Лекция 7. Недиагональная регуляризация обобщенных линейных моделей

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения»

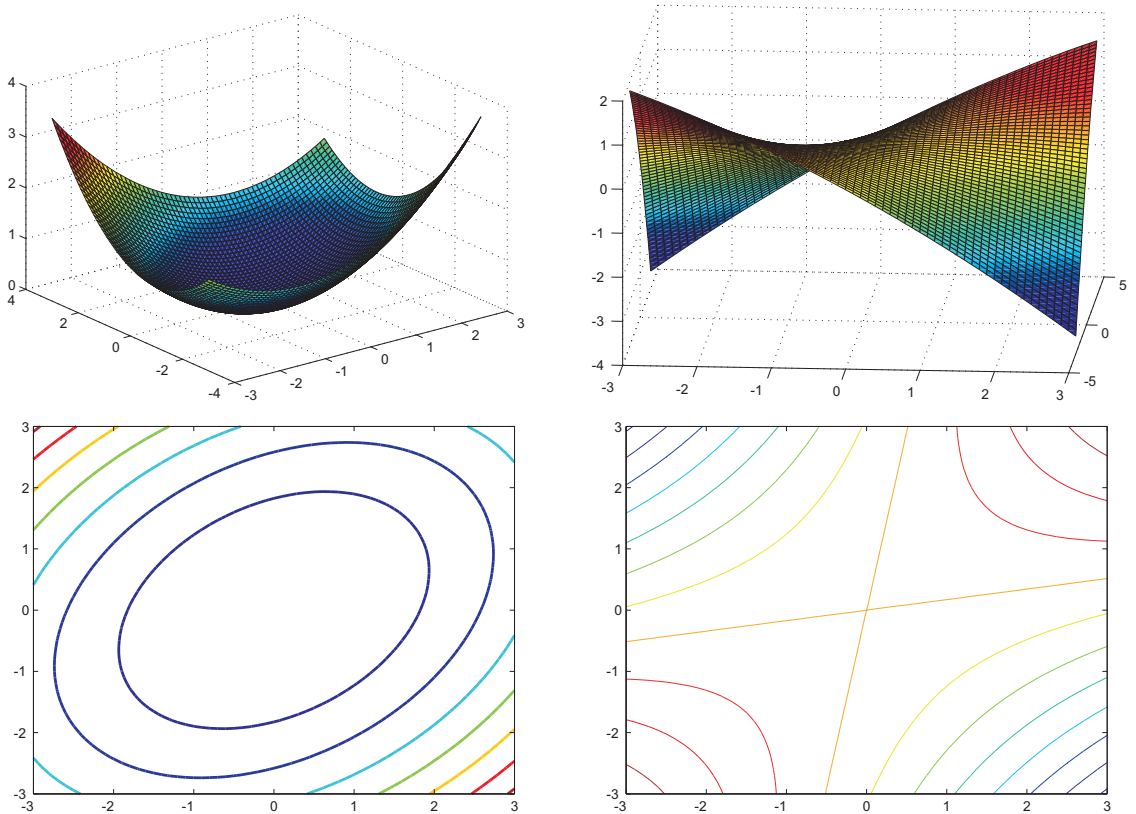
План лекции

Ликбез

Метод релевантных собственных векторов

Неотрицательно определенные матрицы I

$$A \in \mathbb{R}^{n\times n},\ A = A^T \leftrightarrow f(\boldsymbol{x}) = \langle A\boldsymbol{x}, \boldsymbol{x}\rangle = \boldsymbol{x}^T A\boldsymbol{x}$$



Канонический вид квадратичной формы:

$$\exists P:\ P^T = P^{-1},\ \det(P) \neq 0,\ \boldsymbol{y} = P\boldsymbol{x},\ f(\boldsymbol{y}) = \sum_{j=1}^{n} \lambda_j y_j^2,\ A = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}$$

Неотрицательно определенные матрицы II

Матрица — неотрицательно определена, если соответствующая ей квадратичная форма всегда неотрицательна:

$$\forall \boldsymbol{x} \in \mathbb{R}^n \ \boldsymbol{x}^T A \boldsymbol{x} \geq 0 \text{ и } \ = 0 \Leftrightarrow \boldsymbol{x} = 0.$$

У неотрицательно определенной матрицы A все собственные значения $\lambda_j \geq 0$
Количество ненулевых λ_j называется рангом матрицы A.

