Смеси экспертов

Грабовой Андрей Валериевич

Московский физико-технический институт

МФТИ, г. Долгопрудный

План

1. EM–алгоритм:
   - Классический,
   - Вариационный,
2. Смесь моделей:
   - Постановка задачи,
   - Итерационные формулы полученные при помощи вариационного EM–алгоритма,
   - Илюстрация сходимости,
3. Смесь экспертов
   - Постановка задачи,
   - Итерационные формулы полученные при помощи вариационного EM–алгоритма,
   - Илюстрация сходимости,

Максимизация обоснованости:

$$\boldsymbol{\Theta} = \arg\max_{\boldsymbol{\Theta}} p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}\right) \tag{1}$$

ELBO:

$$\begin{aligned}\mathcal{L}\left(q(\mathbf{Z}), \boldsymbol{\Theta}\right) &= \int q(\mathbf{Z}) \log \frac{p\left(\mathbf{y}, \mathbf{Z}|\boldsymbol{\Theta}, \mathbf{X}\right)}{q\left(\mathbf{Z}\right)} d\mathbf{Z} \\ &= p\left(\mathbf{y}|\mathbf{X}, \boldsymbol{\Theta}\right) - \mathsf{D}_{KL}\left(q\big(\mathbf{Z}\big)||p\big(\mathbf{Z}|\mathbf{y}, \mathbf{X}, \boldsymbol{\Theta}\big)\right)\end{aligned} \tag{2}$$

EM–алгоритм:

1. E–шаг:

$$q^{s}\big(\mathbf{Z}\big) = \arg\max_{q\left(\mathbf{z}\right)\in Q} \mathcal{L}\left(q\big(\mathbf{Z}\big), \boldsymbol{\Theta}^{s-1}\right) \tag{3}$$

2. M–шаг:

$$\boldsymbol{\Theta}^{s} = \arg\max_{\boldsymbol{\Theta}} \mathcal{L}\left(q^{s}\big(\mathbf{Z}\big), \boldsymbol{\Theta}\right) \tag{4}$$

Вариационный EM–алгоритм (Mean Field Approximation[1]):

1. E–шаг:

$$\log q\left(\mathbf{Z}_k^s\right) \propto \mathsf{E}_{q/k} \log p\left(\mathbf{y}, \mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta}^{s-1}\right) \tag{5}$$

2. M–шаг:

$$\boldsymbol{\Theta}^{s} = \arg\max_{\boldsymbol{\Theta}} \mathbf{E}_{q^s} \log p\left(\mathbf{y}, \mathbf{Z}|\mathbf{X}, \boldsymbol{\Theta}\right) \tag{6}$$

[1] https://github.com/andriygav/EMprior/blob/master/Lecture/Grabovoy2019MeanField.pdf

Definition

Смесь моделей — мультимодель, ответы которой представляют собой взвешенную сумму ответов всех задействованных моделей независимо от объекта.

$$\hat{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k, \qquad \pi_k = const, \quad \sum_{k=1}^{K} \pi_k = 1, \tag{7}$$

где $\mathbf{f}$ — мультимодель, а $\mathbf{f}_k$ — локальная модель.

Пример 1:

1. Веса моделей в смеси $\boldsymbol{\pi}$ получены из априорного распределения
$$p\left(\boldsymbol{\pi}|\mu\right); \tag{8}$$
2. Вектора параметров $\mathbf{w}_k$ получены из нормального распределения
$$p\left(\mathbf{w}_k|\mathbf{A}_k\right) = \mathcal{N}\left(\mathbf{w}_k|\mathbf{0}, \mathbf{A}_k\right), \ k = 1, \cdots K; \tag{9}$$
3. Для каждого объекта $\mathbf{x}_i$ существует модель $\mathbf{f}_{k_i}$, которой он описывается, причем $p\left(k_i = k\right) = \pi_k$;
4. Для каждого объекта $\mathbf{x}_i$ класс y_i определен в соответсвии с моделью
$$\mathbf{f}_{k_i} : \ y_i \sim \ \mathcal{N}\left(\mathbf{w}_{k_i}^{\mathsf{T}}\mathbf{x}_i + b_k, \beta^{-1}\right) \tag{10}$$

Правдоподобие модели:

$$p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}|\mathbf{X}, \mathbf{A}, \beta, \boldsymbol{\mu}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\mu}) \prod_{k=1}^{K} N(\mathbf{w}_k|\mathbf{0}, \mathbf{A}_k) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(y_i|\mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i, \beta^{-1}) \right) \tag{11}$$

Введем скрытые переменные $\mathbf{Z} = ||z_{ik}||$, где $z_{ik} = 1 \Leftrightarrow k_i = k$:

$$p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}, \beta, \boldsymbol{\mu}) = \text{Dir}(\boldsymbol{\pi}|\boldsymbol{\mu}) \prod_{k=1}^{K} N(\mathbf{w}_k|\mathbf{0}, \mathbf{A}_k) \prod_{i=1}^{N} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}(y_i|\mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i, \beta^{-1}) \right)^{z_{ik}} \tag{12}$$

Вариационный EM–алгоритм $q(\mathbf{Z}, \mathbf{W}, \boldsymbol{\pi}) = q(\mathbf{Z})\, q(\mathbf{W})\, q(\boldsymbol{\pi})$:

1. E–шаг:
$$\begin{aligned} \log q(\mathbf{Z}^s) &\propto \mathsf{E}_{q/\mathbf{Z}} \log p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}^{s-1}, \beta^{s-1}, \boldsymbol{\mu}) \\ \log q(\mathbf{W}^s) &\propto \mathsf{E}_{q/\mathbf{W}} \log p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}^{s-1}, \beta^{s-1}, \boldsymbol{\mu}) \\ \log q(\boldsymbol{\pi}^s) &\propto \mathsf{E}_{q/\boldsymbol{\pi}} \log p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}^{s-1}, \beta^{s-1}, \boldsymbol{\mu}) \end{aligned} \tag{13}$$

