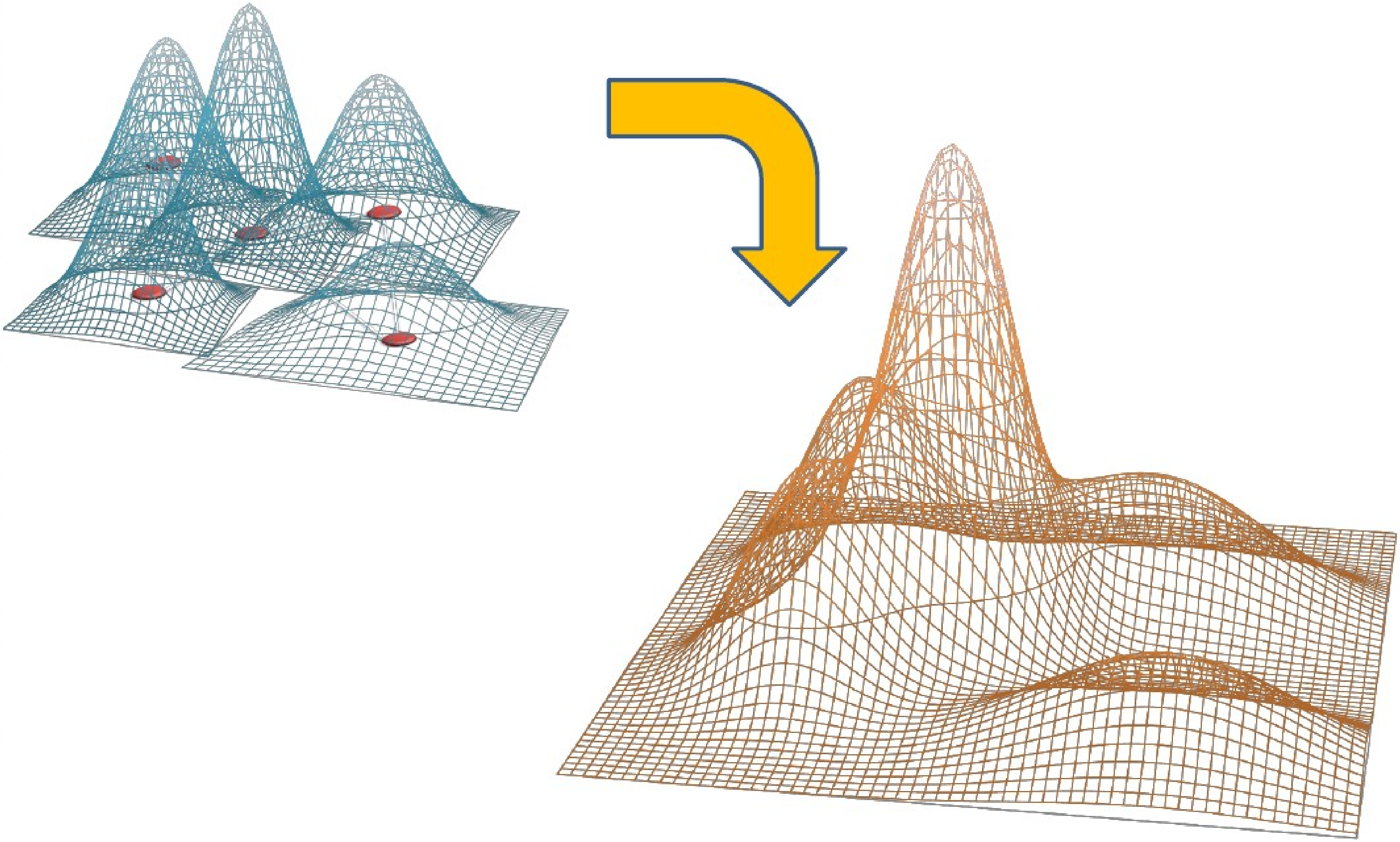
Машинное обучение
Логистическая регрессия.
Смеси распределений



Содержание лекции

- Логистическая регрессия
- Бинаризация признаков
- Скоринг
- Смеси распределений
- ЕМ-алгоритм восстановления смеси