Линейная регрессия I

Выборка $(X, \boldsymbol{t}) = \{\boldsymbol{x}_n, t_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^d$, $t_n \in \mathbb{R}$
Решающее правило:

$$y(\boldsymbol{x}) = \sum_{j=1}^{K} w_j \phi_j(\boldsymbol{x})$$

Обучение в детерминистской постановке:

$$\sum_{n=1}^{N} (t_n - y(\boldsymbol{x}_n))^2 + \lambda \|\boldsymbol{w}\|^2 = \sum_{n=1}^{N} \left(t_n - \sum_{j=1}^{K} w_j \phi_j(\boldsymbol{x}_n) \right)^2 + \lambda \|\boldsymbol{w}\|^2 =$$

$$= \|\boldsymbol{t} - \Phi \boldsymbol{w}\|^2 + \lambda \|\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}$$

$$\boldsymbol{w}_{opt} = (\Phi^T \Phi + \lambda I)^{-1} \Phi^T \boldsymbol{t}$$

Линейная регрессия II

Обучение в вероятностной постановке:

$$p(t|\boldsymbol{x},\boldsymbol{w}) = \mathcal{N}(\boldsymbol{t}|y(\boldsymbol{x}),\sigma^2),\ p(\boldsymbol{t}|X,\boldsymbol{w}) = \prod_{n=1}^{N} p(t_n|\boldsymbol{x}_n,\boldsymbol{w}),\ p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0,\lambda I)$$

$$p(\boldsymbol{w}|X,\boldsymbol{t}) \to \max_{\boldsymbol{w}} \Leftrightarrow p(\boldsymbol{t}|X,\boldsymbol{w})p(\boldsymbol{w}) \to \max_{\boldsymbol{w}} \Leftrightarrow$$

$$\log p(\boldsymbol{t}|X,\boldsymbol{w}) + \log p(\boldsymbol{w}) \to \max_{\boldsymbol{w}} \Leftrightarrow \frac{1}{2\sigma^2}\|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2 + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}$$

$$\boldsymbol{w}_{MP} = \left(\frac{1}{\sigma^2}\Phi^2\Phi + \lambda I\right)^{-1}\frac{1}{\sigma^2}\Phi^T\boldsymbol{t}$$

Логистическая регрессия

Выборка $(X, \boldsymbol{t}) = \{\boldsymbol{x}_n, t_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^d$, $t_n \in \{-1, 1\}$
Решающее правило:

$$y(\boldsymbol{x}) = \text{sign}(f(\boldsymbol{x})) = \text{sign}\left(\sum_{j=1}^K w_j \phi_j(\boldsymbol{x})\right)$$

Обучение в детерминистской постановке:

$$\sum_{n=1}^N \log(1 + \exp(-t_n f(\boldsymbol{x}_n))) + \lambda \|\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}$$

Обучение в вероятностной постановке:

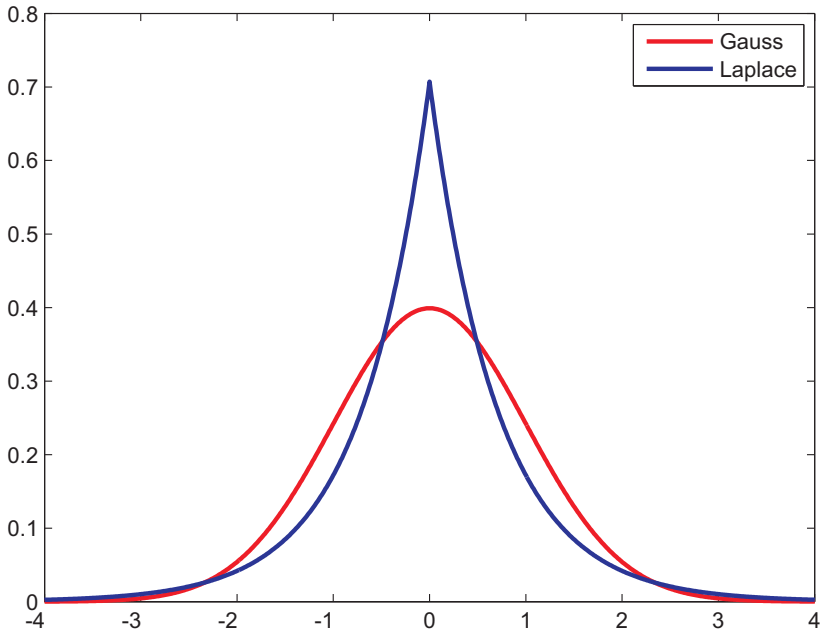
$$p(t|\boldsymbol{x}, \boldsymbol{w}) = \frac{1}{1 + \exp(-tf(\boldsymbol{x}))},\ p(\boldsymbol{t}|X, \boldsymbol{w}) = \prod_{n=1}^N p(t_n|\boldsymbol{x}_n, \boldsymbol{w}),\ p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \lambda I)$$

$$p(\boldsymbol{w}|X, \boldsymbol{t}) \to \max_{\boldsymbol{w}} \quad \Leftrightarrow \quad \sum_{n=1}^N \log(1 + \exp(-t_n f(\boldsymbol{x}_n))) + \frac{\lambda}{2}\|\boldsymbol{w}\|^2 \to \min_{\boldsymbol{w}}$$

Распределение Лапласа I

Плотность распределения:

$$\mathcal{L}(x|\lambda) = \frac{\lambda}{4}\exp\left(-\frac{\lambda}{2}|x|\right)$$

