Evidence: байесовский подход к выбору моделей

Адуенко Александр

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель:
д.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

22 октября 2015 года

План

- Линейная и логистическая регрессия
- Регуляризаторы: влияние на переобучение и разрешимость
- Априорное распределение, поощряющее разреженность
- Evidence (обоснованность)
- Принцип максимума evidence (обоснованности)
- Evidence для линейной регрессии
- Evidence для логистической регрессии

Линейная регрессия

$\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n \times d}$, $\mathbf{w} \in \mathbb{R}^d$.

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

$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = N(\mathbf{y}|\mathbf{X}\mathbf{w}, \beta^{-1}\mathbf{I})$.

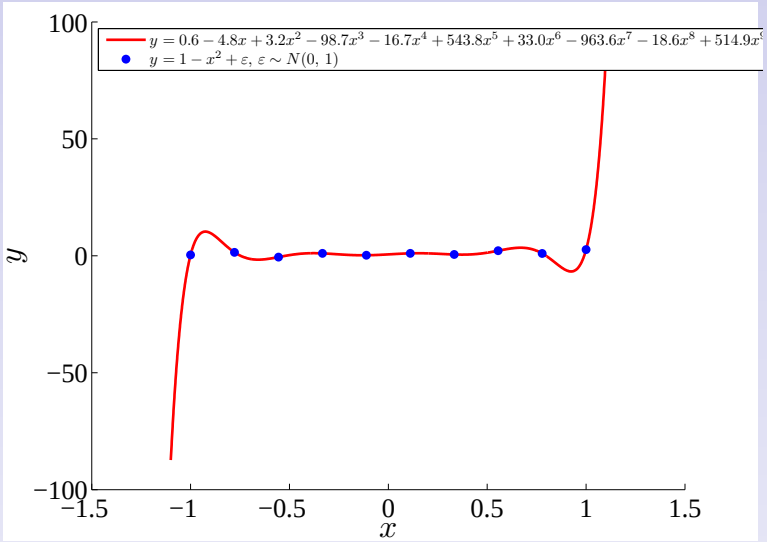
Логистическая регрессия

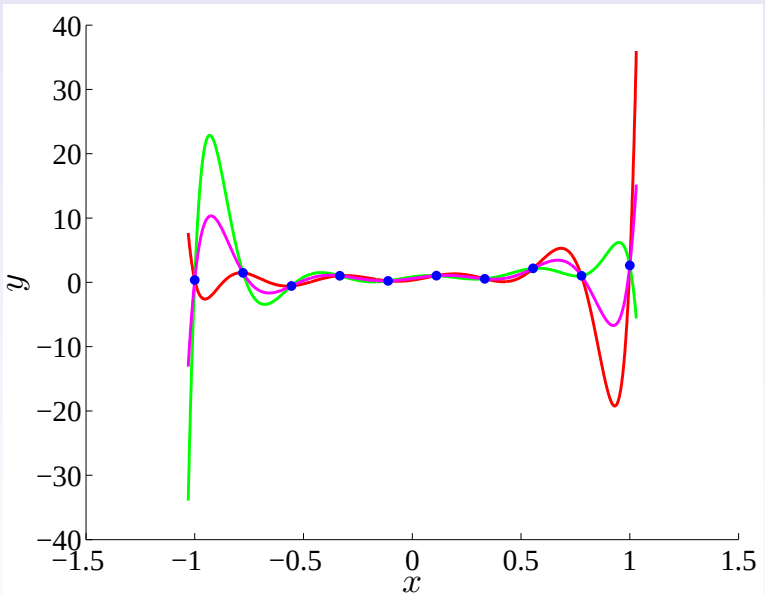
Два класса: $\{-1, 1\}$.

Для объекта $\mathbf{x}_i$ вероятность принадлежать классу y_i есть

$$p(y_i|\mathbf{x}_i, \mathbf{w}) = \frac{1}{1 + \exp(-y_i \mathbf{w}^\mathsf{T} \mathbf{x}_i)}.$$

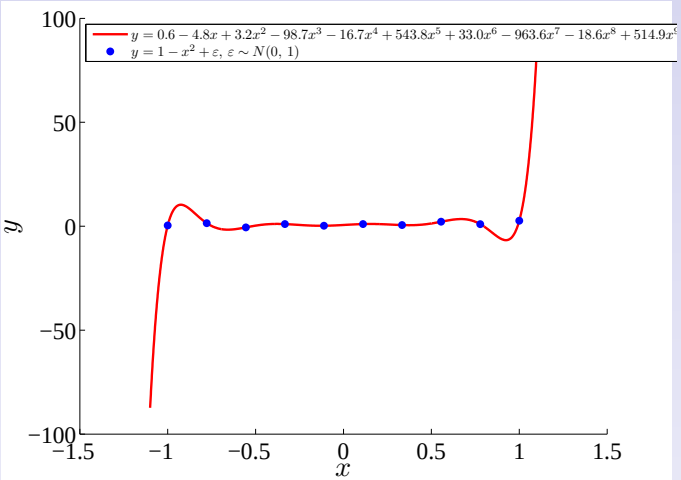
$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \prod_{i=1}^{n} \sigma(y_i \mathbf{w}^\mathsf{T} \mathbf{x}_i).$$