2. M–шаг:
$$\mathbf{A}^s, \beta^s = \arg\max_{\mathbf{A}, \beta} \mathbf{E}_{q^s} \log p(\mathbf{y}, \mathbf{W}, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}, \beta, \boldsymbol{\mu}) \tag{14}$$

Итерационные формулы EM–алгоритма[1]:

1. E–шаг:

$$p(z_{ik}=1)=\frac{\exp\left(\mathsf{E}\log\pi_k-\frac{\beta}{2}\left[y_i^2-2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k+\mathbf{x}_i^{\mathsf{T}}\left(\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\right)\mathbf{x}_i\right]\right)}{\sum_k p(z_{ik}=1)},$$

$$q(\boldsymbol{\pi})=\mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\mu}+\boldsymbol{\gamma}),\quad q(\mathbf{w}_k)=\mathcal{N}(\mathbf{w}_k|\mathbf{m}_k,\mathbf{B}_k),$$

$$\gamma_k=\sum_{i=1}^{N}\mathsf{E}z_{ik},\quad \mathbf{m}_k=\beta\mathbf{B}_k\left(\sum_{i=1}^{N}\mathbf{x}_i y_i\mathsf{E}z_{ik}\right),\quad \mathbf{B}_k=\left(\mathbf{A}_k^{-1}+\beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}z_{ik}\right)^{-1}. \tag{15}$$

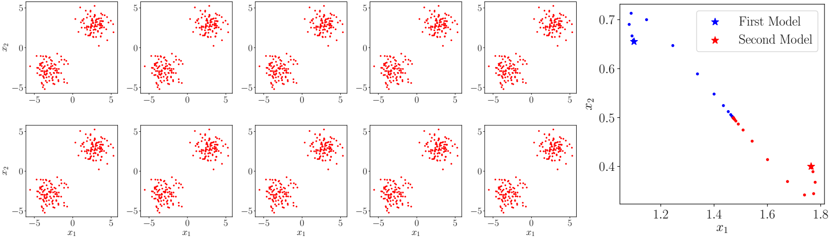
2. M–шаг:

$$\mathbf{A}_k=\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}$$
$$\frac{1}{\beta}=\frac{\sum\sum\left[y_i^2-2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k+\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\mathsf{E}z_{ik}}{\sum\sum\mathsf{E}z_{ik}} \tag{16}$$

Некоторые математические ожидания:

1. $\mathsf{E}z_{ik}=p(z_{ik}=1)$,
2. $\mathsf{E}\log\pi_k=\psi^0(\mu_k+\gamma_k)-\psi^0(K\mu_k+N)$,
3. $\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}=\mathbf{B}_k+\mathbf{m}_k\mathbf{m}_k^{\mathsf{T}}$.

[1] https://github.com/andriygav/EMprior/blob/master/paper/Grabovoy2019Draft.pdf



(a) (b)

На рис показано, что в случае смеси моделей, предсказать то, какую модель использовать в каждой точке нельзя.
На рис. показана зависимость векторов $\mathbf{w}_k$ — параметры локальных моделей в процессе обучения.

[1] https://github.com/andriygav/MixtureLib

Definition

Смесь экспертов — мультимодель, определяющая правдоподобие веса π_k каждой локальной модели $\mathbf{f}_k$ на признаковом описании объекта $\mathbf{x}$.

$$\hat{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k, \qquad \pi_k\left(\mathbf{x}, \mathbf{V}\right) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0, 1], \qquad \sum_{k=1}^{K} \pi_k\left(\mathbf{x}, \mathbf{V}\right) = 1 \tag{17}$$

где $\hat{\mathbf{f}}$ — мультимодель, а $\mathbf{f}_k$ является некоторой моделью,
π_k — параметрическая модель, $\mathbf{w}_k$ — параметры k-й локальной модели,
$\mathbf{V}$ — параметры шлюзовой функции.

Пример 2:

1. Правдоподобие k-й локальной модели $p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right) = \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)$,
2. Априорное распределение параметров k-й локальной модели $p^k\left(\mathbf{w}_k\right) = \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right)$,
3. Шлюзовая функция $\boldsymbol{\pi}\left(\mathbf{x}, \mathbf{V}\right) = \text{softmax}\big(\mathbf{V}_1^{\mathsf{T}}\boldsymbol{\sigma}\left(\mathbf{V}_2^{\mathsf{T}}\mathbf{x}\right)\big)$.

Правдоподобие модели:

$$p\big(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \beta\big) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)\right). \tag{18}$$

Введем скрытые переменные $\mathbf{Z} = ||z_{ik}||$, где $z_{ik} = 1 \Leftrightarrow k_i = k$:

$$p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \beta\big) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right) \prod_{i=1}^{N} \prod_{k=1}^{K} \left(\pi_k \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)\right)^{z_{ik}}. \tag{19}$$

Вариационный EM–алгоритм $q\left(\mathbf{Z}, \mathbf{W}\right) = q\left(\mathbf{Z}\right) q\left(\mathbf{W}\right)$:

1. E–шаг:

$$\begin{aligned} \log q\left(\mathbf{Z}^s\right) &\propto \mathsf{E}_{q/\mathbf{Z}} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}^{s-1}, \mathbf{A}^{s-1}, \mathbf{W}^{0,s-1}, \beta^{s-1}\big) \\ \log q\left(\mathbf{W}^s\right) &\propto \mathsf{E}_{q/\mathbf{W}} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}^{s-1}, \mathbf{A}^{s-1}, \mathbf{W}^{0,s-1}, \beta^{s-1}\big) \end{aligned} \tag{20}$$

2. M–шаг:

$$\mathbf{W}^{0,s}, \mathbf{A}^s, \beta^s = \arg\max_{\mathbf{W}^0, \mathbf{A}, \beta} \mathbf{E}_{q^s} \log p\big(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \beta\big) \tag{21}$$

Итерационные формулы EM–алгоритма[1]:

1. E–шаг:

$$p\left(z_{ik}=1\right)=\frac{\exp\left(\log\pi_k\left(\mathbf{x}_i,\mathbf{V}\right)-\frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i-\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\right)\right)}{\sum_{k'=1}^{K}\exp\left(\log\pi_{k'}\left(\mathbf{x}_i,\mathbf{V}\right)-\frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\mathbf{w}_{k'}^{\mathsf{T}}\mathbf{x}_i-\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\right)\right)},$$

$$q(\mathbf{w}_k)=\mathcal{N}(\mathbf{w}_k|\mathbf{m}_k,\mathbf{B}_k),$$

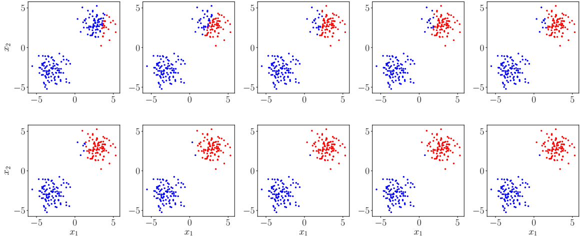
$$\mathbf{m}_k=\mathbf{B}_k\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0+\beta\sum_{i=1}^{N}\mathbf{x}_iy_i\mathsf{E}z_{ik}\right),\quad \mathbf{B}_k=\left(\mathbf{A}_k^{-1}+\beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}z_{ik}\right)^{-1}. \tag{22}$$

2. M–шаг:

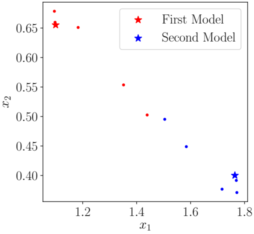
$$\begin{aligned}
\mathbf{A}_k&=\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}-\mathbf{w}_k^0\mathsf{E}\mathbf{w}_k^{\mathsf{T}}-\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{0\mathsf{T}}+\mathbf{w}_k^0\mathbf{w}_k^{0\mathsf{T}},\\
\frac{1}{\beta}&=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_i^2-2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k+\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\mathsf{E}z_{ik},\\
\mathbf{w}_k^0&=\mathsf{E}\mathbf{w}_k,\\
\mathbf{V}&=\arg\max_{\mathbf{V}}\mathbf{E}_{q^s}\log p\big(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\beta\big).
\end{aligned} \tag{23}$$

[1] https://github.com/andriygav/EMprior/blob/master/paper/Grabovoy2019MixtureOfExpert.pdf

Пример 2



(c)

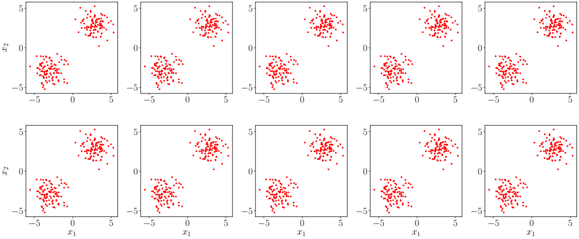


(d)

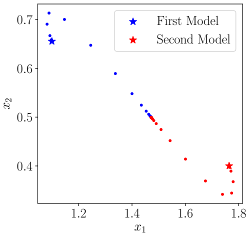
На рис. показано, что в случае смеси экспертов гипермодель предсказывает к какому классу относится каждая точка в пространстве объектов.
На рис. показана зависимость векторов $\mathbf{w}_k$ — параметры локальных моделей в процессе обучения.

[1] `https://github.com/andriygav/MixtureLib`

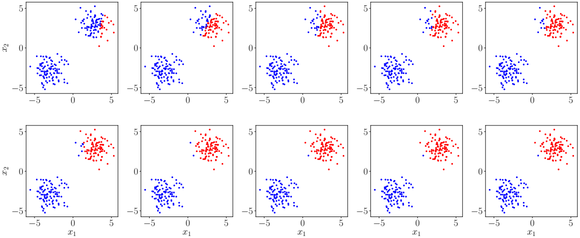
Сравнение результатов



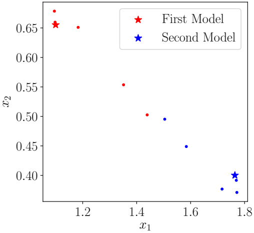
(e) Смесь Моделей



(f) Смесь Моделей



(g) Смесь Экспертов



(h) Смесь Экспертов

[1] https://github.com/andriygav/MixtureLib

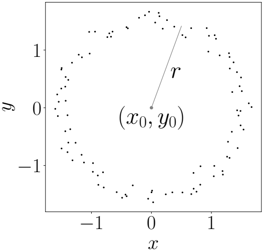
Задано бинарное изображение:

$$\mathbf{M} \in \{0, 1\}^{m_1 \times m_2},$$

где 1 — черная точка, 0 — белая точка фона.
По изображению $\mathbf{M}$ строится выборка $\mathbf{C}$:

$$\mathbf{C} \in \mathbb{R}^{N \times 2},$$

где N — число черных точек на изображении $\mathbf{M}$.



Пусть x_0, y_0 — центр окружности, которую требуется найти, а r ее радиус.
Точки $(x_i, y_i) \in \mathbf{C}$ должны удовлетворять уравнению окружности:

$$(x_i - x_0)^2 + (y_i - y_0)^2 = r^2 \Rightarrow (2x_0) \cdot x_i + (2y_0) \cdot y_i + (r^2 - x_0^2 - y_0^2) \cdot 1 = x_i^2 + y_i^2.$$

Задачу линейной регрессии для нахождения окружности:

$$\mathbf{Xw} \approx \mathbf{y}, \quad \mathbf{X} = [\mathbf{C}, \mathbf{1}], \quad \mathbf{y} = [x_1^2 + y_1^2, x_2^2 + y_2^2, \cdots, x_N^2 + y_N^2]^\mathsf{T},$$

где найденые оптимальные параметры линейной регрессии $\mathbf{w} = \left[w_1, w_2, w_3\right]^\mathsf{T}$ восстанавливают параметры окружности:

$$x_0 = \frac{w_1}{2}, \quad y_0 = \frac{w_2}{2}, \quad r = \sqrt{w_3 + x_0^2 + y_0^2}.$$

Задана выборка:

$$\mathbf{X} \in \mathbb{R}^{N \times n},$$

где N — число объектов в выборке, а n — размерность признакового пространства.