Предположение 1

- $X=R^n$, $Y=\{+1,-1\}$, X^ℓ
- Распределение $p(x|y)$ из экспоненциального семейства:

  $p(x|y)=\exp(\ c(\delta)\langle\theta_y,x\rangle+b_y(\delta,\theta_y)+d(x,\delta)\ )$

  $\theta_y \in R^n$ – параметр сдвига

  δ – параметр разброса

  b_y,c,d – произвольные числовые функции
- Экспоненциальное семейство распределений широко: равномерное, нормальное, Лапласа, Пуассона, Парето, Дирихле, биномиальное, Г-распределение, χ^2-распределение, и др.

Предположение 2

- Плотности p(x|y) имеют равные значения параметров c, d и δ, но отличаются значениями параметра сдвига θ_y.

Теорема

Если выполняются предположения 1 и 2 и среди признаков есть константа, то

- оптимальный байесовский классификатор для заданных штрафов λ_+ и λ_- является линейным:

$$a(x) = \mathrm{sign}\big(\langle w, x\rangle - w_0\big)$$

- апостериорные вероятности классов вычисляются по формуле: $\mathsf{P}(y|x) = \sigma\big(\langle w, x\rangle\, y\big)$

где $\sigma(z) = \frac{1}{1+e^{-z}}$ - логистическая функция

Доказательство

$$a(x) = 1 \quad \Leftrightarrow \quad \lambda_+ P(+1|x) > \lambda_- P(-1|x) \quad \Leftrightarrow \quad \frac{P(+1|x)}{P(-1|x)} > \frac{\lambda_-}{\lambda_+}$$

$$\Leftrightarrow \quad \frac{P(x|+1)P(+1)}{P(x|-1)P(-1)} > \frac{\lambda_-}{\lambda_+} \quad \Leftrightarrow \quad \ln \frac{P(x|+1)P(+1)}{P(x|-1)P(-1)} > \ln \frac{\lambda_-}{\lambda_+}$$

подставим сюда $p(x|\pm 1) = \exp\left(c_\pm(\delta)\langle\theta_\pm, x\rangle + b_\pm(\delta, \theta_\pm) + d(x, \delta)\right)$

$$\ln \frac{\mathrm{P}(+1|x)}{\mathrm{P}(-1|x)} = \langle \underbrace{c(\delta)(\theta_+ - \theta_-)}_{w=\mathrm{const}(x)}, x\rangle + \underbrace{b_+(\delta, \theta_+) - b_-(\delta, \theta_-) + \ln \frac{P_+}{P_-}}_{\beta=\mathrm{const}(x)}$$

Добавим β к коэффициенту w_j при константном признаке $f_j = 1$

Доказательство

Получим:

$$\frac{\mathsf{P}(+1|x)}{\mathsf{P}(-1|x)} = e^{\langle w,x\rangle}$$

По формуле полной вероятности $\mathsf{P}(-1|x) + \mathsf{P}(+1|x) = 1$,

следовательно

$$\mathsf{P}(+1|x) = \frac{1}{1+e^{-\langle w,x\rangle}}; \qquad \mathsf{P}(-1|x) = \frac{1}{1+e^{\langle w,x\rangle}}$$

$$\mathsf{P}(y|x) = \frac{1}{1+e^{-\langle w,x\rangle y}} = \sigma\big(\langle w,x\rangle y\big)$$

Разделяющая классы поверхность – линейна:

$$\lambda_-\,\mathsf{P}(-1|x) = \lambda_+\,\mathsf{P}(+1|x),$$

$$\langle w,x\rangle - \ln\frac{\lambda_-}{\lambda_+} = 0.$$

Поиск w

- Максимизация логарифма правдоподобия обучающей выборки:

$$\ln \prod_{i=1}^{\ell} p(x_i, y_i) = \sum_{i=1}^{\ell} \ln p(x_i, y_i) \to \max_{\mathrm{w}}$$

- Для логистического распределения:

