и качество пиццы X. При этом рапределение по качеству в разных пиццериях может быть разным. Давайте предложим следующую идею:



Все переменные X_i, а также все Y_i независимы в совокупности, но X_i и Y_i зависимы: мы предполагаем что сначала семплируется пиццерия Y, а затем на основе результата семплируется пицца. Здесь мы использовали сокращённую нотацию, обозначив за θ все параметры вероятностной модели, а p_θ обозначает конкретное распределение (вероятность или плотность), которое понятно из контекста. У нас всего две пиццерии, поэтому вначале семплируется Y из распределения $Ber(p)$, затем семплируется качество трёх видов из распределения на конечном множестве с параметрами $\alpha_1, \alpha_2, 1 - \alpha_1 - \alpha_2$ (для пиццерии ГоГоПицца) и $\beta_1, \beta_2, 1 - \beta_1 - \beta_2$ (для пиццерии БотманПицца). Так $\theta = [p, \alpha_1, \alpha_2, \beta_1, \beta_2]$, все X получены независимо прогоном через такую цепочку семплирования.

3.2 EM-алгоритм

Как оценивать такую модель? У нас есть только выборка $X_1, .., X_n$, при этом модель включает в себя *скрытые* $Y_1, .., Y_n$, которые в данных не наблюдаются.



EM-алгоритм [3] решает эту проблему, предлагая интересный подход на основе метода максимального правдоподобия, но с естественным способом адресовать неопределённость вокруг набора Y. Помните, мы недавно рассматривали метод максимального правдоподобия как попытку минимизировать аппроксимированную отрицательную кросс-энтропию?

$$-CE(p_{\theta_0}||p_\theta) = \mathbb{E}_{\theta_0}\left[\ln p_\theta(X)\right] \approx \frac{1}{n}\mathbb{E}_{\theta_0}\left[\ln p_\theta(X)|X\right] = \frac{1}{n}l(\theta, X).$$

Мы будем опускать константу $1/n$ как несущественную для оптимизации. В основе знака приближённого равенства лежала идея того, что мы можем в условие матожидания поставить то, что известно из данных, а по остальному взять матожидание. В нашем случае со скрытыми переменными мы получим с тем же подходом

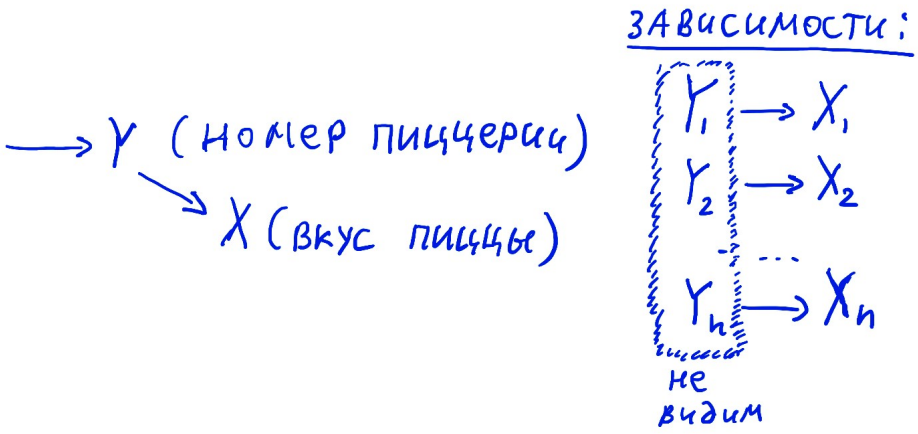
$$\mathbb{E}_{\theta_0}\left[\ln p_\theta(X, Y)\right] \approx \mathbb{E}_{\theta_0}\left[\ln p_\theta(X, Y)|X\right],$$

где при известных θ_0 мы могли бы явно посчитать это выражение и прооптимизировать по θ.

В этом смысле EM-алгоритм предлагает следующее:

1. Задать произвольный стартовый θ_0;
2. (E-шаг) Посчитать матожидание справа;
3. (M-шаг) Максимизировать $\mathbb{E}_{\theta_0}\left[\ln p_\theta(X, Y)|X\right]$ по θ, получив $\hat{\theta}$;
4. Задать $\theta_0 = \hat{\theta}$ и повторить с пункта 2.

