Байесовский выбор моделей: обоснованность и отбор признаков в линейной и логистической регрессии

Александр Адуенко

10е 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 семейство распределений. Достаточные статистики.
- Наивный байесовский классификатор. Связь целевой функции и вероятностной модели.
- Линейная регрессия: классический подход, связь МНК и принципа максимума правдоподобия, связь регуляризации и MAP-оценки для вектора параметров $\mathbf{w}$.
- Свойство сопряженности априорного распределения правдоподобию.
- Прогноз для одиночной модели:
$$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}}|\mathbf{w}, \mathbf{X}_{\text{test}})p(\mathbf{w}|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})d\mathbf{w}.$$
- Связь апостериорной вероятности модели и обоснованности
$$p(M_i|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) \propto p(M_i)p_i(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}}).$$

Пример выбора модели

a – applicant, r – reviewer

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

d – decision

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

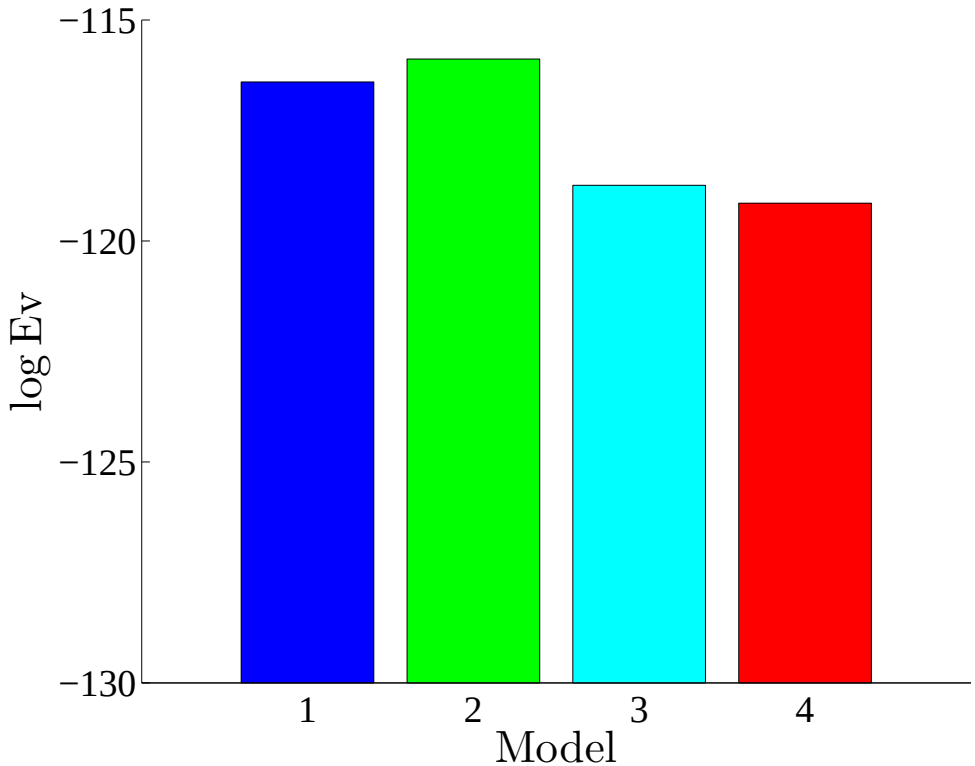
$r = 0$	$d = 0$	$d = 1$
$a = 0$	9	0
$a = 1$	132	19

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

Случаи:

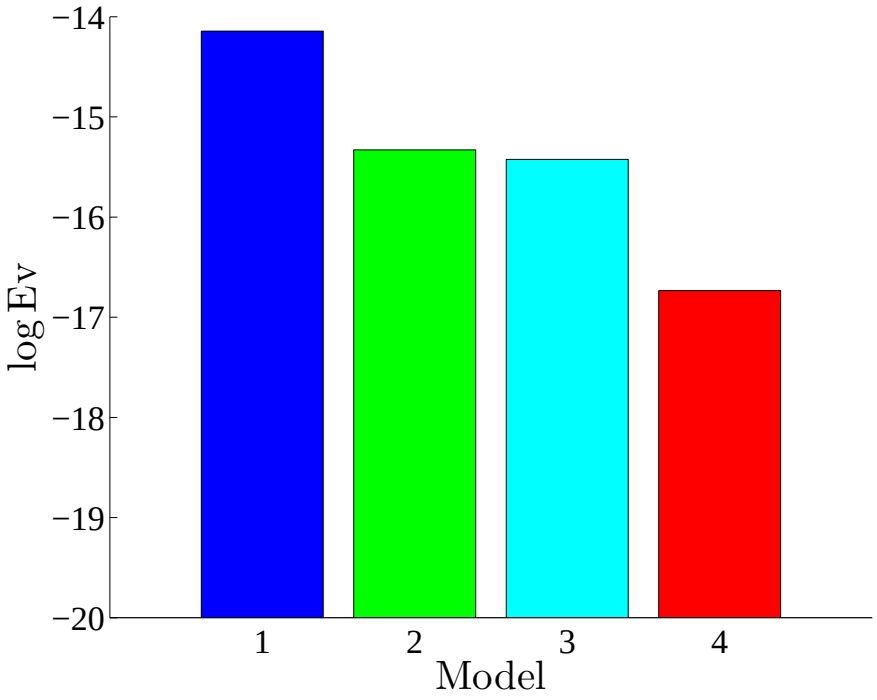
1. $p(d|a,\ r) = p(d)$
2. $p(d|a,\ r) = p(d|a)$
3. $p(d|a,\ r) = p(d|r)$
4. $p(d|a,\ r) = p(d|a,\ r)$

