Байесовский выбор моделей: обоснованность и отбор признаков в логистической регрессии

Александр Адуенко

24е октября 2018

Содержание предыдущих лекций

- Формула Байеса: $\mathrm{P}(A|B) = \dfrac{\mathrm{P}(B|A)\mathrm{P}(A)}{\mathrm{P}(B)}$;
- Формула полной вероятности: $\mathrm{P}(B) = \mathrm{P}(B|A)\mathrm{P}(A) + \mathrm{P}(B|\overline{A})\mathrm{P}(\overline{A})$;
- Определение априорных вероятностей и selection bias;
- (Множественное) тестирование гипотез
- Экспоненциальное семейства. Достаточные статистики.
- Наивный байесовский классификатор. Связь целевой функции и вероятностной модели.
- Линейная регрессия: связь МНК и $\mathbf{w}_{\mathrm{ML}}$, регуляризации и $\mathbf{w}_{\mathrm{MAP}}$.
- Свойство сопряженности априорного распределения правдоподобию.
- Прогноз для одиночной модели:
  $$p(\mathbf{y}_{\mathrm{test}}|\mathbf{X}_{\mathrm{test}}, \mathbf{X}_{\mathrm{train}}, \mathbf{y}_{\mathrm{train}}) = \int p(\mathbf{y}_{\mathrm{test}}|\mathbf{w}, \mathbf{X}_{\mathrm{test}})p(\mathbf{w}|\mathbf{X}_{\mathrm{train}}, \mathbf{y}_{\mathrm{train}})d\mathbf{w}.$$
- Связь апостериорной вероятности модели и обоснованности $p(M_i|\mathbf{X}_{\mathrm{train}}, \mathbf{y}_{\mathrm{train}}) \propto p(M_i)p_i(\mathbf{y}_{\mathrm{train}}|\mathbf{X}_{\mathrm{train}})$.
- Обоснованность: понимание и связь со статистической значимостью.
- Логистическая регрессия: проблемы ML-оценки $\mathbf{w}$ и связь априорного распределения с отбором признаков.

Апостериорное распределение в логистической регрессии

$p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = p(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A})$, где $p(\mathbf{y}|\mathbf{X},\ \mathbf{w}) = \prod_{j=1}^{m} \sigma(y_j \mathbf{w}^\mathsf{T} \mathbf{x}_j)$.

$$p(\mathbf{w}|\mathbf{X},\ \mathbf{y},\ \mathbf{A}) = \frac{p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A})}{p(\mathbf{y}|\mathbf{X},\ \mathbf{A})} = \frac{\prod_{j=1}^{m} \sigma(y_j \mathbf{w}^\mathsf{T} \mathbf{x}_j) N(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1})}{p(\mathbf{y}|\mathbf{X},\ \mathbf{A})}.$$

$$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}}|\mathbf{w},\ \mathbf{X}_{\text{test}}) p(\mathbf{w}|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) d\mathbf{w}.$$

Вопрос 1: Как определить $\mathbf{w}_{\text{MAP}}$? Единственное ли решение?

$q(\mathbf{w}) = -\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = -\log p(\mathbf{w}|\mathbf{A}) - \log p(\mathbf{y}|\mathbf{X},\ \mathbf{w}) =$
$q(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\mathsf{T} \mathbf{H}^{-1} (\mathbf{w} - \mathbf{w}_{\text{MAP}}) + O(\|\mathbf{w} - \mathbf{w}_{\text{MAP}}\|^3)$, где
$\mathbf{H}^{-1} = \mathbf{A} + \mathbf{X}^\mathsf{T} \mathbf{R} \mathbf{X}$, где $\mathbf{R} = \text{diag}(\sigma(\mathbf{w}_{\text{MAP}}^\mathsf{T} \mathbf{x}_j)\sigma(-\mathbf{w}_{\text{MAP}}^\mathsf{T} \mathbf{x}_j))$.
Нормальная аппроксимация: $p(\mathbf{w}|\mathbf{X},\ \mathbf{y},\ \mathbf{A}) \approx N(\mathbf{w}|\mathbf{w}_{\text{MAP}},\ \mathbf{H}^{-1})$.

Пример. Пусть $n = 1$, $\mathbf{w}_{\text{MAP}} = 1$.
Вопрос 2: Что можно сказать про принадлежность объектов с $x = 0;\ 1;\ -1;\ 5;\ -5$ к классу 1?
Вопрос 3: Как результат зависит от неопределенности h^{-1}? Что происходит при $h \to 0$ и при $h \to \infty$?