3.3 Где пицца лучше?



Давайте вернёмся к примеру. Модель была устроена так: сначала семплируется пиццерия (из 0,1), затем семплируется качество пиццы (из $0, 1, 2$). Предположим, что параметр $\theta_0 = [p^0, \alpha_1^0, \alpha_2^0, \beta_1^0, \beta_2^0]$ мы задали, тогда первая наша задача – вычислить аналитически матожидание

$$Q(\theta_0, \theta) := \mathbb{E}_{\theta_0} \left[\ln p_\theta(X, Y) | X \right].$$

Для этого воспользуемся сначала независимостью по индексу наблюдения:

$$Q(\theta_0, \theta) = \mathbb{E}_{\theta_0} \left[\sum_{i=1}^{n} \ln p_\theta(X_i, Y_i) \ \ | \, X \right].$$

Здесь нам нужно посчитать условное матожидание; обозначим за x_i реализацию выборки и заметим, что для этого нужно знать условные вероятности

$$p_{\theta_0}(Y_i = y | X_i = x_i) = \frac{p_{\theta_0}(X_i = x_i | Y_i = y) p_{\theta_0}(Y_i = y)}{p_{\theta_0}(X_i = x_i)} =$$
$$= \frac{p_{\theta_0}(X_i = x_i | Y_i = y) p_{\theta_0}(Y_i = y)}{p_{\theta_0}(X_i = x_i | Y_i = 0) p_{\theta_0}(Y_i = 0) + p_{\theta_0}(X_i = x_i | Y_i = 1) p_{\theta_0}(Y_i = 1)} =: \gamma_{Y_i}(y).$$

В силу того, что нам дан θ_0, всё это можно вычислить. К примеру, для события $Y_i = 0, X_i = 1$

$$\gamma_{Y_i}(0) = p_{\theta_0}(Y_i = 0 | X_i = 1) = \frac{\alpha_2^0 (1 - p^0)}{\alpha_2^0 (1 - p^0) + \beta_2^0 p^0}.$$

Вернёмся в Q; в силу независимости по индексу i мы можем, используя только что полученное, переписать матожидание в другом виде:

$$\mathbb{E}_{\theta_0} \left[\sum_{i=1}^{n} \ln p_\theta(X_i, Y_i) \ \ | \, X \right] = \sum_{i=1}^{n} \gamma_{Y_i}(0) \ln \left(p_\theta(x_i | Y_i = 0) p_\theta(Y_i = 0) \right) + \gamma_{Y_i}(1) \ln \left(p_\theta(x_i | Y_i = 1) p_\theta(Y_i = 1) \right)$$

и более компактно получаем

$$Q(\theta_0, \theta) = \sum_{i=1}^{n} \gamma_{Y_i}(0) \ln \left(p_\theta(x_i | Y_i = 0)(1 - p) \right) + \gamma_{Y_i}(1) \ln \left(p_\theta(x_i | Y_i = 1) p \right).$$

Далее для М-шага нужно максимизировать Q по θ, но это оказывается не очень сложно, так как переменные p и α, β разделяются на отдельные суммы – по p и $\alpha_1, \alpha_2, \beta_1, \beta_2$ можно оптимизировать отдельно. Из-за весов γ_{Y_i} решение будет отличаться от привычного по методу максимального правдоподобия. Для краткости введём

$$\Gamma_0 = \sum_{i=1}^{n} \gamma_{Y_i}(0), \quad \Gamma_1 = \sum_{i=1}^{n} \gamma_{Y_i}(1),$$

тогда в итоге получим оценки

$$\begin{aligned}
&\hat{p} = \frac{\Gamma_1}{n}, \\
&\hat{\alpha}_1 = \frac{1}{\Gamma_0} \sum_{i=1}^{n} \gamma_{Y_i}(0) \mathbb{1}(X_i = 0), \ \hat{\alpha}_2 = \frac{1}{\Gamma_0} \sum_{i=1}^{n} \gamma_{Y_i}(0) \mathbb{1}(X_i = 1), \\
&\hat{\beta}_1 = \frac{1}{\Gamma_1} \sum_{i=1}^{n} \gamma_{Y_i}(1) \mathbb{1}(X_i = 0), \ \hat{\beta}_2 = \frac{1}{\Gamma_1} \sum_{i=1}^{n} \gamma_{Y_i}(1) \mathbb{1}(X_i = 1).
\end{aligned}$$