Definition

Смесь экспертов — мультимодель, определяющая правдоподобие веса π_k каждой локальной модели $\mathbf{f}_k$ на признаковом описании объекта $\mathbf{x}$.

$$\hat{\mathbf{f}} = \sum_{k=1}^{K} \pi_k \mathbf{f}_k, \qquad \pi_k\left(\mathbf{x}, \mathbf{V}\right) : \mathbb{R}^{n \times |\mathbf{V}|} \to [0, 1], \qquad \sum_{k=1}^{K} \pi_k\left(\mathbf{x}, \mathbf{V}\right) = 1,$$

где $\hat{\mathbf{f}}$ — мультимодель, а $\mathbf{f}_k$ является локальной моделью, π_k — шлюзовая функция, $\mathbf{w}_k$ — параметры k-й локальной модели, $\mathbf{V}$ — параметры шлюзовой функции.

В качестве локальных моделей $\mathbf{f}_k$ и шлюзовой функции $\boldsymbol{\pi}$ рассматриваются следующие функции:

$$\mathbf{f}_k\left(\mathbf{x}\right) = \mathbf{w}_k^{\mathsf{T}} \mathbf{x}, \quad \boldsymbol{\pi}\left(\mathbf{x}, \mathbf{V}\right) = \text{softmax}\Big(\mathbf{V}_1^{\mathsf{T}} \boldsymbol{\sigma}\big(\mathbf{V}_2^{\mathsf{T}} \mathbf{x}\big)\Big),$$

где $\mathbf{V} = \big\{\mathbf{V}_1, \mathbf{V}_2\big\}$ — параметры шлюзовой функции.

Параметры локальных моделей оптимизируются согласно принципу максимального правдоподобия модели:

$$p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr) = \prod_{k=1}^{K} p^k\left(\mathbf{w}_k\right) \prod_{i=1}^{N} \left(\sum_{k=1}^{K} \pi_k p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right)\right),$$

где $\mathbf{W} = \bigl[\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_K\bigr]^{\mathsf{T}}$.
Задача оптимизации параметров локальных моделей и параметров смеси:

$$\hat{\mathbf{W}}, \hat{\mathbf{V}} = \arg\max_{\mathbf{W}, \mathbf{V}} p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}\bigr).$$

Рассматривается вероятностная постановка задачи:

1) правдоподобие выборки $p_k\left(y_i|\mathbf{w}_k, \mathbf{x}_i\right) = \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)$, где β уровень шума,
2) априорное распределение параметров $p^k\left(\mathbf{w}_k\right) = \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right)$, где $\mathbf{w}_k^0$ — вектор размера $n \times 1$, $\mathbf{A}_k$ — ковариационная матрица параметров,
3) регуляризация априорного распределения $p\left(\boldsymbol{\varepsilon}_{k,k'}|\boldsymbol{\alpha}\right) = \mathcal{N}\left(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\right)$, где $\boldsymbol{\Xi}$ — ковариационная матрица общего вида, $\boldsymbol{\varepsilon}_{k,k'} = \mathbf{w}_k^0 - \mathbf{w}_{k'}^0$.

Правдоподобие модели включает правдоподобие выборки, априорное распределение параметров, а также их регуляризацию

$$p\bigl(\mathbf{y}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr) = \prod_{k,k'=1}^{K} \mathcal{N}\left(\boldsymbol{\varepsilon}_{k,k'}|\mathbf{0}, \boldsymbol{\Xi}\right) \cdot \\ \cdot \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{w}_k|\mathbf{w}_k^0, \mathbf{A}_k\right) \prod_{i=1}^{N}\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(y_i|\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i, \beta^{-1}\right)\right),$$

где $\mathbf{A} = \bigl\{\mathbf{A}_1, \cdots, \mathbf{A}_K\bigr\}$. Введем скрытые переменные $\mathbf{Z} = [z_{ik}]$, где $z_{ik} = 1$ тогда и только тогда, когда $k_i = k$:

$$\log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W}|\mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr) = \\ = \sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left[\log \pi_k\left(\mathbf{x}_i, \mathbf{V}\right) - \frac{\beta}{2}\left(y_i - \mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right)^2 + \frac{1}{2}\log\frac{\beta}{2\pi}\right] + \\ + \sum_{k=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k - \mathbf{w}_k^0\right)^{\mathsf{T}}\mathbf{A}_k^{-1}\left(\mathbf{w}_k - \mathbf{w}_k^0\right) + \frac{1}{2}\log\det\mathbf{A}_k^{-1} - \frac{n}{2}\log 2\pi\right] + \\ + \sum_{k=1}^{K}\sum_{k'=1}^{K}\left[-\frac{1}{2}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right)^{\mathsf{T}}\hat{\boldsymbol{\alpha}}^{-1}\left(\mathbf{w}_k^0 - \mathbf{w}_{k'}^0\right) + \frac{1}{2}\log\det\boldsymbol{\Xi} - \frac{n}{2}\log 2\pi\right].$$

Задача оптимизации параметров локальных моделей и параметров смеси принимает следующий вид:

$$\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta = \arg\max_{\mathbf{W}, \mathbf{Z}, \mathbf{V}, \mathbf{W}^0, \mathbf{A}, \beta} \log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr).$$

Для оптимизации используется вариационный EM–алгоритм с аппроксимацией среднего поля $q(\mathbf{Z}, \mathbf{W}) = q(\mathbf{Z})\, q(\mathbf{W})$.
E и M шаги алгоритма имеют следующий вид:

1. E–шаг:

$$\log q(\mathbf{Z}^s) \propto \mathsf{E}_{q/\mathbf{Z}} \log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}^{s-1}, \mathbf{A}^{s-1}, \mathbf{W}^{0,s-1}, \boldsymbol{\Xi}, \beta^{s-1}\bigr),$$
$$\log q(\mathbf{W}^s) \propto \mathsf{E}_{q/\mathbf{W}} \log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}^{s-1}, \mathbf{A}^{s-1}, \mathbf{W}^{0,s-1}, \boldsymbol{\Xi}, \beta^{s-1}\bigr),$$

2. M–шаг:

$$\mathbf{W}^{0,s}, \mathbf{A}^s, \mathbf{V}^s, \beta^s = \arg\max_{\mathbf{W}^0, \mathbf{A}, \mathbf{V}, \beta} \mathsf{E}_{q^s} \log p\bigl(\mathbf{y}, \mathbf{Z}, \mathbf{W} | \mathbf{X}, \mathbf{V}, \mathbf{A}, \mathbf{W}^0, \boldsymbol{\Xi}, \beta\bigr),$$

где s — номер итерации.