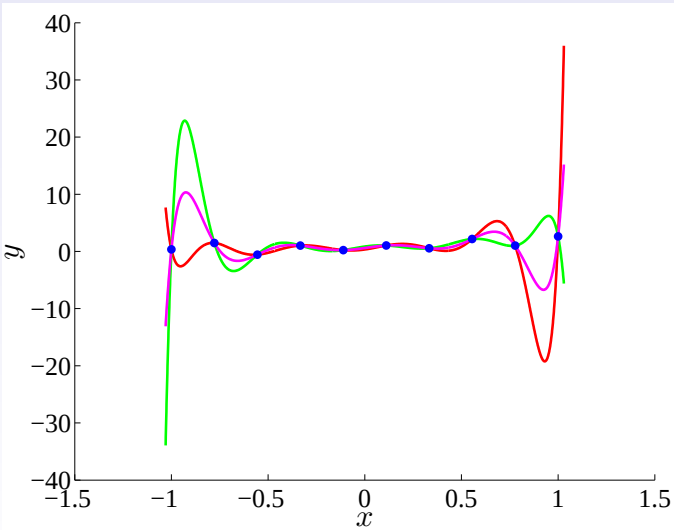




Переобучение и разрешимость

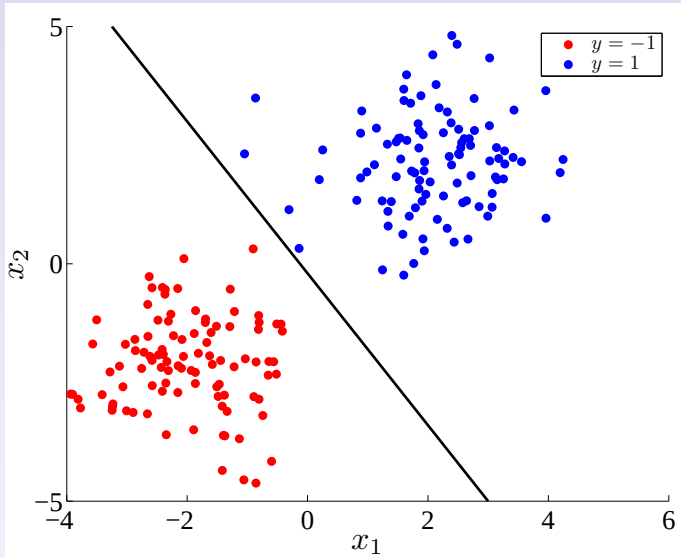
Линейная регрессия и избыточные признаки

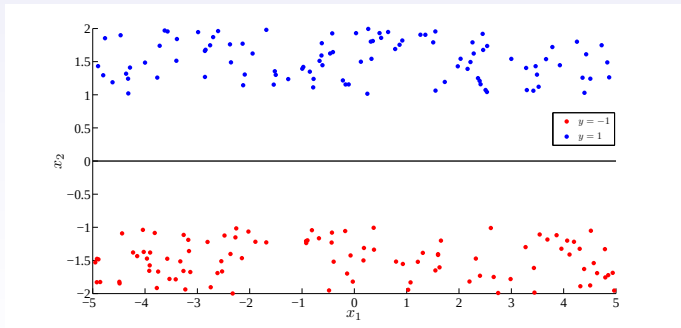




Два линейно-разделимых класса для логистической регрессии

Проблема: оптимальное решение имеет $\|\mathbf{w}\|_2 = \infty$

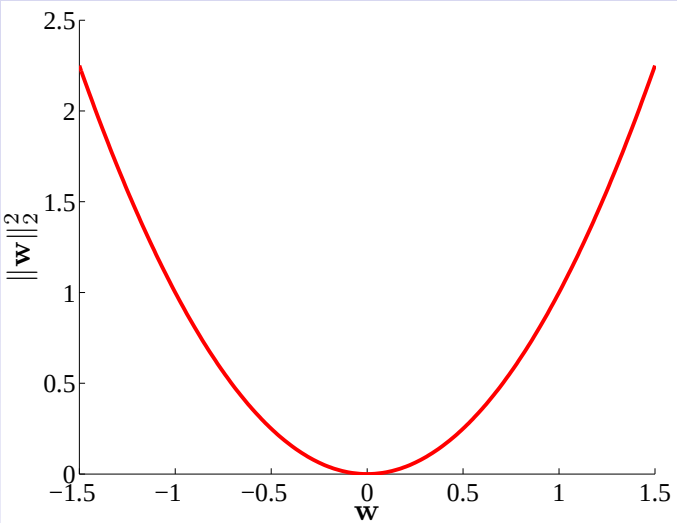




Регуляризация

Квадратическая регуляризация

Prior: $\mathbf{w} \sim N(\mathbf{0},\ \tau^{-1}\mathbf{I})$



Свойства:

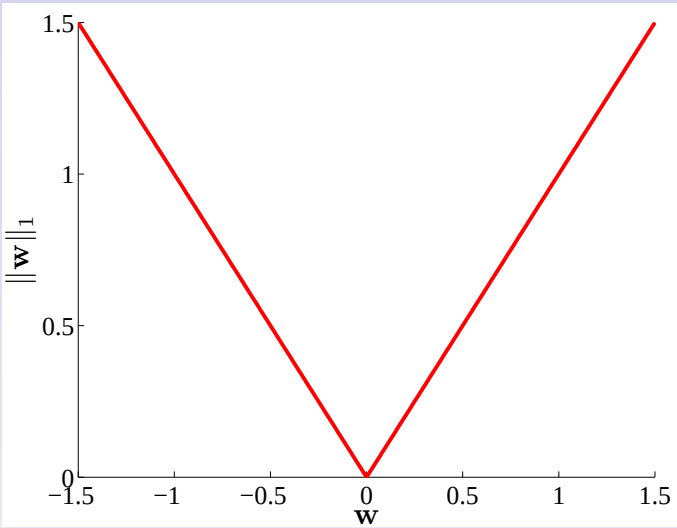
\+ Обе задачи становятся разрешимы

\+ Для линейной регрессии есть аналитическое решение

— Слабо поощряет разреженность

l_1–regularization

Prior: $\mathbf{w} \sim Laplace(\mathbf{0},\ \tau^{-1})$



Свойства:

\+ Обе задачи становятся разрешимы

— Для линейной регрессии нет аналитического решения

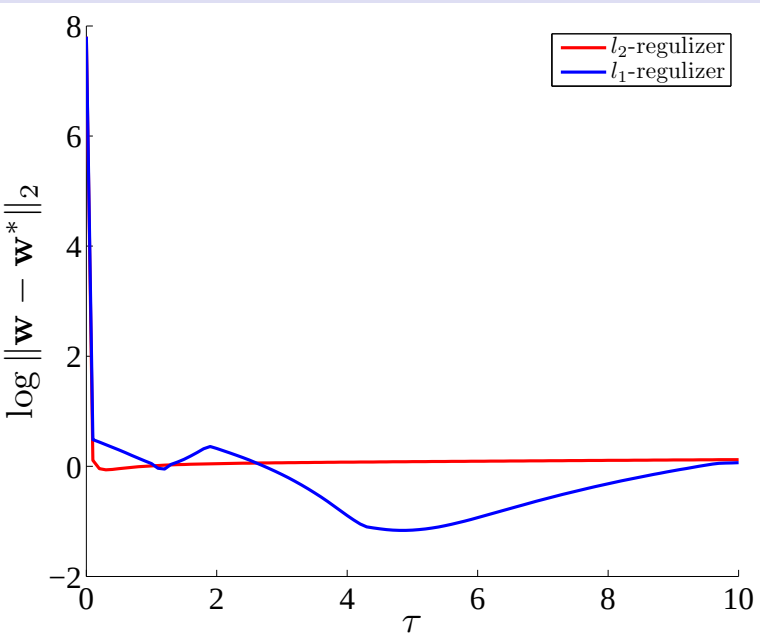
— Недифференцируемая целевая функция

\+ Поощряет разреженность

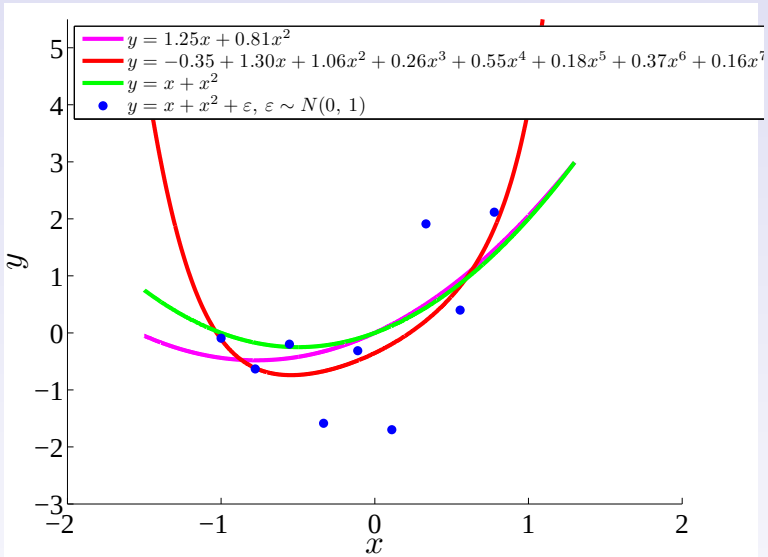
Пример с регрессией на полиномы

Данные

$y = x + x^2 + \varepsilon,\ \varepsilon \sim N(0,\ 1),$
$y_i \sim p(y|x_i),\ i = 1,\ \ldots,\ 10,$ где $x_1,\ \ldots,\ x_{10}$
выбраны равномерно на $[-1,\ 1]$.



Зависимость точности от параметра регуляризации τ

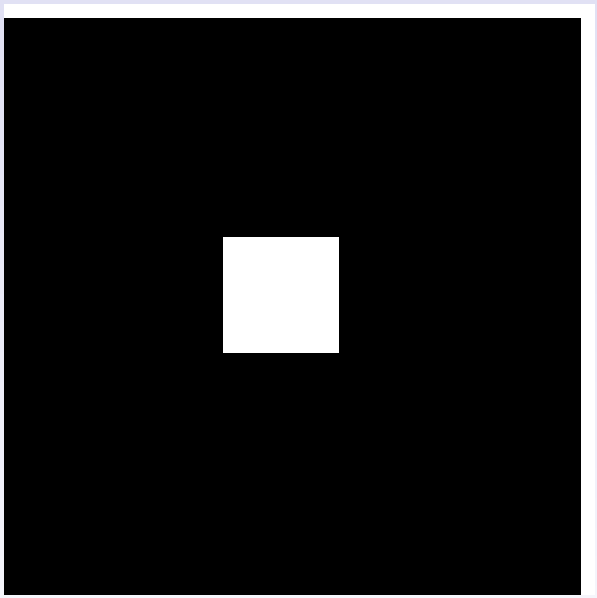


Наилучшие полиномы

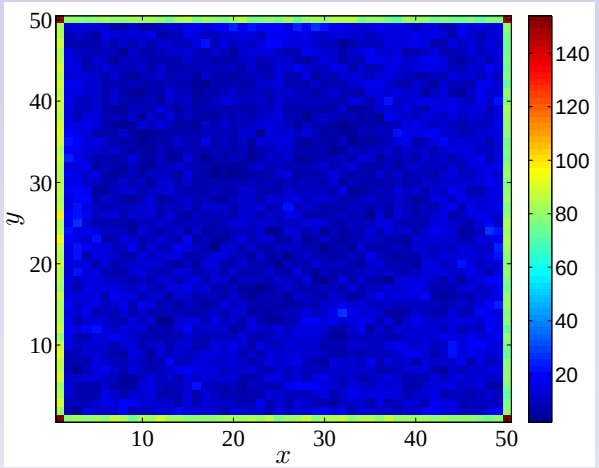
Пример "томография"

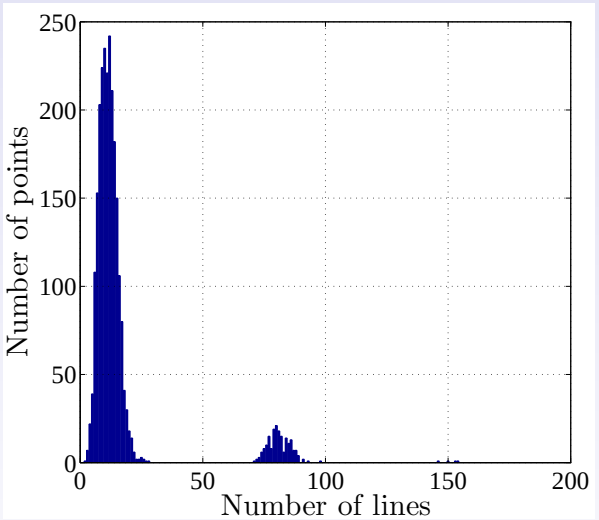
Постановка задачи

$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \beta^{-1}\mathbf{I})$,
$\mathbf{y} \in \mathbb{R}^m$, $\mathbf{X} \in \mathbb{R}^{m \times n^2}$, $m < n^2$.
$\mathbf{w} \in [0, 1]^{n^2}$.
Параметры: $m = 1000$, $n = 50$.