Нелинейная разделяющая поверхность

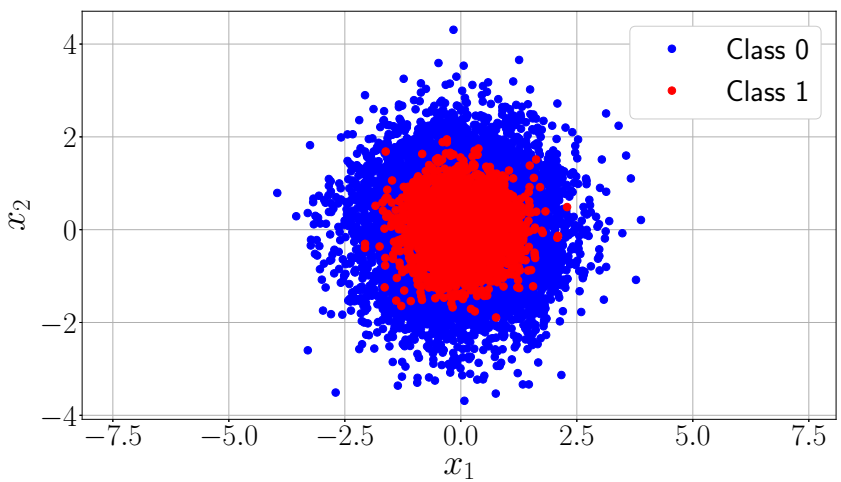
$$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}}|\mathbf{w},\ \mathbf{X}_{\text{test}})p(\mathbf{w}|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})d\mathbf{w}.$$

$$p(\mathbf{w}|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) = N(w|1,\ h^{-1}).$$

Прогноз вероятности класса 1 в зависимости от неопределенности h^{-1}

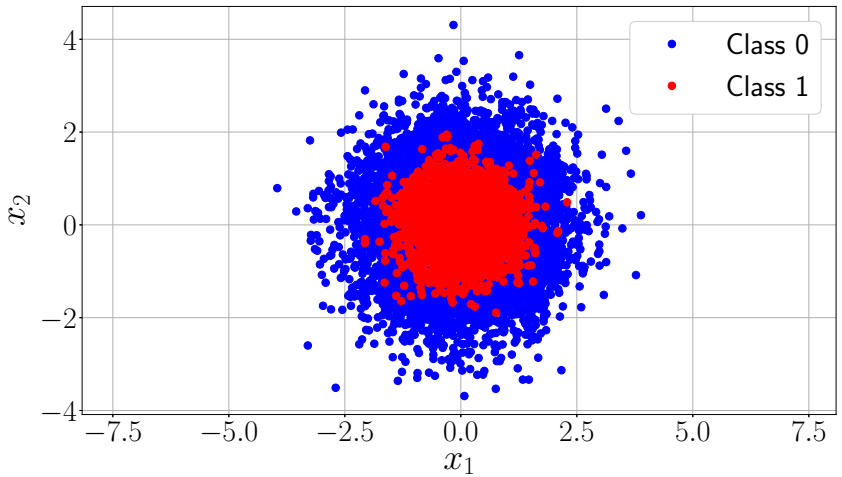
	$x=5$	$x=1$	$x=0$	$x=-1$	$x=-5$
$h=\infty$	0.0067	0.269	0.5	0.731	0.9933
$h=1$	0.169	0.301	0.5	0.699	0.831
$h=0$	0.5	0.5	0.5	0.5	0.5

Вопрос 1: как учесть в модели, что классы не сбалансированы?



Вопрос 2: что делать, если разделяющая поверхность нелинейна?

Выбросы и пропуски в данных



Вопрос 1: что делать, если разделяющая поверхность нелинейна?

Идея:

$$\mathbf{x} \mapsto \varphi(\mathbf{x}) = [K(\mathbf{x}, \mathbf{x}_i), i = 1, \ldots, m].$$

Вопрос 2: Чему соответствует отбор признаков при замене $\mathbf{x} \mapsto \varphi(\mathbf{x}) = [K(\mathbf{x}, \mathbf{x}_i), i = 1, \ldots, m]$?

Вопрос 3: Что если значения части признаков не заданы или некорректны? Что происходит при замене на среднее / медиану?