Итерационные формулы EM–алгоритма:

1. E–шаг:

$$p\left(z_{ik}=1\right)=\frac{\exp\left(\log\pi_k\left(\mathbf{x}_i,\mathbf{V}\right)-\frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i-\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\right)\right)}{\sum_{k'=1}^{K}\exp\left(\log\pi_{k'}\left(\mathbf{x}_i,\mathbf{V}\right)-\frac{\beta}{2}\left(\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\mathbf{w}_{k'}^{\mathsf{T}}\mathbf{x}_i-\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_{k'}\right)\right)},$$

$$q(\mathbf{w}_k)=\mathcal{N}(\mathbf{w}_k|\mathbf{m}_k,\mathbf{B}_k),$$

$$\mathbf{m}_k=\mathbf{B}_k\left(\mathbf{A}_k^{-1}\mathbf{w}_k^0+\beta\sum_{i=1}^{N}\mathbf{x}_iy_i\mathsf{E}z_{ik}\right),\quad \mathbf{B}_k=\left(\mathbf{A}_k^{-1}+\beta\sum_{i=1}^{N}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}z_{ik}\right)^{-1}.$$

2. M–шаг:

$$\mathbf{A}_k=\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}-\mathbf{w}_k^0\mathsf{E}\mathbf{w}_k^{\mathsf{T}}-\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{0\mathsf{T}}+\mathbf{w}_k^0\mathbf{w}_k^{0\mathsf{T}},$$

$$\frac{1}{\beta}=\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K}\left[y_i^2-2y_i\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k+\mathbf{x}_i^{\mathsf{T}}\mathsf{E}\mathbf{w}_k\mathbf{w}_k^{\mathsf{T}}\mathbf{x}_i\right]\mathsf{E}z_{ik},$$

$$\mathbf{w}_k^0=\left[\mathbf{A}_k^{-1}+(K-1)\,\boldsymbol{\Xi}\right]^{-1}\left(\mathbf{A}_k^{-1}\mathsf{E}\mathbf{w}_k+\boldsymbol{\Xi}\sum_{k'=1,\ k'\neq k}^{K}\mathbf{w}_{k'}^0\right),$$

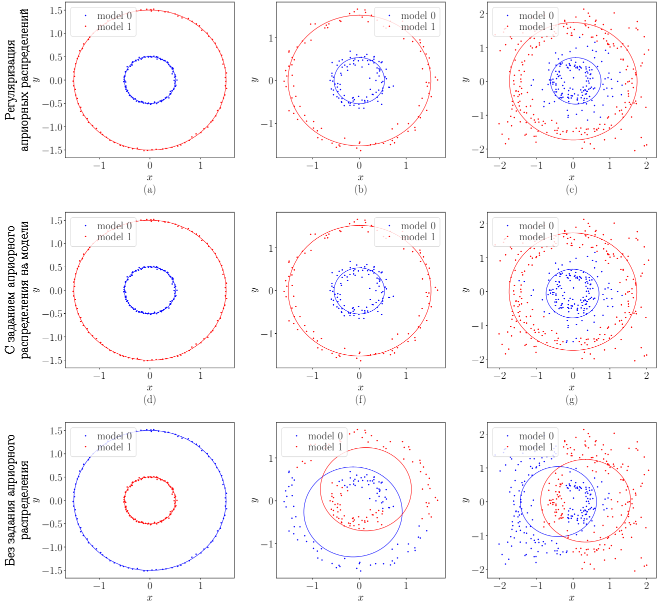
$$\mathbf{V}=\arg\max_{\mathbf{V}}\mathsf{E}_{q^s}\log p\big(\mathbf{y},\mathbf{Z},\mathbf{W}|\mathbf{X},\mathbf{V},\mathbf{A},\mathbf{W}^0,\boldsymbol{\Xi},\beta\big).$$

Вычислительный эксперимент делится на следующие этапы:

1. Анализ синтетических данных с разным типом шума в изображении;
2. Анализ изменения параметров локальных моделей во время обучения;
3. Анализ мультимоделей в зависимости от уровня шума в изображении;
4. Анализ качества модели на реальных данные.

Гиперпараметры заданы следующим образом:

1. Априорные распределения на параметры локальных моделей в эксперименте задано следующим образом:
$$p^1\left(\mathbf{w}_1\right) \sim \mathcal{N}\left(\mathbf{w}_1^0, \mathbf{I}\right), \quad p^2\left(\mathbf{w}_2\right) \sim \mathcal{N}\left(\mathbf{w}_2^0, \mathbf{I}\right),$$
где $\mathbf{w}_1^0 = [0, 0, 0.1]$, $\mathbf{w}_2^0 = [0, 0, 2]$.
2. Параметр регуляризации:
$$\boldsymbol{\Xi} = \begin{bmatrix} 0.01 & 0 & 0 \\ 0 & 0.01 & 0 \\ 0 & 0 & 1 \end{bmatrix},$$
что указывает на концентричность окружностей.