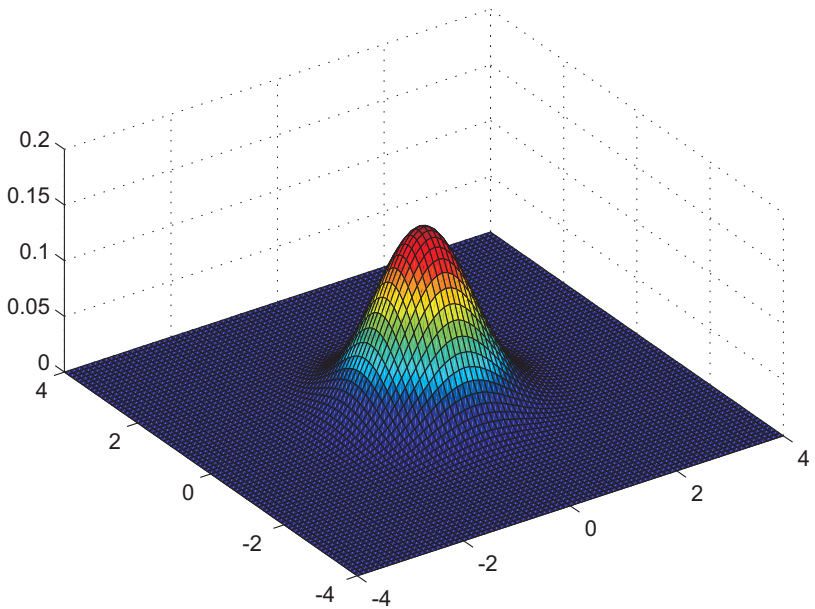


Распределение Лапласа имеет сингулярность в нуле и более тяжелые хвосты, чем нормальное распределение

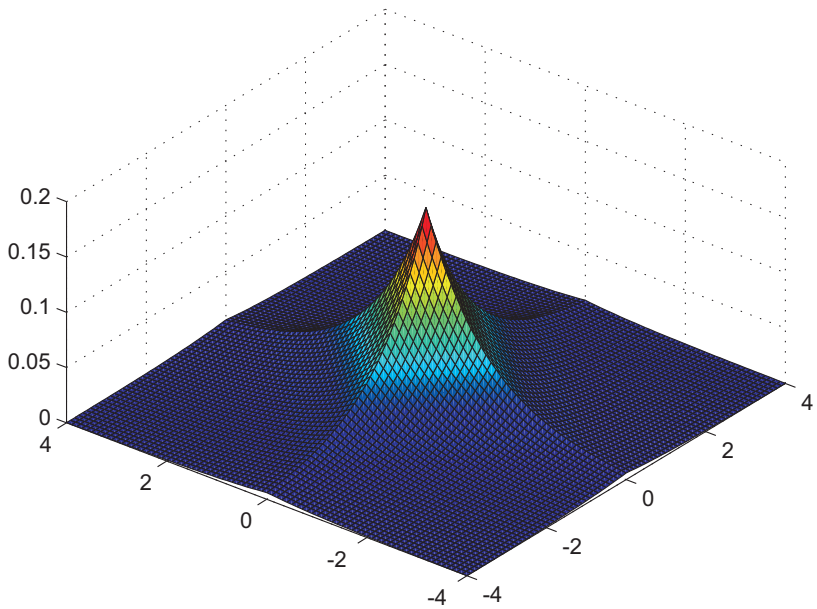
Распределение Лапласа II

Многомерное распределение Лапласа:

$$\mathcal{L}(\boldsymbol{x}|\lambda) = \left(\frac{\lambda}{4}\right)^d \exp\left(-\frac{\lambda}{2}\sum_{j=1}^{d}|x_j|\right)$$



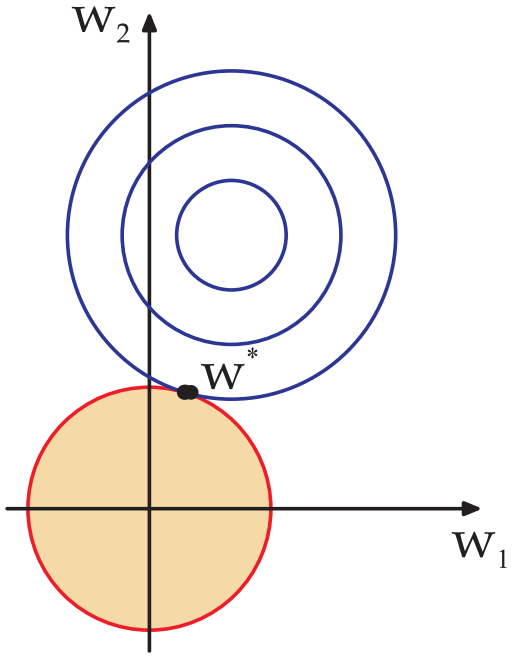
Нормальное распределение
$\mathcal{N}(\boldsymbol{x}|\lambda) \propto \exp\left(-\frac{\lambda}{2}\|\boldsymbol{x}\|_{\ell_2}^2\right)$