Настоящий $\mathbf{w}$

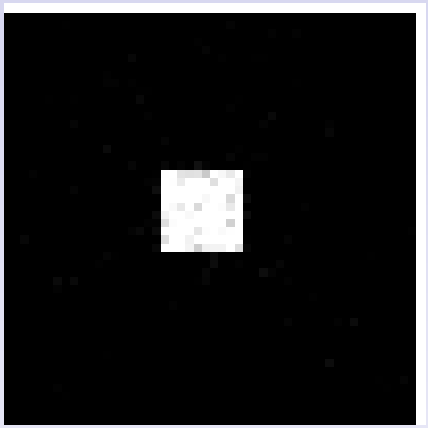




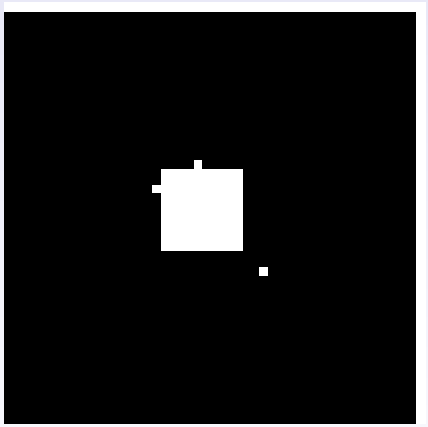
Распределение точек по числу линий

Пример "томография $\beta = 100$

l_1–регуляризация

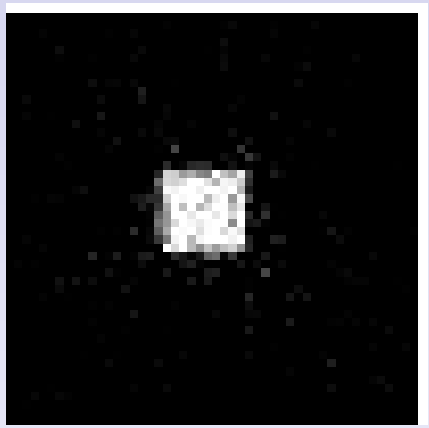


$\hat{\mathbf{w}}$

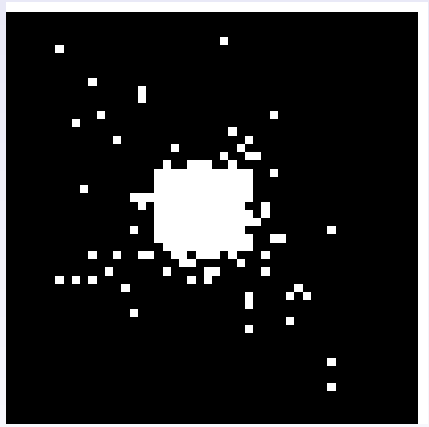


$[\hat{\mathbf{w}} > 0.05]$

Квадратическая регуляризация



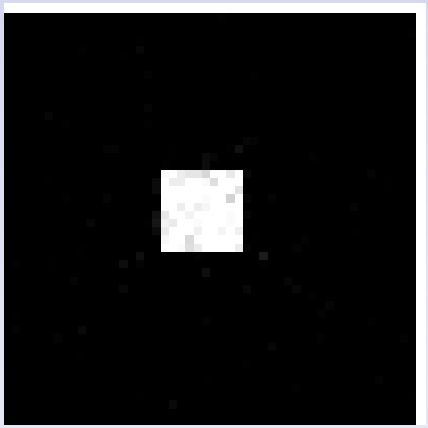
$\hat{\mathbf{w}}$



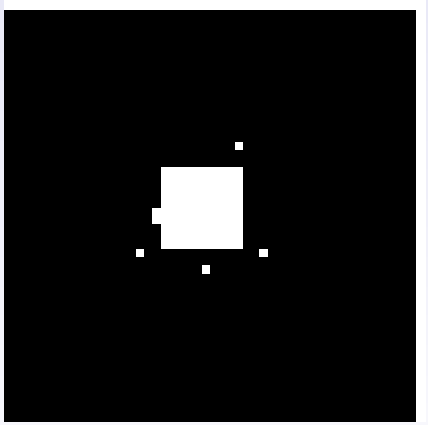
$[\hat{\mathbf{w}} > 0.05]$

Пример "томография $\beta = 4$

l_1–регуляризация

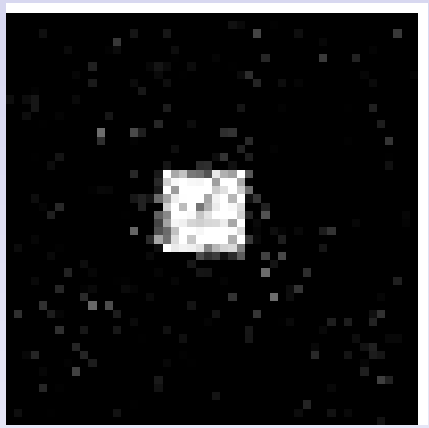


$\hat{\mathbf{w}}$

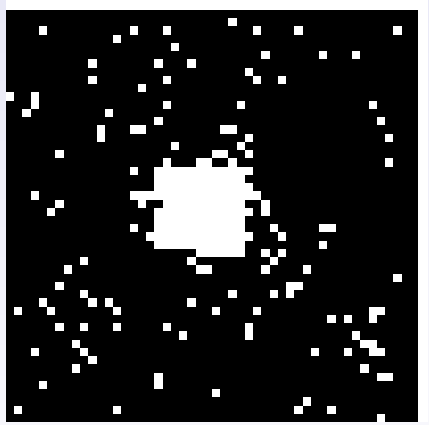


$[\hat{\mathbf{w}} > 0.05]$

Квадратическая регуляризация



$\hat{\mathbf{w}}$



$[\hat{\mathbf{w}} > 0.05]$

Evidence

Модель M_i: $p_i(T,\ \theta|X) = p_i(T|X,\ \theta)p(\theta)$

Шаг	Наблюдаемые	Скрытые	Результат
Обучение	$(X_{\text{train}},\ T_{\text{train}})$	θ	$p(\theta\|X_{\text{train}},\ T_{\text{train}})$
Контроль	X_{test}	T_{test}	$p(T_{\text{test}}\|X_{\text{test}},\ X_{\text{train}},\ T_{\text{train}})$

$$p(\theta|X_{\text{train}},\ T_{\text{train}}) = \frac{p(T_{\text{train}},\ \theta|X_{\text{train}})}{\int p(T_{\text{train}},\ \theta^*|X_{\text{train}})d\theta^*}$$

$$p(T_{\text{test}}|X_{\text{test}},\ X_{\text{train}},\ T_{\text{train}}) = \int p(T_{\text{test}},\ \theta|X_{\text{test}},\ X_{\text{train}},\ T_{\text{train}})$$

$$p(\theta|X_{\text{test}}, X_{\text{train}}, T_{\text{train}})d\theta = \int p(T_{\text{test}}, \theta|X_{\text{test}})p(\theta|X_{\text{train}}, T_{\text{train}})d\theta$$

Evidence.

Модель M_i: $p_i(T, \theta|X) = p_i(T|X, \theta)p_i(\theta)$
Пусть имеется $K > 1$ моделей.