Коэффциенты γ_{Y_i} выступают в роли весов: чем правдоподобнее полученные наблюдения в рамках модели θ_0, тем больше вес и, следовательно, соответствующая компонента смеси будет играть большую роль при оптимизации.
Далее, согласно EM-алгоритму, нужно задать $\theta_0 = \hat{\theta}$ и повторять, пока мы не поймём, что сошлись. Вопрос сходимости EM-алгоритма мы оставим за рамками, в общем случае сходимость не гарантирована. Всё же...

Утверждение 3.1. *С ростом итерации θ_k EM-алгоритма правдоподобие модели при данных $x_1, .., x_n$ улучшается: для каждого наблюдения x*

$$p_{\theta_k}(x) \geq p_{\theta_{k-1}}(x).$$

▷ Заметим, что из определения условных вероятностей следует также и

$$p_\theta(x) = \frac{p_\theta(x, y)}{p_\theta(y|x)},$$

а взяв логарифм, мы получим

$$\ln p_\theta(x) = \ln p_\theta(x, y) - \ln p_\theta(y|x).$$

Теперь мы можем просуммировать (или проинтегрировать) обе части по y по распределению $p_{\theta_{k-1}}(y|x)$:

$$\sum_j p_{\theta_{k-1}}(Y = j|x) \ln p_\theta(x) = \sum_j p_{\theta_{k-1}}(Y = j|x) \left(\ln p_\theta(x, Y = j) - \ln p_\theta(Y = j|x) \right).$$

Заметим, что сумма с первым слагаемым – в точности $Q(\theta_{k-1}, \theta)$. Если подставим $\theta = \theta_{k-1}$, то получим другое равенство:

$$\sum_j p_{\theta_{k-1}}(Y = j|x) \ln p_{\theta_{k-1}}(x) = \sum_j p_{\theta_{k-1}}(Y = j|x) \left(\ln p_{\theta_{k-1}}(x, Y = j) - \ln p_{\theta_{k-1}}(Y = j|x) \right).$$

Обратите внимание на левые части: логарифм выносится, а сумма равна единице. Теперь если вычтем из первого равенства второе, то получим

$$\ln p_\theta(x) - \ln p_{\theta_{k-1}}(x) = Q(\theta_{k-1}, \theta) - Q(\theta_{k-1}, \theta_{k-1}) - \sum_j p_{\theta_{k-1}}(Y = j|x) \ln \frac{p_\theta(Y = j|x)}{p_{\theta_{k-1}}(Y = j|x)}.$$

Для суммы справа мы можем использовать неравенство Йенсена для логарифма, тогда получим неравенство

$$\ln p_\theta(x) - \ln p_{\theta_{k-1}}(x) \geq Q(\theta_{k-1}, \theta) - Q(\theta_{k-1}, \theta_{k-1}) - \ln \sum_j p_{\theta_{k-1}}(Y = j|x) \frac{p_\theta(Y = j|x)}{p_{\theta_{k-1}}(Y = j|x)} =$$

$$= Q(\theta_{k-1}, \theta) - Q(\theta_{k-1}, \theta_{k-1}).$$

Так, мы получили неравенство

$$\ln p_\theta(x) - \ln p_{\theta_{k-1}}(x) \geq Q(\theta_{k-1}, \theta) - Q(\theta_{k-1}, \theta_{k-1}).$$

Теперь заметим, что согласно нашей конструкции $\theta_k = \arg\max_\theta Q(\theta_{k-1}, \theta)$, поэтому $Q(\theta_{k-1}, \theta_k) \geq Q(\theta_{k-1}, \theta_{k-1})$ и подставив $\theta = \theta_k$, получим в результате

$$\ln p_{\theta_k}(x) - \ln p_{\theta_{k-1}}(x) \geq 0.$$

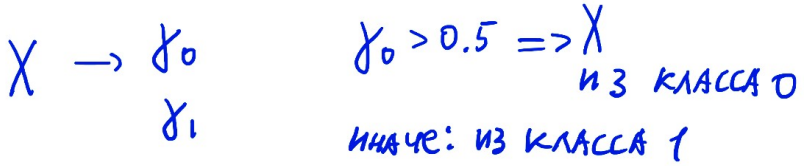
□

Поскольку лог-правдоподобие выборки в нашем случае в силу независимости раскладывается в произведение, то этот же результат верен и для выборки.