$p(x, y) = p(y|x)p(x) = \sigma(\langle \mathrm{w}, x\rangle y)p(x)$ отсюда

$$\sum_{i=1}^{\ell} \ln\left(1 + e^{-\langle \mathrm{w}, x_i\rangle y_i}\right) + \mathrm{Const}(\mathrm{w}) \to \min_{\mathrm{w}}$$

- Напоминает минимизацию функционала эмпирического риска $Q(a, X^\ell) = \frac{1}{\ell}\sum_{i=1}^{\ell} \mathscr{L}(a, x_i)$

Сравнение с другими видами функционала эмпирического риска

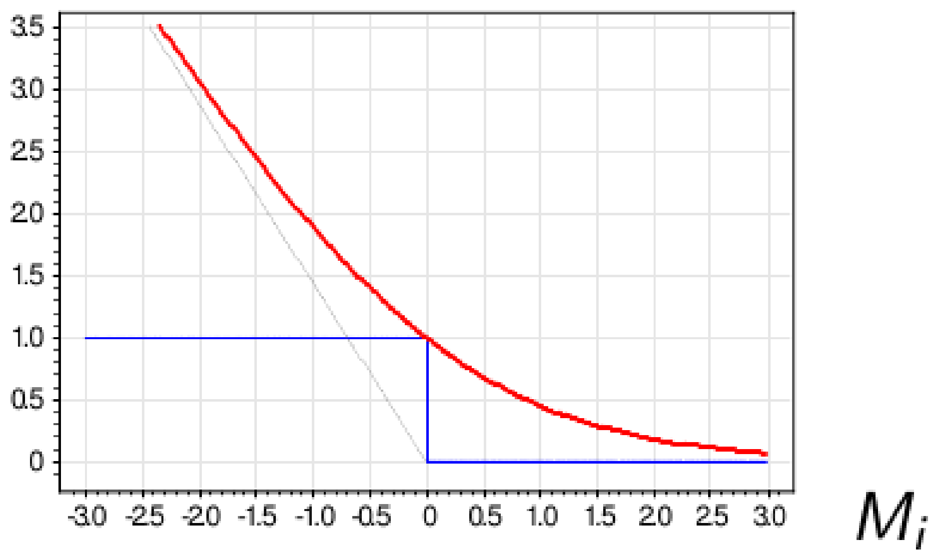
- Определим выступ объекта как:

$$M_i = \langle \mathrm{w}, x_i \rangle y_i$$

- В случае логистической регрессии:

$$\mathscr{L}(M) = \ln\left(1 + e^{-M}\right)$$

- Сравним:



Поиск w

Метод первого порядка — стохастический градиент:

$$w^{(t+1)} := w^{(t)} + \eta_t\, y_i x_i \,(1 - \sigma_i),$$

η_t — градиентный шаг,
$\sigma_i = \sigma(y_i w^{\mathsf{T}} x_i) = \mathsf{P}(y_i|x_i)$ — вероятность правильной классификации x_i.

Метод второго порядка (Ньютона-Рафсона) приводит к IRLS, Iteratively Reweighted Least Squares:

$$w^{(t+1)} := w^{(t)} + \eta_t (F^{\mathsf{T}} \Lambda F)^{-1} F^{\mathsf{T}} \tilde{y},$$

F — матрица объекты–признаки $\ell \times n$,
$\tilde{y} = \big(y_i(1 - \sigma_i)\big)$,
$\Lambda = \mathrm{diag}\big((1 - \sigma_i)/\sigma_i\big)$,

Бинаризация признаков (One Hot encoding)

- Пусть x - единственный признак (номинальный, закодированный: 0,1,2,...k)
- Классификатор: $a(x) = \text{sign}(wx+w_0)$
- Проблема: вес w нельзя подобрать так, чтобы классификатор был не монотонным.
- Для любых w и w0 значения $a(x) > 0$ когда $x > w_0/w$ и $a(x) \leq 0$ в противном случае

Бинаризация признаков (One Hot encoding)