Пример выбора модели

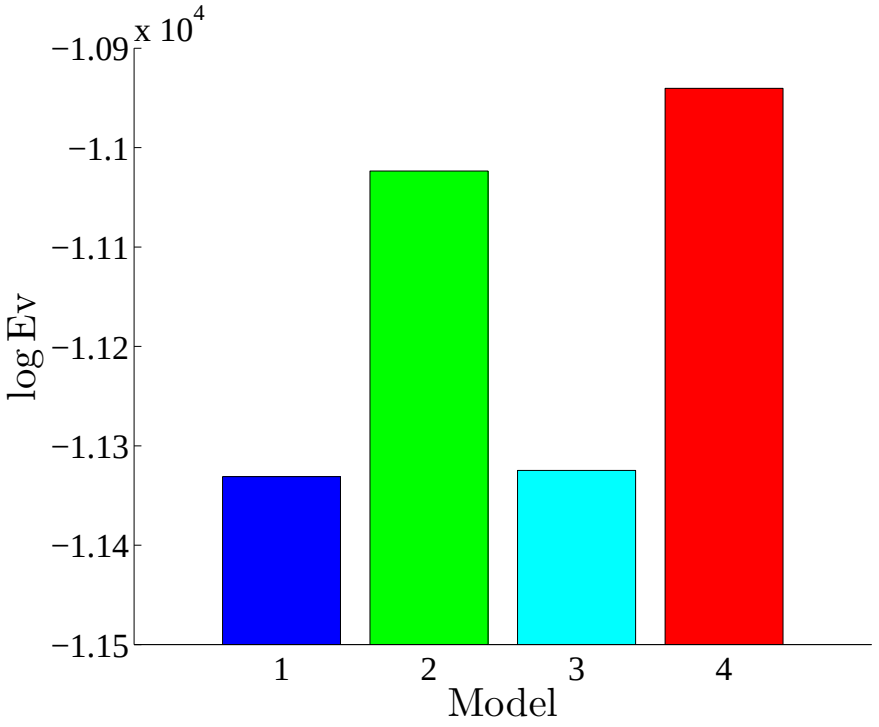


Сравнение обоснованностей, 326 объектов в выборке

Выбор модели: зависимость от размера выборки



Сравнение обоснованностей, 33 объекта в выборке



Сравнение обоснованностей, 32600 объектов в выборке

Простое понимание обоснованности

$$\text{Evidence} : p_i(\mathbf{y}|\mathbf{X}) = \int p_i(\mathbf{y}|\mathbf{X},\ \mathbf{w}) p_i(\mathbf{w}) d\mathbf{w}$$

$$p_i(\mathbf{w}|\mathbf{X},\ \mathbf{y}) = \frac{p_i(\mathbf{y}|\mathbf{X}, \mathbf{w}) p_i(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})}.$$

Предположения:

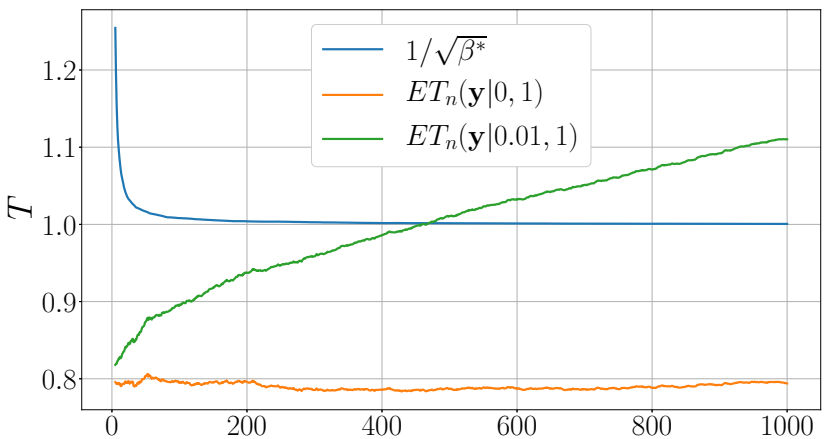
- $\mathbf{w}$ одномерный
- Априорное распределение $p_i(w)$ плоское с шириной Δw_{prior}
- Апостериорное распределение $p_i(w|\mathbf{X},\ \mathbf{y})$ сконцентрировано вокруг w_{MP} с шириной Δw_{post}