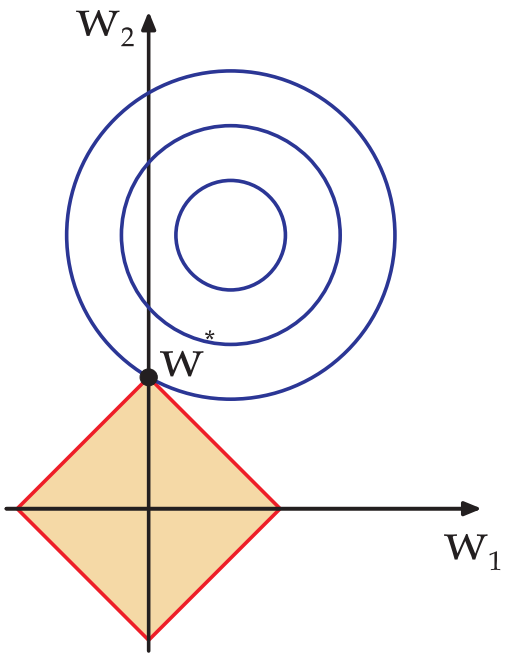


Распределение Лапласа
$\mathcal{L}(\boldsymbol{x}|\lambda) \propto \exp\left(-\frac{\lambda}{2}\|\boldsymbol{x}\|_{\ell_1}\right)$

Лапласовская регуляризация

$$\log p(\boldsymbol{t}|\boldsymbol{w}, X) - \frac{\alpha}{2}\sum_{j=1}^{K}|w_j| \to \max_{\boldsymbol{w}} \quad \leftrightarrow \quad \begin{array}{l}\log p(\boldsymbol{t}|\boldsymbol{w}, X) \to \max\limits_{\boldsymbol{w}} \\ -\frac{1}{2}\sum_{j=1}^{K}|w_j| \le \gamma\end{array}$$





Метод релевантных векторов

Регуляризатор:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \sqrt{\frac{\det(A)}{(2\pi)^K}} \exp\left(-\frac{1}{2}\boldsymbol{w}^T A \boldsymbol{w}\right),$$

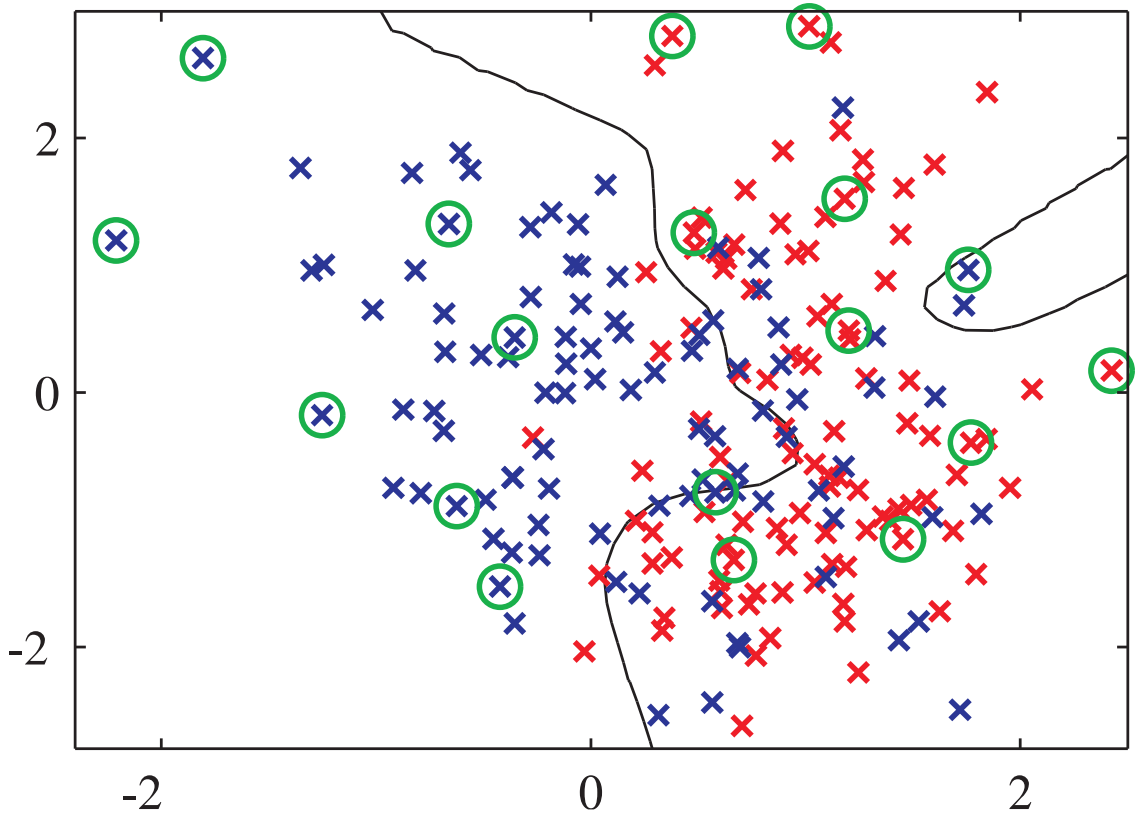
где $A = \operatorname{diag}(\alpha_1, \ldots, \alpha_K)$, $\alpha_j \geq 0$.