Процесс порождения выборки:

- Природа выбирает модель из K доступных моделей с априорными вероятностями $p(M_i)$, $i = 1, \ldots, K$.
- Для выбранной модели i^* природа сэмплирует вектор параметров θ^* из априорного распределения $p_{i^*}(\theta)$
- Имея i^*, θ^* природа выбирает X_{train} и сэмплирует T_{train} из $p_{i^*}(T|X_{\text{train}}, \theta^*)$
- $(X_{\text{train}}, T_{\text{train}})$ даны наблюдателю.
- Природа выбирает X_{test} и сэмплирует T_{test} из $p_{i^*}(T|X_{\text{test}}, \theta^*)$

Evidence

Модель M_i: $p_i(T,\ \theta|X) = p_i(T|X,\ \theta)p_i(\theta)$
Общая модель M: $p(T,\ \theta,\ M_i|X) = p(M_i)p_i(\theta)p_i(T|X,\ \theta)$

$$p(T_{\text{test}}|X_{\text{test}},\ X_{\text{train}},\ T_{\text{train}}) =$$

$$\sum_{i=1}^{K} p_i(T_{\text{test}}|X_{\text{test}},\ X_{\text{train}},\ T_{\text{train}})p(M_i|X_{\text{test}},\ X_{\text{train}},\ T_{\text{train}}) =$$

$$\sum_{i=1}^{K} p_i(T_{\text{test}}|X_{\text{test}},\ X_{\text{train}},\ T_{\text{train}})p(M_i|X_{\text{train}},\ T_{\text{train}})$$

$$p(M_i|X_{\text{train}}, T_{\text{train}}) = \frac{p(T_{\text{train}},\ M_i|X_{\text{train}})}{P(T_{\text{train}}|X_{\text{train}})} \propto p(T_{\text{train}}, M_i|X_{\text{train}}) =$$

$$\int p(T_{\text{train}},\ \theta,\ M_i|X_{\text{train}})d\theta = p(M_i)p_i(T_{\text{train}}|X_{\text{train}})$$

Пример выбора модели

a – applicant, r – reviewer

$$a,\ r = \begin{cases} 0, \text{ нет PhD}, \\ 1, \text{ PhD}. \end{cases}$$

d – decision

$$d = \begin{cases} 1, \text{ принять}, \\ 0, \text{ отвергнуть}. \end{cases}$$

$r = 0$	$d = 0$	$d = 1$
$a = 0$	9	0
$a = 1$	132	19

$r = 1$	$d = 0$	$d = 1$
$a = 0$	97	6
$a = 1$	52	11

Случаи:

1. $p(d|a,\ r) = p(d)$
2. $p(d|a,\ r) = p(d|a)$
3. $p(d|a,\ r) = p(d|r)$
4. $p(d|a,\ r) = p(d|a,\ r)$

Пример выбора модели

1) $p(d|a,\ r) = p(d)$

Поэтому $p(d|\theta) = \mathrm{Be}(\theta)$. **Prior** : $p(\theta) = U[0,\ 1]$

$$p(T|X) = \int p(T|X,\ \theta)p(\theta)d\theta = \int_0^1 C_9^0(1-\theta)^9$$

$$C_{103}^{97}\theta^6(1-\theta)^{97}C_{151}^{132}\theta^{19}(1-\theta)^{132}C_{63}^{52}\theta^{11}(1-\theta)^{52}d\theta = 2.8\cdot10^{-51}CCCC$$

2) $p(d|a,\ r) = p(d|a)$

Поэтому $p(d|a=0) = \mathrm{Be}(\theta_1)$, $p(d|a=1) = \mathrm{Be}(\theta_2)$.

Prior : $p(\theta_1) = U[0,\ 1]$, $p(\theta_2) = U[0,\ 1]$

$$p(T|X) = \int p(T|X,\theta_1,\theta_2)p(\theta_1)p(\theta_2)d\theta_1 d\theta_2 = \int_0^1\int_0^1 C_9^0(1-\theta_1)^9C_{103}^{97}$$

$$\theta_1^6(1-\theta_1)^{97}C_{151}^{132}\theta_2^{19}(1-\theta_2)^{132}C_{63}^{52}\theta_2^{11}(1-\theta_2)^{52}d\theta_1 d\theta_2 = 4.7\cdot10^{-51}CCCC$$

Пример выбора модели

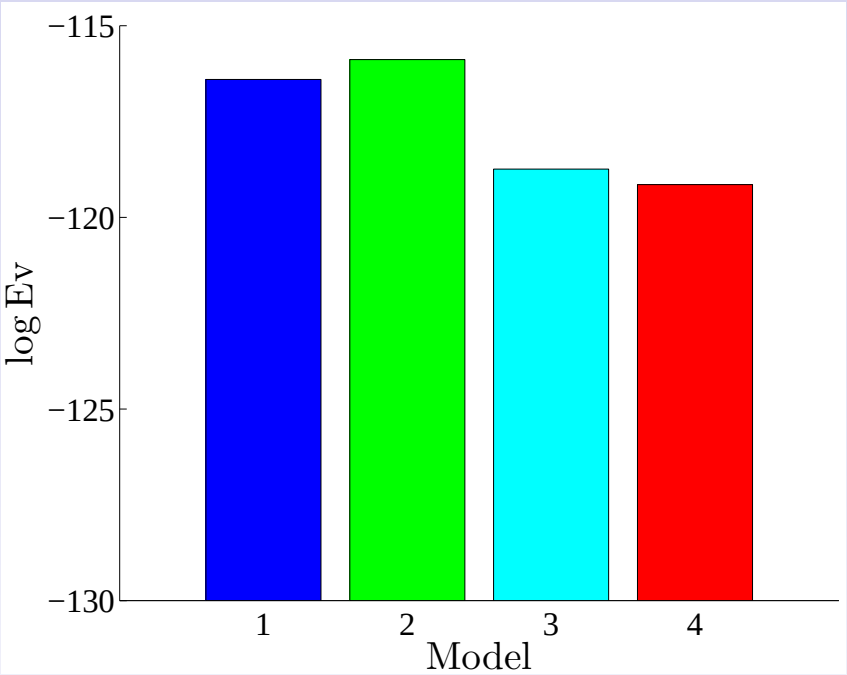
3)$p(d|a,\ r) = p(d|r)$
Поэтому $p(d|r = 0) = \mathrm{Be}(\theta_1)$, $p(d|r = 1) = \mathrm{Be}(\theta_2)$.
Prior : $p(\theta_1) = U[0,\ 1]$, $p(\theta_2) = U[0,\ 1]$

$$p(T|X) = 0.27 \cdot 10^{-51} CCCC$$

4) $p(d|a,\ r) = p(d|a,\ r)$
Поэтому $p(d|a = 0,\ r = 0) = \mathrm{Be}(\theta_1)$, $p(d|a = 0,\ r = 1) = \mathrm{Be}(\theta_2)$,
$p(d|a = 1,\ r = 0) = \mathrm{Be}(\theta_3)$, $p(d|a = 1,\ r = 1) = \mathrm{Be}(\theta_4)$.
Prior : $p(\theta_1) = U[0,\ 1]$, $p(\theta_2) = U[0,\ 1]$,
$p(\theta_3) = U[0,\ 1]$, $p(\theta_4) = U[0,\ 1]$