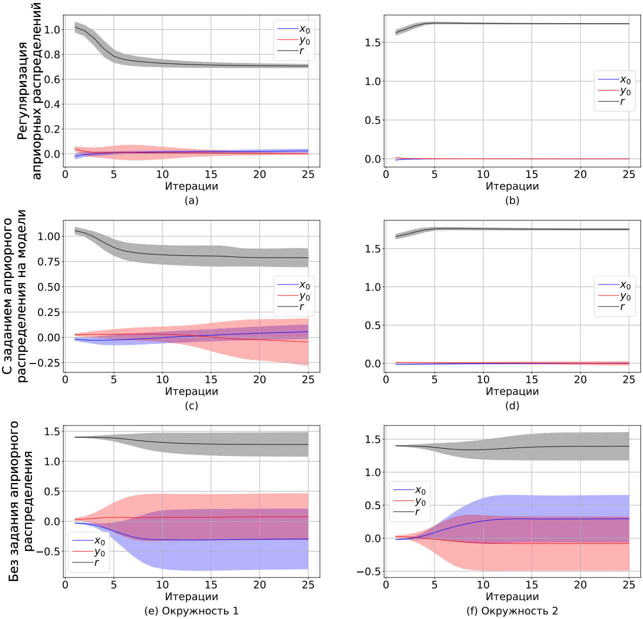


(a) (b) (c)

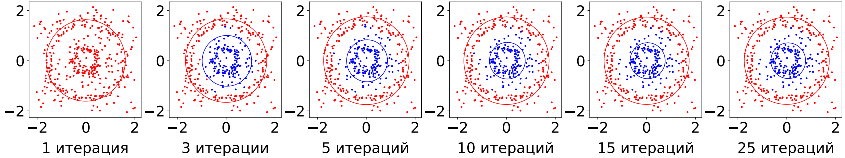
(d) (f) (g)

(e) Без шума (h) Шум вблизи окружности (j) Равномерный шум

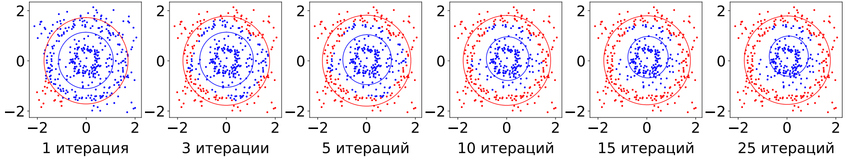
Параметры локальных моделей в процессе обучения



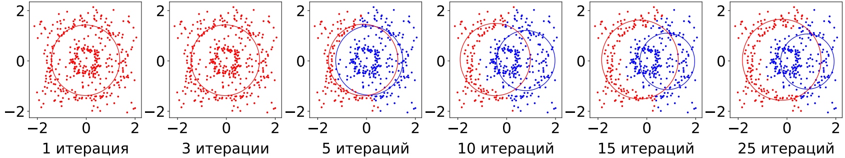
Регуляризация априорных распределений



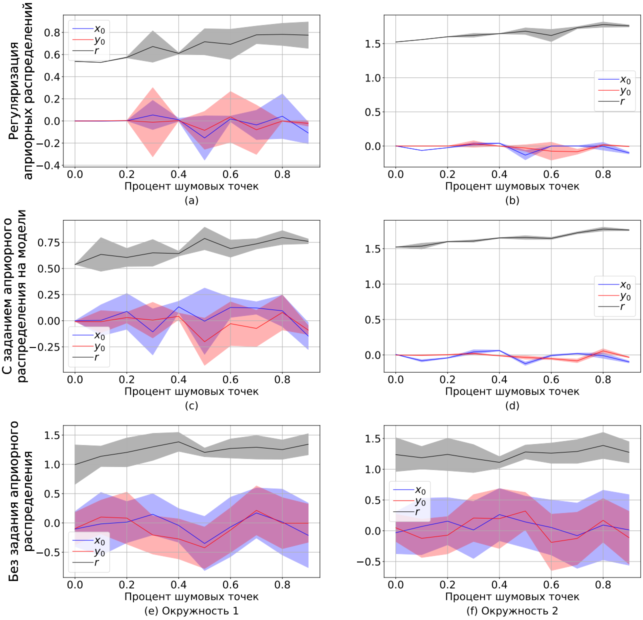
С заданием априорного распределения на модели



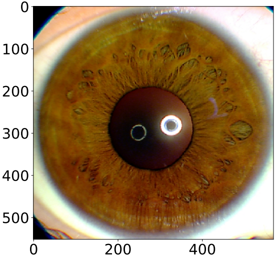
Без задания априорного распределения



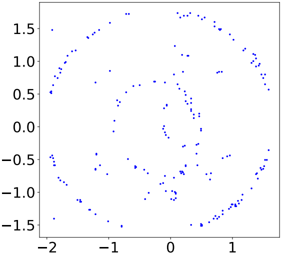
Анализ мультимоделей в зависимости от уровня шума



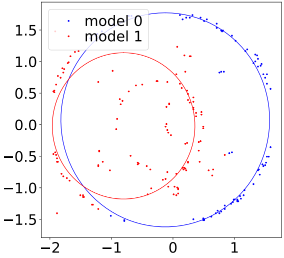
Результаты на реальных данных



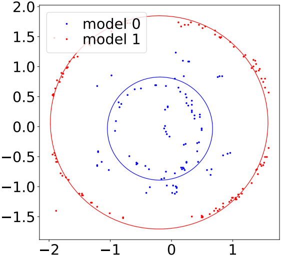
(a) Исходное изображение



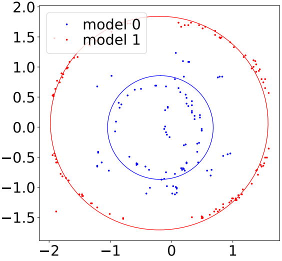
(b) Бинаризованное изображение



(c) Без задания априорного распределения



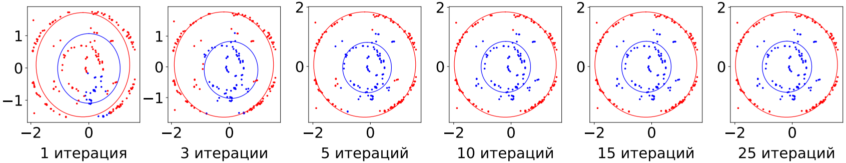
(d) С заданием априорного распределения на модели