Для определения значений коэффициентов регуляризации используется принцип наибольшей обоснованности

$$\boldsymbol{\alpha}_{ME} = \arg\max_{\boldsymbol{\alpha}} p(\boldsymbol{t}|X, \boldsymbol{\alpha}) = \arg\max_{\boldsymbol{\alpha}} \int p(\boldsymbol{t}|X, \boldsymbol{w}) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w}.$$

Разреженность RVM

В процессе оптимизации обоснованности большинство α_j уходят в $+\infty$, исключая нерелевантные базисные функции из решающего правила.



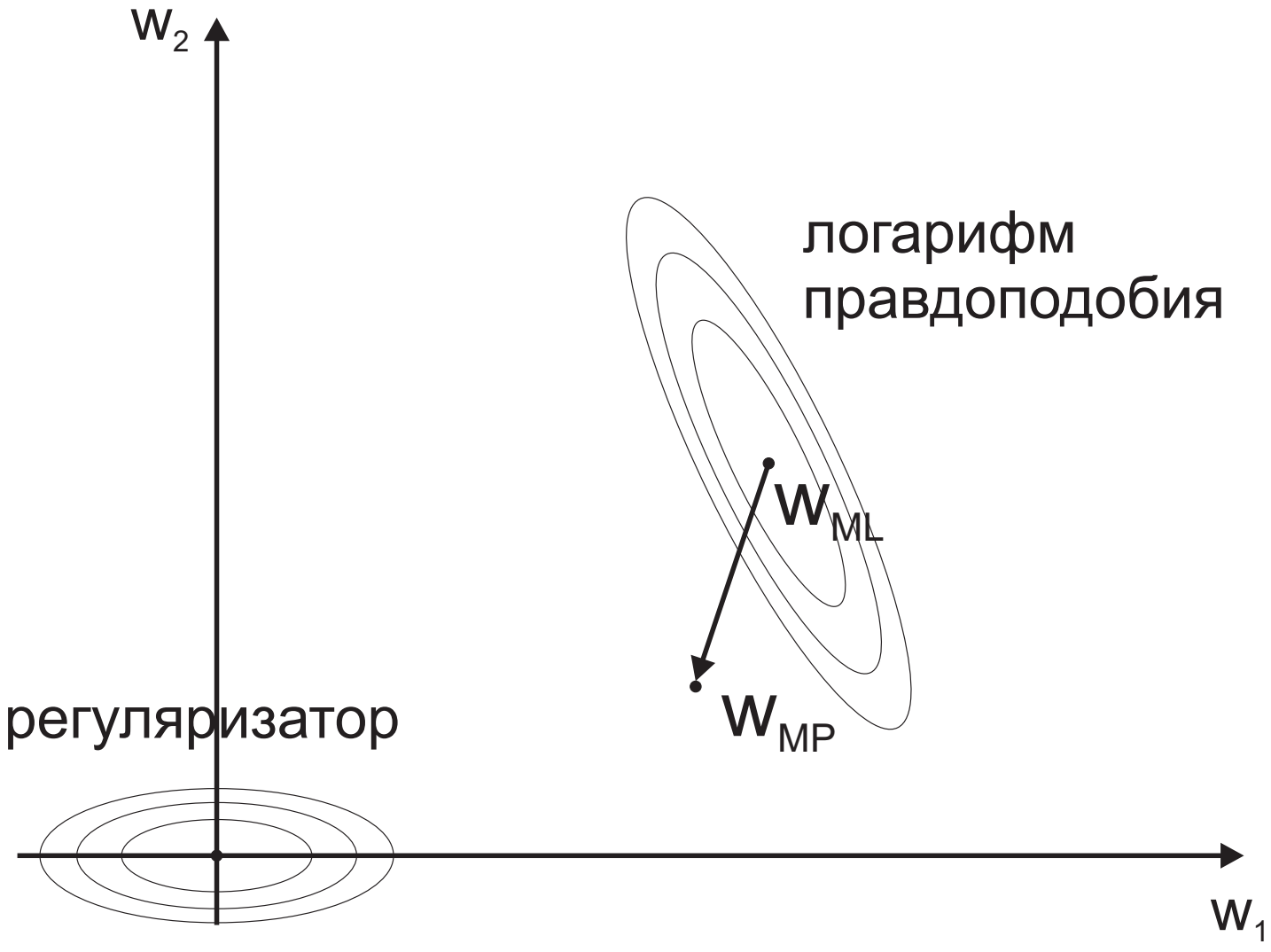
Недостатки RVM

- При обучении классификатора RVM требуется порядка 20–50 итераций для настройки $\boldsymbol{\alpha}$, на каждой из которых приходится обучать метод регуляризованной линейной/логистической регрессии
- RVM не может быть напрямую применен для лапласовского априорного распределения на веса $\boldsymbol{w}$. В то же время известно, что лапласовское априорное распределение обычно приводит к более разреженным решающим правилам
- И регуляризованная линейная/логистическая регрессия, и метод релевантных векторов не инвариантны относительно линейных преобразований базисных функций