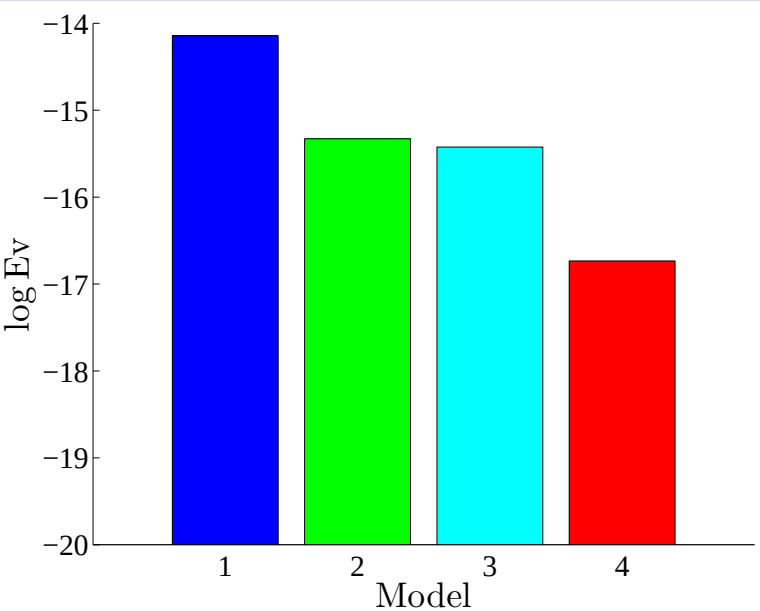
$$p(T|X) = 0.18 \cdot 10^{-51} CCCC$$

Пример выбора модели

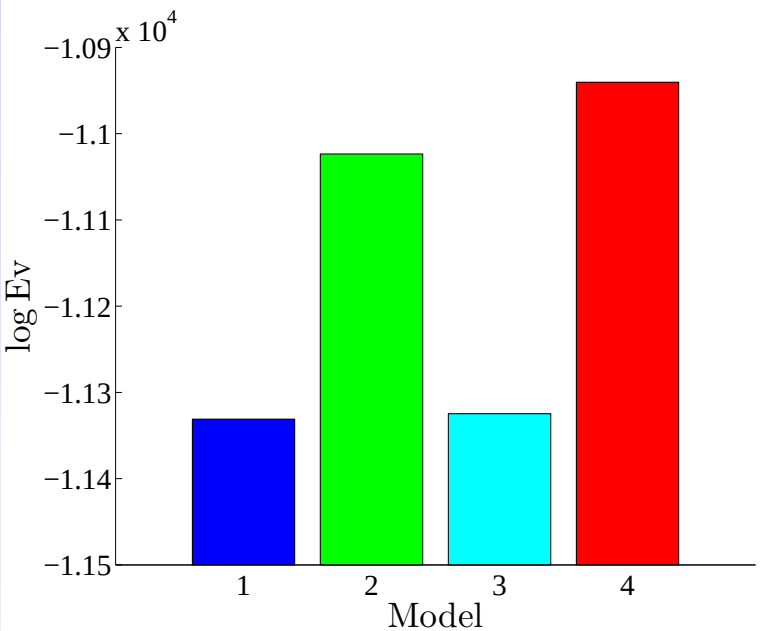


Сравнение обоснованностей, 326 объектов в выборке

Model selection example



Сравнение обоснованностей, 33 объекта в выборке



Сравнение обоснованностей, 32600 объектов в выборке

Простое понимание обоснованности

$$\text{Evidence} : p_i(T|X) = \int p_i(T|X,\ \theta)p_i(\theta)d\theta$$

$$p_i(\theta|X,\ T) = \frac{p_i(T|X,\theta)p_i(\theta)}{p(T|X)}.$$

Предположения:

- θ одномерный
- Априорное распределение $p_i(\theta)$ плоское с шириной $\Delta\theta_{\text{prior}}$
- Апостериорное распределение $p_i(\theta|X,\ T)$ сконцентрировано вокруг θ_{MP} с шириной $\Delta\theta_{\text{post}}$

Тогда: $\log p_i(T|X) \approx \log p_i(T|X,\ \theta_{MP}) + \log\left(\frac{\Delta\theta_{\text{post}}}{\Delta\theta_{\text{prior}}}\right).$

Для M-мерного θ:

$$\log p_i(T|X) \approx \log p_i(T|X,\ \theta_{MP}) + M\log\left(\frac{\Delta\theta_{\text{post}}}{\Delta\theta_{\text{prior}}}\right).$$

Evidence для линейной регрессии

$T = X\theta + \varepsilon,\ \theta \sim N(\theta|\mathbf{0},\ \alpha^{-1}\mathbf{I}),\ \varepsilon \sim N(\mathbf{0},\ \beta^{-1}\mathbf{I})$

Совместное правдоподобие:

$p(T,\ \theta|X,\ \alpha,\ \beta) = p(T|X,\ \theta,\ \beta)p(\theta|\alpha).$

Evidence: $p(T|X,\ \alpha,\ \beta)$

$T|X,\ \alpha,\ \beta \sim N(T|\mathbf{0},\ \beta^{-1}\mathbf{I} + \alpha^{-1}XX^{\mathsf{T}})$

Поэтому: $\log p(T|X,\ \alpha,\ \beta) \propto$

$-\frac{1}{2}\log\det(\beta^{-1}\mathbf{I} + \alpha^{-1}XX^{\mathsf{T}}) - \frac{1}{2}T^{\mathsf{T}}(\beta^{-1}\mathbf{I} + \alpha^{-1}XX^{\mathsf{T}})^{-1}T.$

Пример

$y_i = \sin x_i + \varepsilon_i$, x_i равномерно выбрано на $[-\pi/2,\ \pi/2]$,

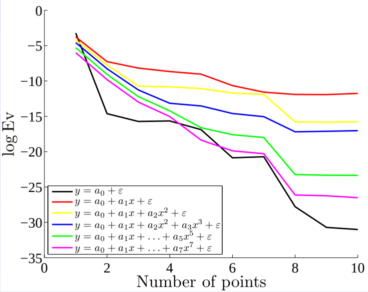
$\varepsilon_i \sim N(0,\ \beta^{-1})$

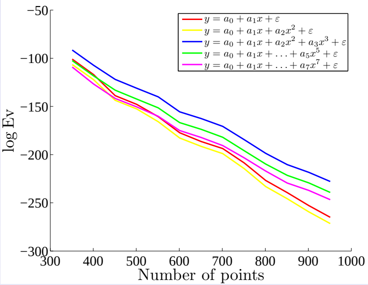
$\theta \sim N(\theta|\mathbf{0},\ \alpha^{-1}\mathbf{I})$