- Вместо одного номинального признака вводим k бинарных признаков. Пример (k=5):

	x1	x2	x3	x4	x5
Азов	0	0	0	0	1
Аксай	0	0	0	1	0
Ростов	0	0	1	0	0
Новочеркасск	0	1	0	0	0
Таганрог	1	0	0	0	0

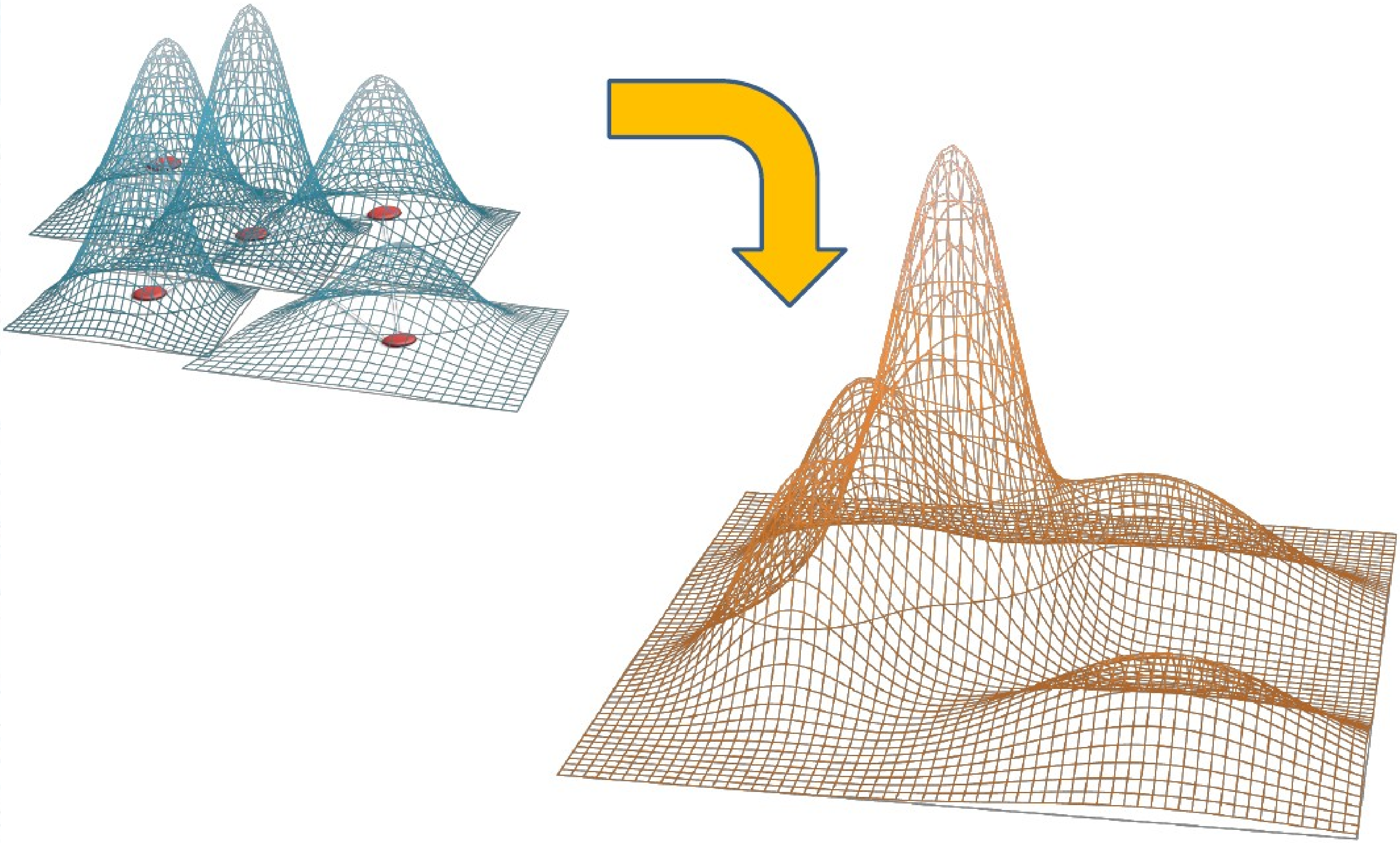
- Возможна бинаризация и количественных признаков путем предварительной дискретизации