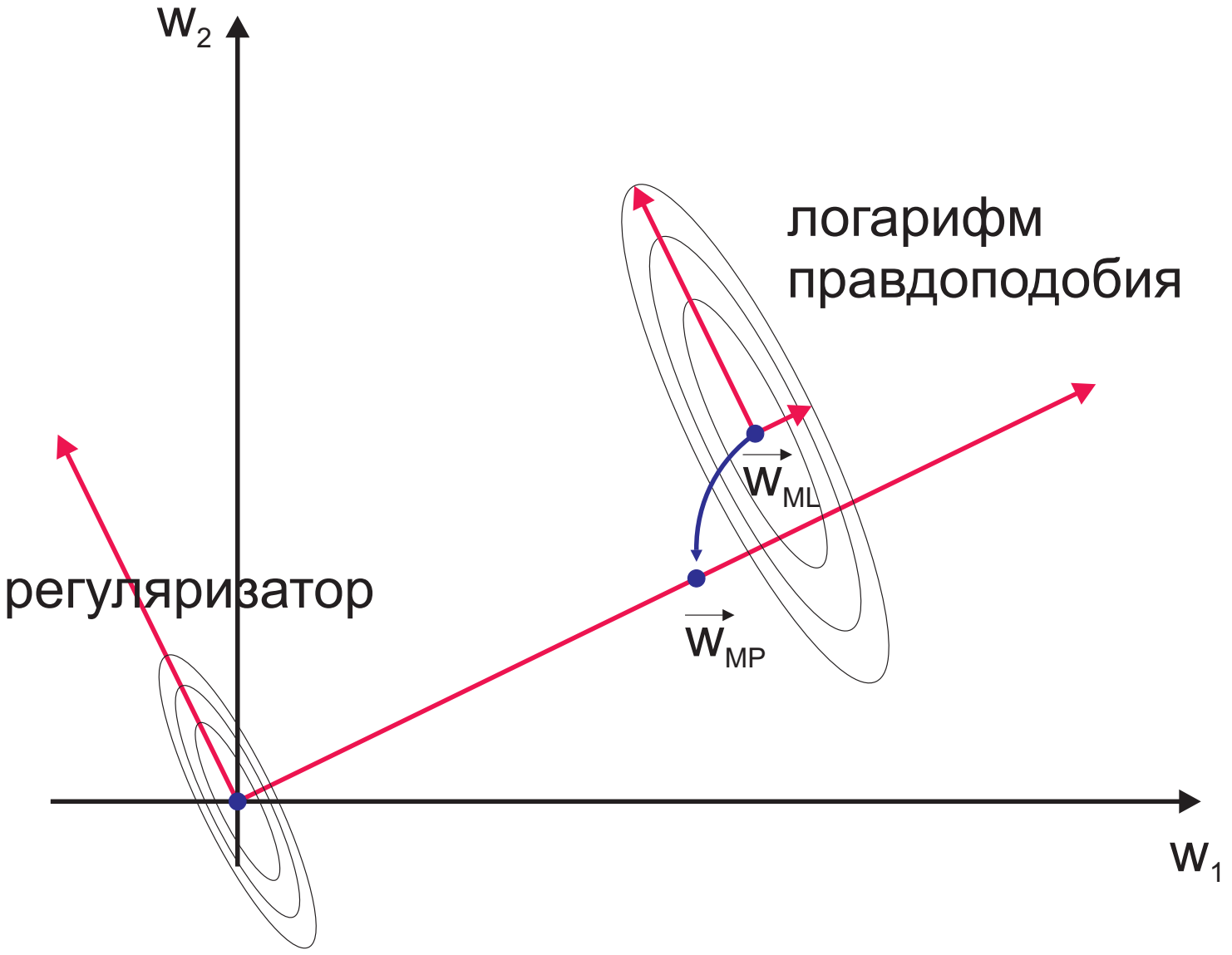
Линейная неинвариантность

- Рассмотрим невырожденную матрицу $L \in \mathbb{R}^{m \times m}$
- Пусть $\boldsymbol{\psi}(\boldsymbol{x}) = L\boldsymbol{\phi}(\boldsymbol{x})$ — новое множество базисных функций
- Поскольку наш классификатор линеен по базисным функциям, вполне естественно требовать, чтобы классификатор, обученный по базисным функциям $\boldsymbol{\psi}(\boldsymbol{x})$, был эквивалентен классификатору, полученному при использовании базисных функций $\boldsymbol{\phi}(\boldsymbol{x})$
- К сожалению, это не так в случае RVM и регуляризованной логистической регрессии

Регуляризация в методе релевантных векторов



Недиагональная регуляризация



Недиагональная регуляризация

$$p(\boldsymbol{t}|X,\boldsymbol{w}) \approx p(\boldsymbol{t}|X,\boldsymbol{w}_{ML}) \exp\left(-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_{ML})^T H (\boldsymbol{w}-\boldsymbol{w}_{ML})\right)$$

$$\exists Q: \; Q^T = Q^{-1}, \; QHQ^T = \Lambda = \operatorname{diag}\{h_1,\ldots,h_K\}$$

Пусть $\boldsymbol{u}$ — новые переменные, ассоциированные с собственными векторами гессиана логарифма правдоподобия: $\boldsymbol{u} = Q\boldsymbol{w}$, $\boldsymbol{w} = Q^T\boldsymbol{u}$.