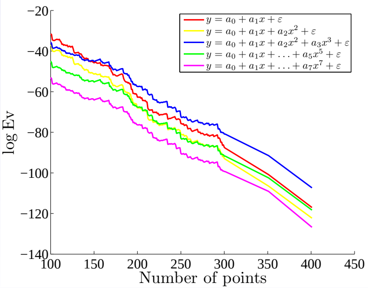
Значения параметров: $\alpha = 0.01,\ \beta = 10.$

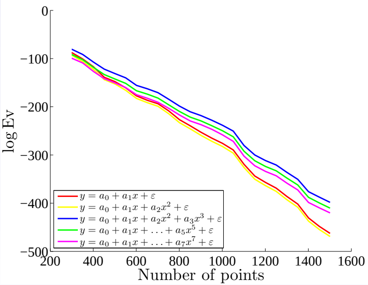
Признаки: $1,\ x_i,\ x_i^2,\ \ldots,\ x_i^k,\ \ldots.$

Пример: сравнение моделей









Пример оптимизации evidence

$$t_i = w + \varepsilon_i,\ \varepsilon_i \sim N(\varepsilon|0,\ \beta^{-1})$$

$$t_1,\ \ldots,\ t_n \sim N(t|\theta,\ \beta^{-1}),\ \theta \sim N(\theta|0,\ \alpha^{-1}).$$

Evidence: $p(t|\alpha,\ \beta)$

$$p(t|\alpha,\ \beta) = \frac{\beta^{n/2}\alpha^{1/2}}{(2\pi)^{n/2}\sqrt{n\beta+\alpha}} \exp\left(-\frac{1}{2}\beta\sum_{i=1}^{n} t_i^2 + \frac{\beta^2(\sum_{i=1}^{n} t_i)^2}{2(n\beta+\alpha)}\right)$$

$$(\alpha^*,\ \beta^*) = \arg\max_{\alpha,\,\beta} p(t|\alpha,\ \beta).$$

$$\alpha^* = \begin{cases} \dfrac{n^2\beta}{\beta(\sum_{i=1}^{n} t_i)^2 - n}, & \beta\left(\sum_{i=1}^{n} t_i\right)^2 > n, \\ +\infty, \text{ иначе.} \end{cases} \qquad \beta^* = \frac{n-1}{\sum_{i=1}^{n}(t_i - \bar{t})^2}.$$

Отбор признаков с помощью evidence для линейной регрессии

$$p(T, \theta | X, A, \beta) = p(T | X, \theta, \beta) p(\theta | A) = N(T | X\theta, \beta^{-1}\mathbf{I}) N(\theta | \mathbf{0}, A^{-1}).$$

Максимизация evidence: $p(T_{\text{train}} | X_{\text{train}},\ A,\ \beta) \to \max\limits_{A,\,\beta}$.

$$\log p(T_{\text{train}} | X_{\text{train}},\ A,\ \beta) \propto$$

$$-\tfrac{1}{2}\log\det(\beta^{-1}\mathbf{I} + XA^{-1}X^{\mathsf{T}}) - \tfrac{1}{2}T^{\mathsf{T}}(\beta^{-1}\mathbf{I} + XA^{-1}X^{\mathsf{T}})^{-1}T \to \max_{A,\,\beta}.$$

Введем $\mu = \beta\Sigma X^{\mathsf{T}} T$, где $\Sigma = (\beta X^{\mathsf{T}} X + A)^{-1}$.

$$\tfrac{1}{2}\log\det\Sigma + \tfrac{n}{2}\log\beta + \tfrac{1}{2}\log\det A - \tfrac{1}{2}\beta\|T - X\mu\|^2 - \tfrac{1}{2}\mu^{\mathsf{T}} A\mu \to \max_{A,\,\beta}.$$

Итеративный алгоритм максимизации evidence

$$\frac{1}{2}\log\det\Sigma+\frac{n}{2}\log\beta+\frac{1}{2}\log\det A-\frac{1}{2}\beta\|T-X\mu\|^2-\frac{1}{2}\mu^{\mathsf{T}}A\mu\to\max_{A,\,\beta}.$$

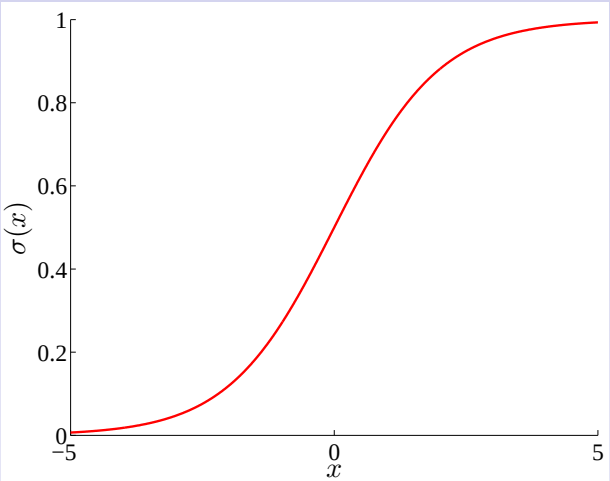
Рассмотрим $f(x,\,g(x))\to\max\limits_{x}$. Предположим

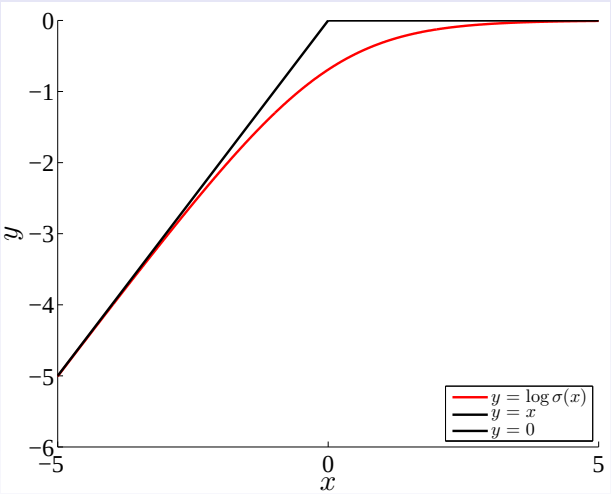
$\dfrac{\partial f}{\partial x}=0,\ \dfrac{\partial f}{\partial g}=0$ легко решаются, а $\dfrac{df}{dx}=0$ сложно.