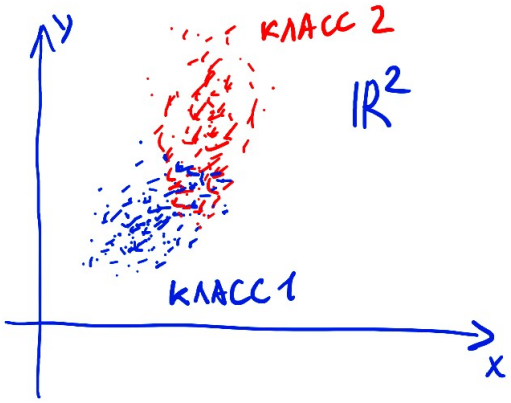
3.4 Чем полезна модель смеси?

Остановимся и подумаем: что можно узнать *после* того, как мы оценили параметры смеси?

При детальном взгляде оказывается, что вес $\gamma_{Y_i}(k)$ как раз отражает вероятность того, что пицца i принадлежит компоненте смеси(пиццерии) номер k. При этом такой вес мы можем вычислить и для новых наблюдений, так как веса зависят от известных посчитанных параметров θ_0. То есть, модели смесей можно использовать для кластеризации.

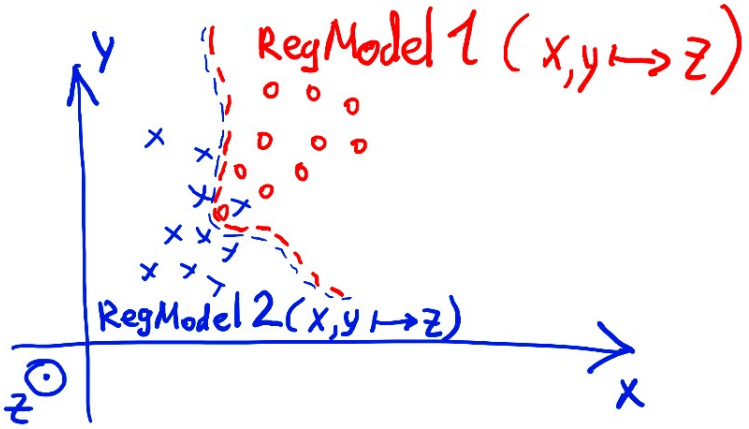


Мы можем свободно изменить модель одной компоненты, поставив другое более подходящее распределение с подходящей моделью параметров, например, гауссовское; такая модель носит название *гауссовской смеси* (Gaussian Mixture, например, реализован в `sklearn`). Тогда мы сможем делать кластеризацию точек в $\mathbb{R}^d$.



Как в кластеризации, мы можем попробовать решать задачу классификации: относить пиццы к одной или к другой пиццерии в зависимости от весов γ. Но тут есть некоторая проблема: номер компоненты не обязательно совпадает с номером пиццерии. Если в инициализации параметров мы поменяем местами параметры α_1, α_2 и β_1, β_2, в результате компоненты 0 и 1 тоже поменяются местами. Если бы мы имели дополнительную информацию, например, кто-то из организаторов сказал бы: "Кажется, мы больше заказывали из ГогоПицца," – то по результату EM-алгоритма мы бы сразу соотнесли ГогоПиццу с компонентой смеси с большим количеством наблюдений, а БотманПиццу – с компонентной с меньшим числом наблюдений.

Можем ли мы пробовать решать задачу регрессии с помощью смеси? Тоже можно, если задать более конкретную параметризацию матожидания в компоненте гауссовской смеси.



Литература

[1] Peter C Austin, Muhammad M Mamdani, David N Juurlink, and Janet E Hux. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, 59(9):964–969, July 2006.

[2] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, Sep 2011.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[4] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[5] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[6] S.J. Levitt, S.D. Dubner. *Freakonomics*. NY: Harper Trophy, 2006.

[7] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[8] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[9] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[10] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных*. Изд. дом Высшей школы экономики, 2012.

[11] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования*. М.: МЦНМО, 2013.