Исходная модель: $p(\mathbf{y}, \mathbf{w}|\mathbf{X}, \mathbf{A}) = p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A})$.

Пусть $\mathbf{X} = \tilde{\mathbf{X}} + \mathbf{Z}$, $\tilde{\mathbf{X}} \cdot \mathbf{Z} = \mathbf{0}$, где $\mathbf{Z}$ – матрица значений пропусков.

Новая модель: $p(\mathbf{y}, \mathbf{w}, \mathbf{Z}|\tilde{\mathbf{X}}, \mathbf{A}) = p(\mathbf{y}|\tilde{\mathbf{X}}, \mathbf{Z}, \mathbf{w})p(\mathbf{w}|\mathbf{A})p(\mathbf{Z}|\tilde{\mathbf{X}})$.

$$p(\mathbf{w}|\mathbf{y}, \tilde{\mathbf{X}}, \mathbf{A}) \propto p(\mathbf{y}, \mathbf{w}|\tilde{\mathbf{X}}, \mathbf{A}) = \int p(\mathbf{y}, \mathbf{w}, \mathbf{Z}|\tilde{\mathbf{X}}, \mathbf{A})d\mathbf{Z} =$$

$$\int p(\mathbf{y}|\tilde{\mathbf{X}}, \mathbf{Z}, \mathbf{w})p(\mathbf{w}|\mathbf{A}) \underbrace{p(\mathbf{Z}|\tilde{\mathbf{X}})}_{\text{вес}} d\mathbf{Z}.$$

EM-алгоритм

Пусть $\mathbf{D} = (\mathbf{X},\ \mathbf{y})$ – наблюдаемые переменные, $\mathbf{Z}$ – скрытые переменные.
$p(\mathbf{D},\ \mathbf{Z}|\mathbf{\Theta}) = p(\mathbf{D}|\mathbf{X},\ \mathbf{\Theta})p(\mathbf{Z}|\mathbf{\Theta})$.

Вопрос 1: как решить задачу $p(\mathbf{D}|\mathbf{\Theta}) = \int p(\mathbf{D},\ \mathbf{Z}|\mathbf{\Theta})d\mathbf{Z} \to \max\limits_{\mathbf{\Theta}}$?

Пример 1. $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},\ \mathbf{w} \sim N(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}),\ \boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \beta^{-1}\mathbf{I})$
$p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta) = p(\mathbf{y}|\mathbf{X},\ \mathbf{w},\ \beta)p(\mathbf{w}|\mathbf{A})$.
$\log p(\mathbf{y}|\mathbf{X},\ \underbrace{\mathbf{A},\ \beta^{-1}}_{\mathbf{\Theta}}) \propto -\frac{1}{2}\log\det(\beta^{-1}\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathsf{T}}) - \frac{1}{2}\mathbf{y}^{\mathsf{T}}(\beta^{-1}\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathsf{T}})^{-1}\mathbf{y}$.

EM-алгоритм
Введем $F(q,\ \mathbf{\Theta}) = -\int q(\mathbf{Z})\log q(\mathbf{Z})d\mathbf{Z} + \int q(\mathbf{Z})\log p(\mathbf{D},\ \mathbf{Z}|\mathbf{\Theta})d\mathbf{Z} =$

$-\int q(\mathbf{Z})\log q(\mathbf{Z})d\mathbf{Z} + \int q(\mathbf{Z})\log p(\mathbf{Z}|\mathbf{D},\ \mathbf{\Theta})d\mathbf{Z} + \int \log p(\mathbf{D}|\mathbf{\Theta})q(\mathbf{Z})d\mathbf{Z} =$

$\log p(\mathbf{D}|\mathbf{\Theta}) - \int q(\mathbf{Z})\log\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{D},\mathbf{\Theta})}d\mathbf{Z} = \log p(\mathbf{D}|\mathbf{\Theta}) - D_{\mathrm{KL}}(q\|p(\mathbf{Z}|\mathbf{D},\ \mathbf{\Theta}))$.

Идея 1: $p(\mathbf{D}|\mathbf{\Theta}) \to \max\limits_{\mathbf{\Theta}}$ заменим на $F(q,\ \mathbf{\Theta}) \to \max\limits_{q,\ \mathbf{\Theta}}$.

Идея 2: Пошагово оптимизируем по $\mathbf{\Theta}$ и q, то есть

1. E-шаг: $q^s = F(q,\ \mathbf{\Theta}^s) \to \max\limits_{q}$;
2. M-шаг: $\mathbf{\Theta}^s = F(q^{s-1},\ \mathbf{\Theta}) \to \max\limits_{\mathbf{\Theta}}$.