Скоринг

- Если все признаки – бинарные, то линейный классификатор удобно рассматривать как суммирование баллов (score): Sum += w_j, если x_j=1
- Рисунок – фрагмент скоринговой карты для вопроса о выдаче кредита

Возраст	до 25	5
	25 - 40	10
	40 - 50	15
	50 и больше	10
Собственность	владелец	20
	совладелец	15
	съемщик	10
	другое	5
Работа	руководитель	15
	менеджер среднего звена	10
	служащий	5
	другое	0
Стаж	1/безработный	0
	1..3	5
	3..10	10
	10 и больше	15
Работа_мужа /жены	нет/домохозяйка	0
	руководитель	10
	менеджер среднего звена	5
	служащий	1

Смеси распределений



$$p(x) = \sum_{j=1}^{k} w_j p_j(x;\theta_j), \qquad \sum_{j=1}^{k} w_j = 1, \qquad w_j \geqslant 0$$

Смеси распределений

Задача 1: имея простую выборку $X^m \sim p(x)$ и зная k, оценить вектор параметров $\Theta = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$.

Задача 2: оценить ещё и k.

Задача максимизации логарифма правдоподобия

$$L(\Theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i; \theta_j) \to \max_{\Theta}$$

при ограничениях $\sum_{j=1}^{k} w_j = 1; \;\; w_j \geqslant 0.$

Решение оптимизационной задачи

$$Q(\Theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j p_j(x_i) \to \max_{\Theta} \qquad \textstyle\sum_{j=1}^{k} w_j = 1$$

$$L(\Theta; X^m) = \sum_{i=1}^{m} \ln\left(\sum_{j=1}^{k} w_j p_j(x_i)\right) - \lambda\left(\sum_{j=1}^{k} w_j - 1\right)$$

$$\frac{\partial L}{\partial w_j} = \sum_{i=1}^{m} \frac{p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} - \lambda = 0, \quad j = 1, \ldots, k$$

Умножим левую и правую части на w_j , просуммируем все k этих равенств, и поменяем местами знаки суммирования по j и по i:

$$\sum_{i=1}^{m} \underbrace{\sum_{j=1}^{k} \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)}}_{=1} = \lambda \underbrace{\sum_{j=1}^{k} w_j}_{=1}, \quad \Rightarrow \quad \lambda = m$$

Решение оптимизационной задачи

$$w_j = \frac{1}{m} \sum_{i=1}^{m} \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} = \frac{1}{m} \sum_{i=1}^{m} g_{ij}, \quad j = 1, \ldots, k$$

где $g_{ij} = \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)}$ "похожи" на вероятности того, что x_i попал в j-тое скопление смеси:

$$g_{ij} = \mathrm{P}(j|x_i) = \frac{P(j)p(x_i|j)}{p(x_i)} = \frac{w_j p_j(x_i;\theta_j)}{p(x_i)} = \frac{w_j p_j(x_i;\theta_j)}{\sum_{s=1}^{k} w_s p_s(x_i;\theta_s)}$$

$$\sum_{j=1}^{k} g_{ij} = 1$$

Решение оптимизационной задачи

Приравняем к нулю производную лагранжиана по θ_j, помня, что $p_j(x) = \varphi(x;\theta_j)$:

$$\frac{\partial L}{\partial \theta_j} = \sum_{i=1}^{m} \frac{w_j}{\sum_{s=1}^{k} w_s p_s(x_i)} \frac{\partial}{\partial \theta_j} p_j(x_i) = \sum_{i=1}^{m} \frac{w_j p_j(x_i)}{\sum_{s=1}^{k} w_s p_s(x_i)} \frac{\partial}{\partial \theta_j} \ln p_j(x_i) =$$
$$= \sum_{i=1}^{m} g_{ij} \frac{\partial}{\partial \theta_j} \ln p_j(x_i) = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} g_{ij} \ln p_j(x_i) = 0, \quad j = 1, \dots, k.$$

Полученное условие совпадает с необходимым условием максимума в задаче максимизации взвешенного правдоподобия:

$$\theta_j := \arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i;\theta), \quad j = 1, \dots, k.$$

при условии, что g_{ij} не зависят от θ. Что, конечно же, не так.