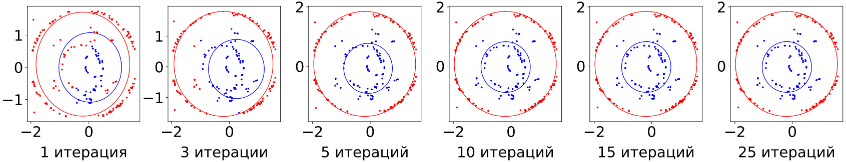
(e) Регуляризация априорных распределений

Обучения на реальных данных

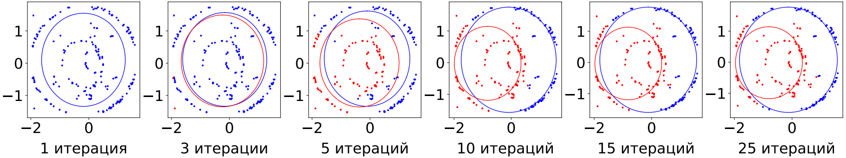
Регуляризация априорных распределений



С заданием априорного распределения на модели



Без задания априорного распределения



Пусть задано множество объектов $\mathbf{\Omega}$, а также подмножество наблюдаемых объектов $\mathbf{\Omega}'$

$$\mathbf{\Omega}' \subset \mathbf{\Omega},$$

где $|\mathbf{\Omega}'| = N$. Пусть для $\mathbf{\Omega}$ задана некоторая экспертная информация $\boldsymbol{E}(\mathbf{\Omega})$. На основе экспертной информации $\boldsymbol{E}(\mathbf{\Omega})$ введем отображения из множества объектов $\mathbf{\Omega}$:

$$K_y^{\boldsymbol{E}(\mathbf{\Omega})} : \mathbf{\Omega} \to \mathbb{R}, \quad K_x^{\boldsymbol{E}(\mathbf{\Omega})} : \mathbf{\Omega} \to \mathbb{R}^n,$$

где n количество признаков, причем предполагаем, что $n \ll N$. Применив отображения $K_x^{\boldsymbol{E}(\mathbf{\Omega})}$ и $K_y^{\boldsymbol{E}(\mathbf{\Omega})}$ для множества наблюдаемых объектов $\mathbf{\Omega}'$ получаем выборку:

$$\mathfrak{D}\left(\mathbf{\Omega}', \boldsymbol{E}(\mathbf{\Omega})\right) = \{(\mathbf{x}, y) \,|\mathbf{x} = K_x^{\boldsymbol{E}(\mathbf{\Omega})}(\omega),\ y = K_y^{\boldsymbol{E}(\mathbf{\Omega})}(\omega),\ \forall \omega \in \mathbf{\Omega}'\}.$$

Предполагается, что существуют нетривиальные отображения $K_y^{\boldsymbol{E}(\mathbf{\Omega})}, K_x^{\boldsymbol{E}(\mathbf{\Omega})}$, и $\mathbf{w} \in \mathbb{R}^n$, такие, что:

$$y \approx \mathbf{x}^{\mathsf{T}}\mathbf{w}, \quad \forall (\mathbf{x}, y) \in \mathfrak{D}\left(\mathbf{\Omega}', \boldsymbol{E}(\mathbf{\Omega})\right),$$

то есть получаем выборку, которая является задачей линейной регрессии по нахождению неизвестного вектора $\mathbf{w}$ (аналогично можно ввести задачу для логистической регрессии).

В случае, когда экспертная информация представляется в виде объединения нескольких экспертов:

$$\boldsymbol{E}\left(\boldsymbol{\Omega}\right)=\boldsymbol{E}_0\left(\boldsymbol{\Omega}\right)\cup\boldsymbol{E}_1\left(\boldsymbol{\Omega}_1\right)\cup\boldsymbol{E}_2\left(\boldsymbol{\Omega}_2\right)\cup\cdots\cup\boldsymbol{E}_K\left(\boldsymbol{\Omega}_K\right),\quad \cup_{i=k}^{K}\boldsymbol{\Omega}_k=\boldsymbol{\Omega}$$

в этом случае будем говорить о задаче смеси K экспертов. Каждая информация $\boldsymbol{E}_k\left(\boldsymbol{\Omega}_k\right)$ описывает локальную информацию о каком-то подмножестве объектов $\boldsymbol{\Omega}_k$ для всех $k=\overline{1...K}$. Информация эксперта $\boldsymbol{E}_0\left(\boldsymbol{\Omega}\right)$ описывает глобальную информацию о всем множестве объектов $\boldsymbol{\Omega}$.
В случае задачи смеси K экспертов вводятся отображения:

$$\begin{aligned}K_y^{\boldsymbol{E}_1\left(\boldsymbol{\Omega}_1\right)}:\boldsymbol{\Omega}\to\mathbb{R},\quad K_x^{\boldsymbol{E}_1\left(\boldsymbol{\Omega}_1\right)}:\boldsymbol{\Omega}\to\mathbb{R}^{n_1},\\ \cdots\\ K_y^{\boldsymbol{E}_K\left(\boldsymbol{\Omega}_K\right)}:\boldsymbol{\Omega}\to\mathbb{R},\quad K_x^{\boldsymbol{E}_K\left(\boldsymbol{\Omega}_K\right)}:\boldsymbol{\Omega}\to\mathbb{R}^{n_K},\end{aligned}$$

где получаем множество отображений во множество локальных моделей, в которых учтены информации от каждого эксперта.
Также, как в и задаче одного эксперта вводится предположения, что каждая локальная модель является линейной:

$$\forall k\in\{1,\cdots K\},\quad y\approx\mathbf{x}^{\mathsf{T}}\mathbf{w}_k,\quad\forall\left(\mathbf{x},y\right)\in\mathfrak{D}\left(\boldsymbol{\Omega}_k',\boldsymbol{E}\left(\boldsymbol{\Omega}_k\right)\right).$$