EM-алгоритм для максимизации обоснованности

$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},\ \mathbf{w} \sim N(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}),\ \boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \beta^{-1}\mathbf{I})$

$p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta) = p(\mathbf{y}|\mathbf{X},\ \mathbf{w},\ \beta)p(\mathbf{w}|\mathbf{A}) =.$

$\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta) \propto \frac{m}{2}\log\beta - \frac{\beta}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2}\log\det\mathbf{A} - \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w}.$

$$F(q,\ \mathbf{A},\ \beta) = -\int q(\mathbf{w})\log q(\mathbf{w})d\mathbf{w} + \int q(\mathbf{w})\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta)d\mathbf{w} =$$

$$\log p(\mathbf{y}|\mathbf{X},\ \mathbf{A},\ \beta) - D_{\mathrm{KL}}(q(\mathbf{w})\|p(\mathbf{w}|\mathbf{y},\ \mathbf{X},\ \mathbf{A},\ \beta)) \to \max_{q,\ \mathbf{A},\ \beta}.$$

E-шаг (считаем $\mathbf{A}$, β фиксированными)

$F(q,\ \mathbf{A},\ \beta) \to \max\limits_{q} \Longleftrightarrow q(\mathbf{w}) = p(\mathbf{w}|\mathbf{y},\ \mathbf{X},\ \mathbf{A},\ \beta) = \mathrm{N}(\mathbf{w}|\mathbf{w}_0,\ \boldsymbol{\Sigma}_0^{-1})$, где

$\boldsymbol{\Sigma}_0 = \mathbf{A} + \beta\mathbf{X}^{\mathsf{T}}\mathbf{X},\ \mathbf{w}_0 = \beta\boldsymbol{\Sigma}_0^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}.$

M-шаг (считаем $q(\mathbf{w})$ фиксированным)

$$\mathsf{E}_{q(\mathbf{w})}\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta) = \int q(\mathbf{w})\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta)d\mathbf{w} \to \max_{\mathbf{A},\ \beta}.$$

$$\tilde{F}(\mathbf{A},\ \beta) = \frac{m}{2}\log\beta - \frac{\beta}{2}\mathsf{E}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2}\sum_{j=1}^{n}\log\alpha_j - \frac{1}{2}\sum_{j=1}^{n}\alpha_j\mathsf{E}w_j^2 \to \max_{\mathbf{A},\ \beta}.$$

$$\frac{\partial F}{\partial\alpha_j} = \frac{1}{2\alpha_k} - \frac{1}{2}\mathsf{E}w_j^2 = 0 \Longleftrightarrow \alpha_j = \frac{1}{\mathsf{E}w_j^2}.$$

Hint: $1 = \alpha_j(\mathsf{E}^2 w_j + \mathsf{D}w_j) \Longrightarrow \alpha_j^{\mathsf{new}} = \frac{1-\alpha_j^{\mathsf{old}}\mathsf{D}w_j}{\mathsf{E}^2 w_j}.$

$$\frac{\partial F}{\partial\beta} = \frac{m}{2\beta} - \frac{1}{2}\mathsf{E}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = 0 \Longleftrightarrow \beta = \frac{m}{\mathsf{E}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2}.$$

EM-алгоритм для максимизации обоснованности

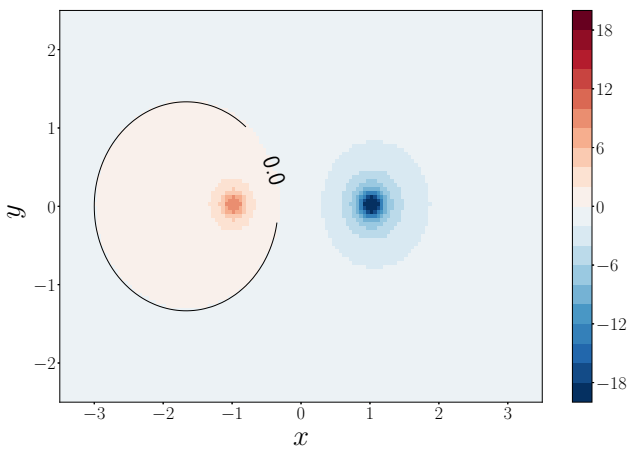
Потенциал поля точечного заряда: $\varphi = k\frac{q}{r}$.