Правдоподобие:

$$p(\boldsymbol{t}|X,\boldsymbol{u}) = p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \exp\left(-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{u}_{ML})^T \Lambda (\boldsymbol{u}-\boldsymbol{u}_{ML})\right)$$

Регуляризатор:

$$p(\boldsymbol{u}|\boldsymbol{\alpha}) = \sqrt{\frac{\det(A)}{(2\pi)^K}} \exp\left(-\frac{1}{2}\boldsymbol{u}^T A \boldsymbol{u}\right), \; A = \operatorname{diag}(\alpha_1,\ldots,\alpha_K), \; \alpha_j \geq 0.$$

Сепарабельные функции в обоснованности

Обоснованность:

$$\int p(\boldsymbol{t}|X,\boldsymbol{u})p(\boldsymbol{u}|\boldsymbol{\alpha})d\boldsymbol{u} \to \max_{\boldsymbol{\alpha}}$$

Сепарабельность правдоподобия:

$$p(\boldsymbol{t}|X,\boldsymbol{u}) = p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \exp\left(-\frac{1}{2}(\boldsymbol{u}-\boldsymbol{u}_{ML})^T \Lambda (\boldsymbol{u}-\boldsymbol{u}_{ML})\right) =$$

$$p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \exp\left(-\frac{1}{2}\sum_{j=1}^{K} h_j (u_j - u_{ML,j})^2\right) = p(\boldsymbol{t}|X,\boldsymbol{u}_{ML}) \prod_{j=1}^{K} \underbrace{\exp\left(-\frac{1}{2} h_j (u_j - u_{ML,j}\right.}_{f(u_j)}$$

Сепарабельность априорного распределения: $p(\boldsymbol{u}|\boldsymbol{\alpha}) = \prod_{j=1}^{K} p(u_j|\alpha_j)$

Обоснованность: $\int p(\boldsymbol{t}|X,\boldsymbol{u})p(\boldsymbol{u}|\boldsymbol{\alpha})d\boldsymbol{u} = \prod_{j=1}^{K} \underbrace{\int f(u_j)p(u_j|\alpha_j)du_j}_{g(\alpha_j)}$

Гауссовское априорное распределение

$$p(u_j|\alpha_j) = \mathcal{N}(u_j|0, \alpha_j^{-1})$$

$$g^G(\alpha_j) = \sqrt{\frac{\alpha_j}{2\pi}} \int \exp\left(-\frac{h_j}{2}(u_j - u_{ML,j})^2 - \frac{\alpha_j}{2}u_j^2\right) du_j =$$

$$\sqrt{\frac{\alpha_j}{h_j + \alpha_j}} \exp\left(-\frac{h_j \alpha_j u_{ML,j}^2}{2(h_j + \alpha_j)}\right),$$

$$\alpha_j^* = \begin{cases} \frac{h_j}{h_j u_{ML,j}^2 - 1}, & \text{если } h_j u_{ML,j}^2 > 1 \\ +\infty, & \text{иначе} \end{cases}$$

В данном случае условие на релевантность степени свободы можно получить в явном виде.

Метод релевантных собственных векторов

Вход: Обучающая выборка $\{\boldsymbol{x}_i, t_i\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \{+1, -1\}$;
Матрица обобщенных признаков $\Phi = \{\phi_j(\boldsymbol{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов $\boldsymbol{w}_{MP}$ для решающего правила
$t_*(\boldsymbol{x}) = \operatorname{sign}\left(\sum_{j=1}^m w_{MP,j}\phi_j(\boldsymbol{x})\right)$;

1: Найти $\boldsymbol{w}_{ML} = \arg\max p(\boldsymbol{t}|X, \boldsymbol{w})$;
2: Вычислить гессиан $H = \nabla\nabla \log p(\boldsymbol{t}|X, \boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{ML}}$;
3: Вычислить собственные вектора и собственные значения гессиана $-H = Q^T \Lambda Q$, $\Lambda = \operatorname{diag}(h_1, \ldots, h_m)$;
4: Вычислить $\boldsymbol{u}_{ML} = Q\boldsymbol{w}_{ML}$;
5: для $j = 1, \ldots, m$
6: если $h_j u_{ML,j}^2 > 1$ то
7: $\alpha_j^* := \frac{h_j}{h_j u_{ML,j}^2 - 1}$;
8: иначе
9: $\alpha_j^* := +\infty$
10: Найти $\boldsymbol{w}_{MP} = \arg\max p(\boldsymbol{t}|X, \boldsymbol{w}) p(Q\boldsymbol{w}|\boldsymbol{\alpha}^*)$