Заметим, что истинного разбиения $\boldsymbol{\Omega}$ на множества $\{\boldsymbol{\Omega}_k\}_{k=1}^{K}$ нету. Рассмотрим вектор функцию $\boldsymbol{\pi}$:

$$\boldsymbol{\pi} : \boldsymbol{\Omega} \to \mathbb{R}^K, \quad \sum_{k=1}^{K} \pi_k\left(\omega\right) = 1,\ \forall \omega \in \boldsymbol{\Omega},$$

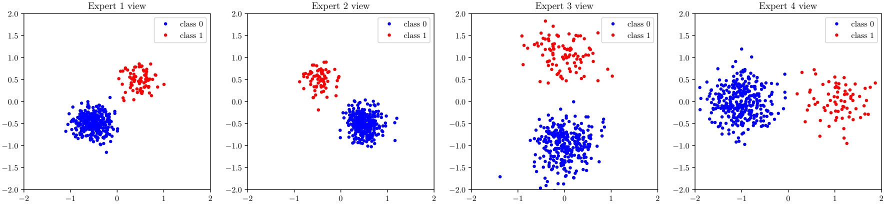
где $\boldsymbol{\pi}$ назовем шлюзовой функцией.

Предположим, что все $\left\{\left(K_x^{\boldsymbol{E}_k(\boldsymbol{\Omega}_k)}, K_y^{\boldsymbol{E}_k(\boldsymbol{\Omega}_k)}\right)\right\}_{k=1}^{K}$ являются заданными отображениями. Используя локальные модели, построим глобальную мультимодель, которая описывает все множество объектов $\boldsymbol{\Omega}$:

$$\sum_{\omega \in \boldsymbol{\Omega}'} \sum_{k=1}^{K} \pi_k\left(\omega, \mathbf{V}\right)\left(K_y^{\boldsymbol{E}_k(\boldsymbol{\Omega}_k)}\left(\omega\right) - \mathbf{w}_k^{\mathsf{T}} K_x^{\boldsymbol{E}_k(\boldsymbol{\Omega}_k)}\left(\omega\right)\right)^2 + R\left(\mathbf{V}, \mathbf{W}, \boldsymbol{E}\left(\boldsymbol{\Omega}\right)\right) \to \min_{\mathbf{V}, \mathbf{W}}$$

где $\mathbf{W} = [\mathbf{w}_1^{\mathsf{T}}, \cdots, \mathbf{w}_K^{\mathsf{T}}]$, а $R\left(\mathbf{V}, \mathbf{W}, \boldsymbol{E}\left(\boldsymbol{\Omega}\right)\right)$ является некоторой регуляризацией параметров, которая также основывается на экспертной информации, $\mathbf{V}$ — параметры шлюзовой функции.

Пример


Expert 1 view
class 0
class 1
Expert 2 view
class 0
class 1
Expert 3 view
class 0
class 1
Expert 4 view
class 0
class 1

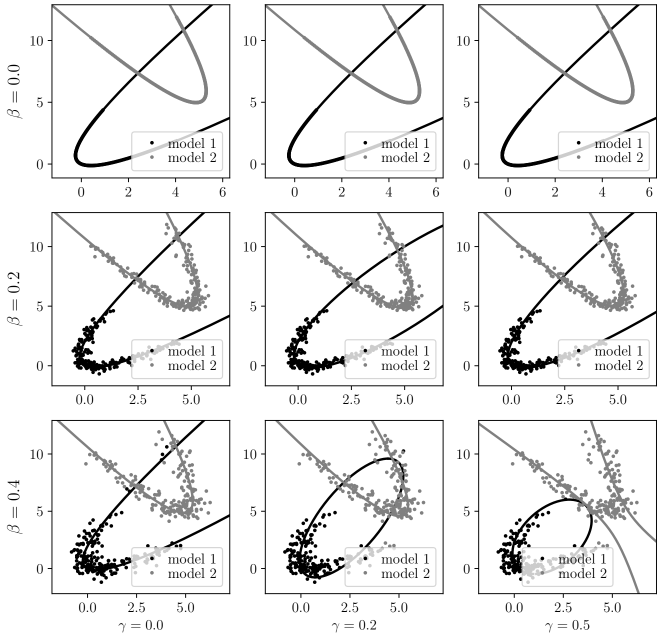
Произвольная кривая второго порядка, главная ось которой не параллельна оси ординат задается следующим выражением:

$$x^2 = B'xy + C'y^2 + D'x + E'y + F',$$

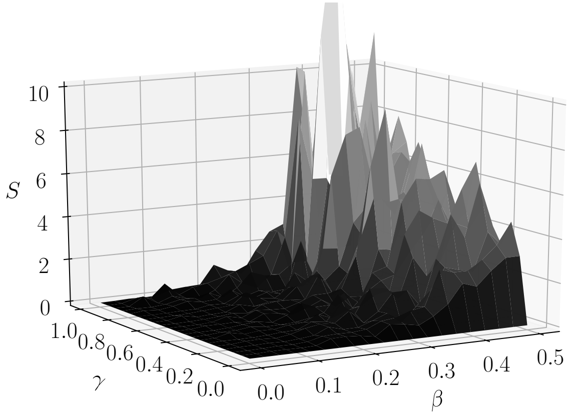
где на коэффициенты B', C' накладываются ограничения, которые зависят от вида кривой. Выражение принимает следующий вид:

$$K_x\big(\mathbf{c}_i\big) = \big[x_i y_i, y_i^2, x_i, y_i, 1\big], \quad K_y\big(\mathbf{c}_i\big) = x_i^2,$$

откуда получаем задачу линейной регрессии для восстановления параметров B', C', D', E', F' по составленной выборке.



$\beta = 0.0$
$\beta = 0.2$
$\beta = 0.4$
$\gamma = 0.0$
$\gamma = 0.2$
$\gamma = 0.5$
model 1
model 2


10
8
6
4
2
0
S
1.0
0.8
0.6
0.4
0.2
0.0
γ
0.0
0.1
0.2
0.3
0.4
0.5
β