Пусть имеется несколько зарядов $q_1, \ldots, q_l$ в точках $\mathbf{z}_1, \ldots, \mathbf{z}_l$.

Тогда $\varphi(\mathbf{x}) = k\sum\limits_{i=1}^{l}\frac{q_l}{\|\mathbf{x}-\mathbf{z}_l\|}$. По набору точек $\mathbf{x}_1, \ldots, \mathbf{x}_m$ и измеренным

$$y_i = \varphi(\mathbf{x}_i) - \underbrace{\varphi(\infty)}_{=0} + \varepsilon_i,\ \varepsilon_i \sim N(\varepsilon_i|0,\ \beta^{-1})$$

требуется оценить $\varphi(\mathbf{x})$ для $\mathbf{x}$ из тестовой выборки.



$$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\boldsymbol{\varepsilon}|\mathbf{0},\ \beta^{-1}\mathbf{I}),\ \text{где}$$
$$\mathbf{\Phi} = \left\|\frac{1}{\delta+\|\mathbf{x}_i-\mathbf{x}_j\|}\right\|,\ i,\ j = \overline{1,\ m};$$
$$\mathbf{w} \sim p(\mathbf{w}|\mathbf{A}) = N(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}).$$

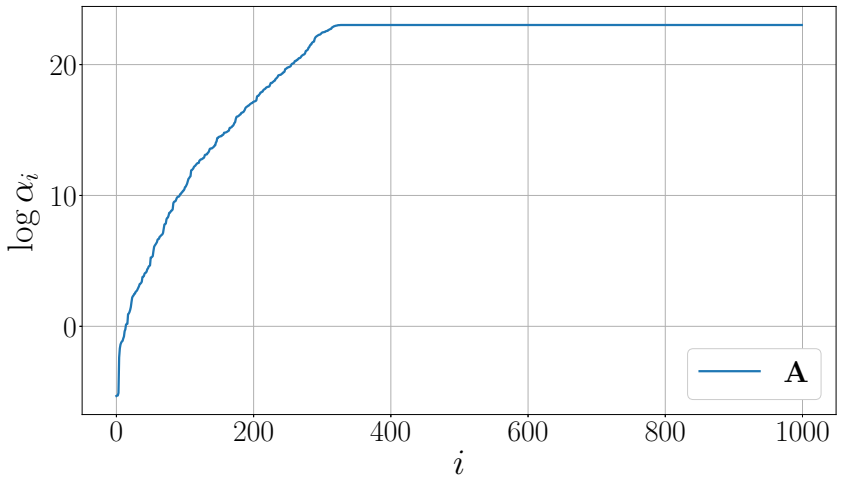
Шаг 1: $p(\mathbf{y}_{\text{train}}|\mathbf{\Phi}_{\text{train}},\ \mathbf{A},\ \beta) \to \max\limits_{\mathbf{A},\,\beta}$ позволит отобрать признаки.

Шаг 2: Прогноз для тестовой выборки:

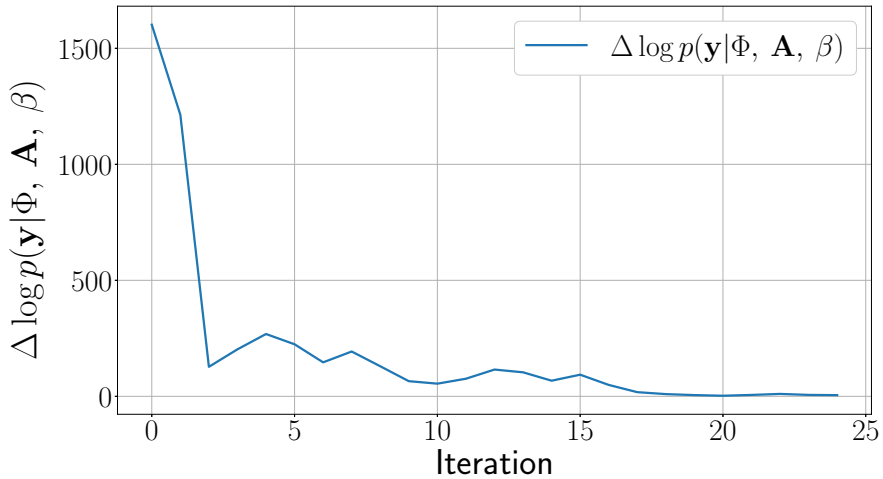
$$p(\mathbf{y}_{\text{test}}|\mathbf{\Phi}_{\text{test}},\ \mathbf{\Phi}_{\text{train}},\ \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}}|\mathbf{w},\ \mathbf{\Phi}_{\text{test}})p(\mathbf{w}|\mathbf{\Phi}_{\text{train}},\ \mathbf{y}_{\text{train}})d\mathbf{w}$$