Лапласовское априорное распределение

$$p(u_j|\alpha_j) = \mathcal{L}(u_j|\alpha_j^{-1})$$

Интеграл обоснованности может быть вычислен аналитически, разобьем интеграл на два

$$g^L(\alpha_j) = \frac{\alpha_j}{4}\int_{-\infty}^{0} \exp\left(-\frac{h_j(u_j - u_{ML,j})^2}{2} + \frac{\alpha_j}{2}u_j\right) du_j +$$

$$\frac{\alpha_j}{4}\int_{0}^{+\infty} \exp\left(-\frac{h_j(u_j - u_{ML,j})^2}{2} - \frac{\alpha_j}{2}u_j\right) du_j =$$

$$\frac{\alpha_j}{4}\sqrt{\frac{\pi}{2h_j}} \exp\left(-\frac{h_i u_{ML,i}^2}{2}\right) \times \left[\exp\left(x_1^2\right)\operatorname{erfc}(x_1) + \exp\left(x_2^2\right)\operatorname{erfc}(x_2)\right]$$

Здесь $x_1 = \sqrt{\frac{h_i}{2}}\left(\frac{\alpha_i}{2h_i} - u_{ML,i}\right)$, $x_2 = \sqrt{\frac{h_i}{2}}\left(\frac{\alpha_i}{2h_i} + u_{ML,i}\right)$,

$$\operatorname{erfc}(x) = \frac{2}{\sqrt{\pi}}\int_{x}^{+\infty} \exp\left(-\xi^2\right) d\xi$$

Особенности численного вычисления

$$g^L(\alpha_j) = \frac{\alpha_j}{4}\sqrt{\frac{\pi}{2h_j}}\exp\left(-\frac{h_i u_{ML,i}^2}{2}\right)\times\left[\exp\left(x_1^2\right)\operatorname{erfc}(x_1) + \exp\left(x_2^2\right)\operatorname{erfc}(x_2)\right]$$

При $x_{1,2}$ существенно отличных от нуля, возникают численные трудности с вычислением этого выражения. При $x > 27$ выражение $\exp(x^2) > 10^{300}$, при $x > 26$ выражение $\operatorname{erfc}(x) < 10^{-300}$.

Пусть $x_j \gg 0$, $j \in \{1, 2\}$, тогда можно показать, что

$$\operatorname{erfcx}(x_j) = \exp(x_j^2)\operatorname{erfc}(x_j) \approx 1/(\sqrt{\pi}x_j)$$

При $x_j \ll 0$ объединяем $\exp(-h_i u_{ML,i}^2/2)$ и $\exp(x_j^2)$

$$\exp(-h_i u_{ML,i}^2/2)\exp(x_j^2) = \exp(y_j),$$

где

$$y_{1,2} = \frac{\alpha_i^2}{8h_i} \mp \frac{\alpha_i u_{ML,i}}{2}.$$

Метод релевантных собственных векторов с лапласовским регуляризатором

Вход: Обучающая выборка $\{\boldsymbol{x}_i, t_i\}_{i=1}^n$, $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \{+1, -1\}$;
Матрица обобщенных признаков $\Phi = \{\phi_j(\boldsymbol{x}_i)\}_{i,j=1}^{n,m}$;

Выход: Набор весов $\boldsymbol{w}_{MP}$ для решающего правила
$t_*(\boldsymbol{x}) = \text{sign}\left(\sum_{j=1}^m w_{MP,j}\phi_j(\boldsymbol{x})\right)$;

1: Найти $\boldsymbol{w}_{ML} = \arg\max p(\boldsymbol{t}|X, \boldsymbol{w})$;
2: Вычислить гессиан $H = \nabla\nabla \log p(\boldsymbol{t}|X, \boldsymbol{w})|_{\boldsymbol{w}=\boldsymbol{w}_{ML}}$;
3: Вычислить собственные вектора и собственные значения гессиана $-H = Q^T \Lambda Q$, $\Lambda = \text{diag}(h_1, \ldots, h_m)$;
4: Вычислить $\boldsymbol{u}_{ML} = Q\boldsymbol{w}_{ML}$;
5: для $j = 1, \ldots, m$
6: $\quad \alpha_j^* := \arg\max f_j^L(h_j, u_{ML,j}, \alpha_j)$;
7: Найти $\boldsymbol{u}_{MP} = \arg\max p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{u}|\boldsymbol{\alpha}^*)$;
8: Найти $\boldsymbol{w}_{MP} = Q^T\boldsymbol{u}_{MP}$

Оптимизация функции $g^L(\alpha_j)$

Точка максимума функции $g^L(\alpha_j)$ не может быть выписана в явном виде, поэтому необходима численная оптимизация (впрочем, не слишком обременительная)