ЕМ-алгоритм

Итерационный алгоритм Expectation–Maximization:

1: начальное приближение вектора параметров Θ;

2: **повторять**

3: $\quad G :=$ E-шаг (Θ); // оцениваются *скрытые переменные* G

4: $\quad \Theta :=$ M-шаг (Θ, G);

5: **пока** Θ и G не стабилизируются.

EM-алгоритм

Вход: $X^m = \{x_1, \dots, x_m\}$, k, δ, начальное $\Theta = (w_j, \theta_j)_{j=1}^k$;

Выход: $\Theta = (w_j, \theta_j)_{j=1}^k$ — параметры смеси распределений

1: **повторять**

2: Е-шаг (expectation):

для всех $i = 1, \dots, m$, $j = 1, \dots, k$

$$g_{ij}^0 := g_{ij}; \qquad g_{ij} := \frac{w_j p_j(x_i; \theta_j)}{\sum_{s=1}^k w_s p_s(x_i; \theta_s)};$$

3: М-шаг (maximization):

для всех $j = 1, \dots, k$

$$\theta_j := \arg\max_\theta \sum_{i=1}^m g_{ij} \ln p_j(x_i; \theta); \qquad w_j := \frac{1}{m} \sum_{i=1}^m g_{ij};$$

4: **пока** $\max\limits_{i,j} |g_{ij} - g_{ij}^0| > \delta$;

5: **вернуть** $(w_j, \theta_j)_{j=1}^k$;

Смеси гауссовских распределений

$$p(x|y) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = \mathcal{N}(x; \mu_{yj}, \Sigma_{yj})$$

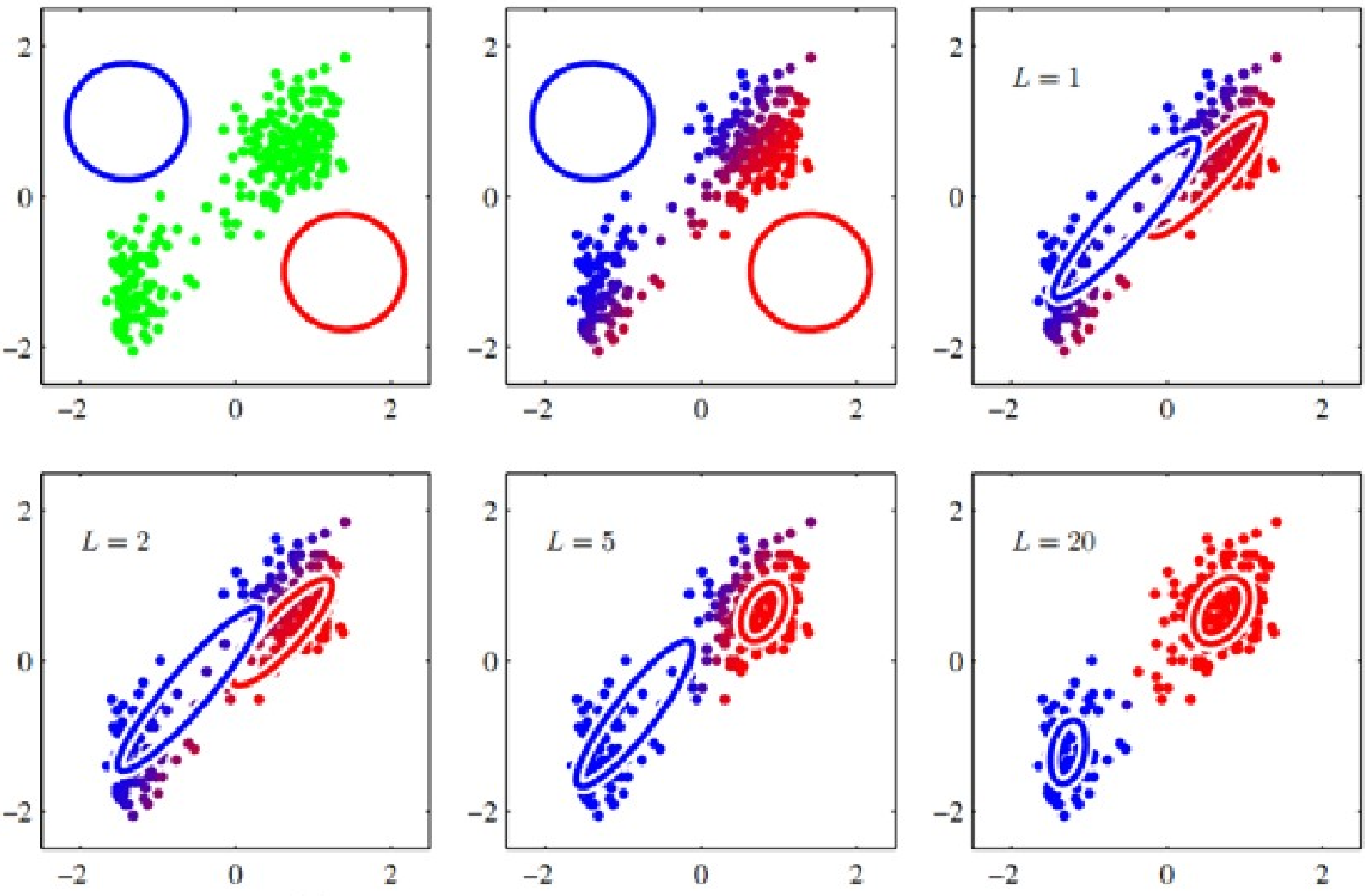
Решение М-шага:

$$\hat{\mu}_j = \frac{1}{m w_j} \sum_{i=1}^{m} g_{ij} x_i, \quad j = 1, \ldots, k;$$

$$\hat{\Sigma}_j = \frac{1}{m w_j} \sum_{i=1}^{m} g_{ij} (x_i - \hat{\mu}_j)(x_i - \hat{\mu}_j)^{\mathsf{T}}, \quad j = 1, \ldots, k.$$

Пример работы алгоритма

- Две гауссовские компоненты k = 2 в R^2.
- Расположение компонент в зависимости от номера итерации:



ЕМ-алгоритм с добавлением и удалением компонент

- Проблемы базового варианта ЕМ-алгоритма:
  - Как выбирать начальное приближение?
  - Как определять число компонент?
  - Как ускорить сходимость?
- Добавление и удаление компонент в ЕМ-алгоритме:
  - Если слишком много объектов x_i имеют слишком низкие правдоподобия $p(x_i)$, то создаём новую k+1-ю компоненту, по этим объектам строим её начальное приближение.
  - Если у j-й компоненты слишком низкий w_j, удаляем её.

Вопросы для самоконтроля

- Пусть плотности вероятности $p(x|y=+1)$ и $p(x|y=-1)$ имеют нормальные распределения (разные для $y=+1$ и $y=-1$). При каких условиях будет справедлива теорема о логистической регрессии?
- Сгенерируйте случайную выборку из 10 объектов с классами +1 и -1 и двумя признаками: x_1, x_2. Придумайте вектор $w=(w_1,w_2)$. Вычислите выступ для каждого объекта выборки

Вопросы для самоконтроля

- Сгенерируйте случайную выборку из 10 объектов с классами +1 и -1 и одним количественным признаком со значениями на интервале [0,20]. По выборке должно быть видно, что точки формируют два отдельных скопления
- Нарисуйте ее график рассеяния (одномерный) и подумайте, какое начальное приближение можно взять для этой выборки в EM-алгоритме разделения смеси из двух нормальных распределений. Задайте веса компонент и параметры распределений.

Вопросы для самоконтроля

- Сгенерируйте случайную выборку из 10 объектов с одним количественным признаком со значениями на интервале [0,20].
- Возьмите в качестве плотностей распределения двух компонент pj(x) два равномерных распределения на разных интервалах. Вычислите коэффициенты gij из EM-алгоритма