Результаты для задачи восстановления потенциала

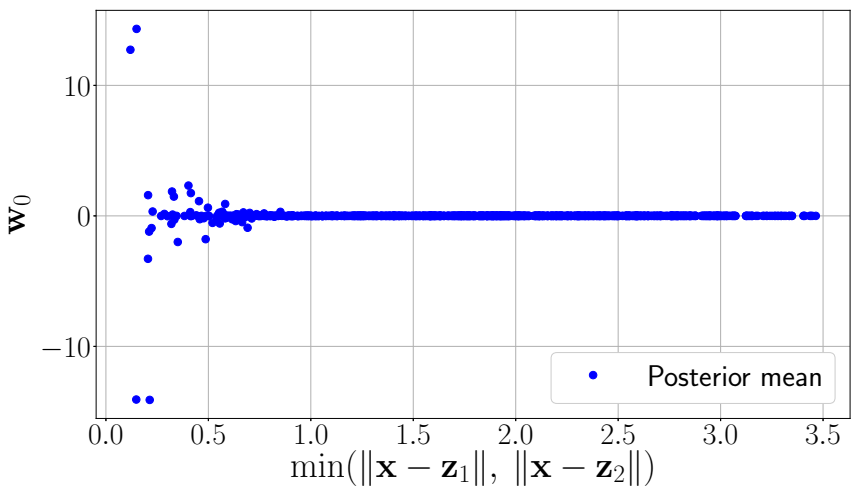
Оптимальный $\boldsymbol{\alpha}$



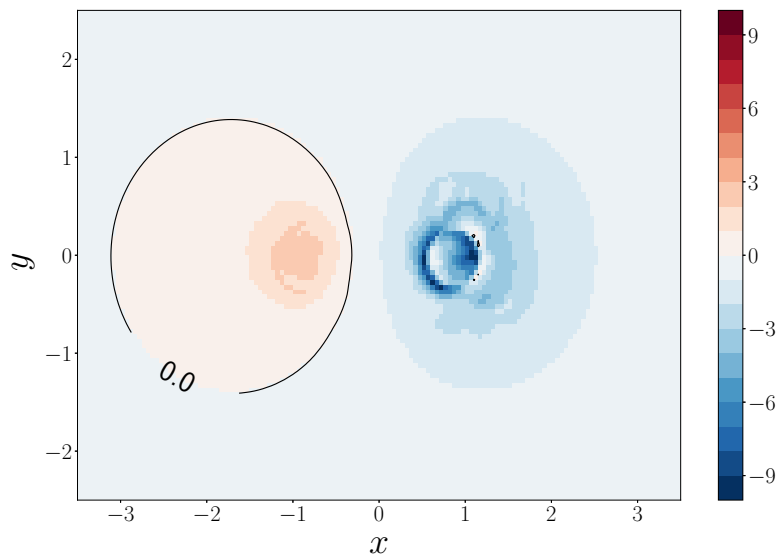
Обоснованность по итерациям



Среднее апостериорного распределения $\mathbf{w}_0$



Восстановленный потенциал



Литература

1. Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 113-120, 161-171.
2. MacKay, David JC. Bayesian methods for adaptive models. Diss. California Institute of Technology, 1992.
3. MacKay, David JC. "The evidence framework applied to classification networks."Neural computation 4.5 (1992): 720-736.
4. Gelman, Andrew, et al. Bayesian data analysis, 3rd edition. Chapman and Hall/CRC, 2013.
5. Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.
6. Chen, Ming-Hui, and Joseph G. Ibrahim. "Conjugate priors for generalized linear models."Statistica Sinica (2003): 461-476.
7. Fahrmeir, Ludwig, and Heinz Kaufmann. "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models."The Annals of Statistics (1985): 342-368.
8. Baghishani, Hossein, and Mohsen Mohammadzadeh. "Asymptotic normality of posterior distributions for generalized linear mixed models."Journal of Multivariate Analysis 111 (2012): 66-77.