Итерационный процесс:
$$\begin{cases} x_n=\arg\max\limits_{x} f(x,\,g_{n-1}),\\ g_n=\arg\max\limits_{g} f(x_n,\,g). \end{cases}$$

1. $\alpha_j^n=\dfrac{1-\alpha_j^{n-1}\Sigma_{jj}^{n-1}}{\mu_j^2}$
2. $\beta^n=\dfrac{n}{\|T-X\mu^{n-1}\|^2+tr(\Sigma^{\mathsf{T}}X^{\mathsf{T}}X)}$
3. $\mu^n=\beta^n(\beta^nX^{\mathsf{T}}X+A^n)^{-1}X^{\mathsf{T}}T.$

Evidence для логистической регрессии





$$p(t_i|x_i,\ \theta) = \frac{1}{1 + \exp(-t\theta^{\mathsf{T}} x)} = \sigma(t\theta^{\mathsf{T}} x).$$

$$p(T|X,\ \theta) = \prod_{i=1}^{n} p(t_i|x_i,\ \theta),$$

$$p(\theta|A) = N(\theta|\mathbf{0},\ A^{-1})$$

$$A^* = \arg\max_A p(T|X,\ A) =$$

$$\arg\max_A \int p(T,\ \theta|X,\ A)d\theta =$$

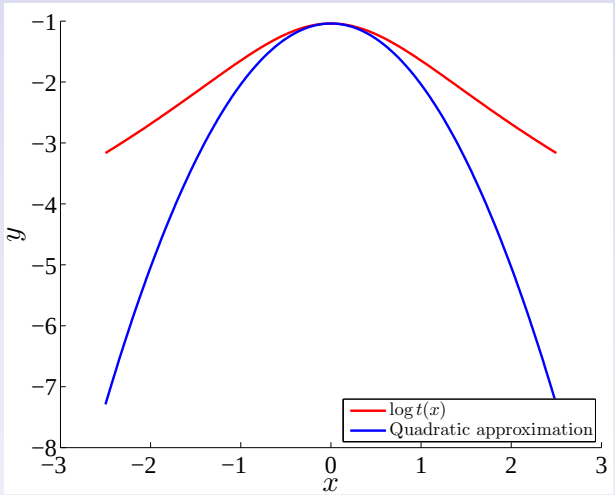
$$\arg\max_A \int \underbrace{p(T|X,\ \theta)p(\theta|A)}_{Q(\theta)}\, d\theta.$$

Аппроксимация

$$\log Q(\theta) \approx \log Q(\theta_{MP}) + \frac{1}{2}(\theta - \theta_{MP})^{\mathsf{T}} \nabla\nabla \log Q(\theta_{MP})(\theta - \theta_{MP})$$

$$\log p(T|X,\ A) \approx \log p(T|X,\ \theta_{MP}) - \frac{d}{2}\log(2\pi) + \frac{1}{2}\log\det(A) + \frac{1}{2}\log\det\Sigma - \frac{1}{2}\theta_{MP}^{\mathsf{T}} A \theta_{MP},\ \text{где}\ \Sigma^{-1} = X^{\mathsf{T}} R X + A.$$

1. $\alpha_j^n = \dfrac{1 - \alpha_j^{n-1}\Sigma_{jj}^{n-1}}{\theta_{MPj}^2}$
2. $\theta_{MP}^n \leftarrow \text{IRLS}$



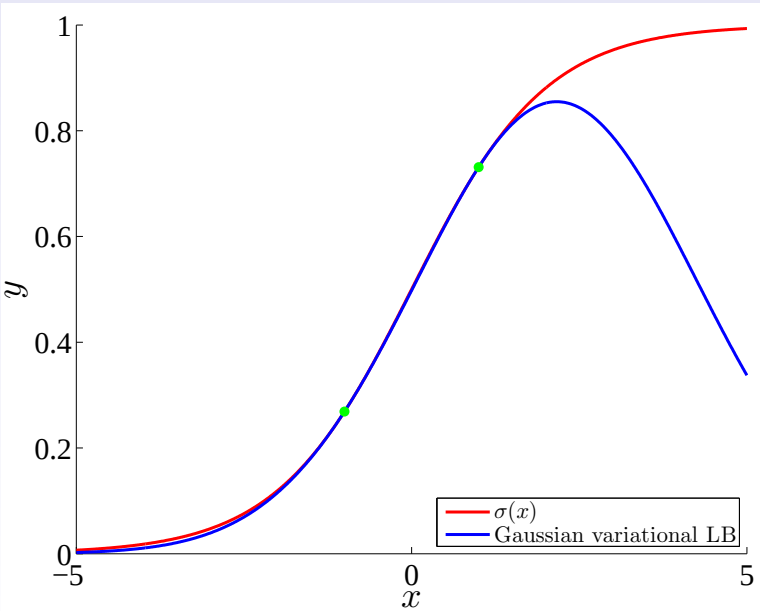
Вариационные нижние оценки

Определение. $g(x,\, \xi)$ вариационная нижняя оценка для $f(x) \Longleftrightarrow$

1. $f(x) \geq g(x,\, \xi)\, \forall\, x,\, \xi$
2. $f(\xi) = g(\xi,\, \xi)$.

Вместо $f(x) \to \max$ рассмотрим:

1. $\xi^n = \arg\max\limits_{\xi} g(x^n,\, \xi)$
2. $x^n = \arg\max\limits_{x} g(x,\, \xi^n)$



VLB для сигмоидной функции

$$\sigma(x) \geq \sigma(\eta) \exp\left(\frac{1}{4\eta}(1 - 2\sigma(\eta))(x^2 - \eta^2) + \frac{x - \eta}{2}\right)$$

$$p(T|X,\, A) \geq LB(A,\, \boldsymbol{\eta}) \to \max_{A,\, \boldsymbol{\eta}}$$

Список литературы

1. Mackay, David. The evidence framework applied to classification networks. Neural computation 4.5 (1992): 720-736.
2. Bishop, Christopher M. Pattern recognition and machine learning. Vol. 1. New York: Springer, 2006.
3. MacKay, David JC. Bayesian methods for adaptive models. Diss. California Institute of Technology, 1992.
4. Yuksel, Seniha Esen, Joseph N. Wilson, and Paul D. Gader. Twenty years of mixture of experts. Neural Networks and Learning Systems, IEEE Transactions on 23.8 (2012): 1177-1193.