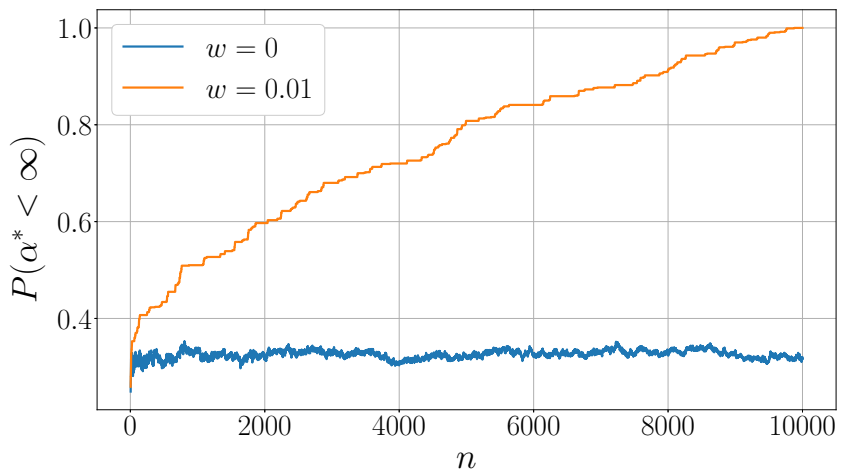
Тогда: $\log p_i(\mathbf{y}|\mathbf{X}) \approx \log p_i(\mathbf{y}|\mathbf{X},\ w_{MP}) + \log \left(\frac{\Delta w_{\text{post}}}{\Delta w_{\text{prior}}} \right).$

Для M-мерного $\mathbf{w}$: $\log p_i(\mathbf{y}|\mathbf{X}) \approx \log p_i(\mathbf{y}|\mathbf{X},\ \mathbf{w}_{MP}) + M \log \left(\frac{\Delta w_{\text{post}}}{\Delta w_{\text{prior}}} \right).$

Пример оптимизации evidence

$$y_i = w + \varepsilon_i,\ \varepsilon_i \sim N(\varepsilon|0,\ \beta^{-1})$$
$$y_1|w,\ \ldots,\ y_n|w \sim N(y_i|w,\ \beta^{-1}),\ w \sim N(w|0,\ \alpha^{-1}).$$
$$p(\mathbf{y}|\alpha,\ \beta) = \frac{\beta^{n/2}\alpha^{1/2}}{(2\pi)^{n/2}\sqrt{n\beta+\alpha}} \exp\left(-\frac{1}{2}\beta\sum_{i=1}^{n} y_i^2 + \frac{\beta^2(\sum_{i=1}^{n} y_i)^2}{2(n\beta+\alpha)}\right).$$
$$(\alpha^*,\ \beta^*) = \arg\max_{\alpha,\ \beta} p(\mathbf{y}|\alpha,\ \beta).$$





$$\alpha^* = \begin{cases} \dfrac{n^2\beta^*}{\beta^*(\sum_{i=1}^{n} y_i)^2 - n}, & \underbrace{\dfrac{|\sum_{i=1}^{n} y_i|}{\sqrt{n}}}_{T_n(\mathbf{y}|w,\beta)} > \dfrac{1}{\sqrt{\beta^*}}, \\ +\infty, \text{ иначе.} \end{cases} \qquad \frac{1}{\beta^*} = \frac{\sum_{i=1}^{n}(y_i - \overline{\mathbf{y}})^2}{n-1}.$$

Обоснованность для линейной регрессии

$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, $\mathbf{w} \sim N(\theta|\mathbf{0},\ \mathbf{A}^{-1})$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0},\ \sigma^2\mathbf{I})$

Совместное правдоподобие: $p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \sigma^2) = p(\mathbf{y}|\mathbf{X},\ \mathbf{w},\ \sigma^2)p(\mathbf{w}|\mathbf{A})$.

Обоснованность:

$$p(\mathbf{y}|\mathbf{X},\ \mathbf{A},\ \beta) = \int p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \sigma^2)d\mathbf{w} = \int p(\mathbf{y}|\mathbf{X},\ \mathbf{w},\ \sigma^2)p(\mathbf{w}|\mathbf{A})d\mathbf{w}.$$

$$\mathbf{y}|\mathbf{X},\ \mathbf{A},\ \sigma^2 \sim N(\mathbf{y}|\mathbf{0},\ \sigma^2\mathbf{I} + \mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\mathsf{T})$$

Поэтому:

$$\log p(\mathbf{y}|\mathbf{X}, \mathbf{A}, \sigma^2) \propto -\tfrac{1}{2}\log\det(\sigma^2\mathbf{I}+\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\mathsf{T}) - \tfrac{1}{2}\mathbf{y}^\mathsf{T}(\sigma^2\mathbf{I}+\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\mathsf{T})^{-1}\mathbf{y}.$$

Пример

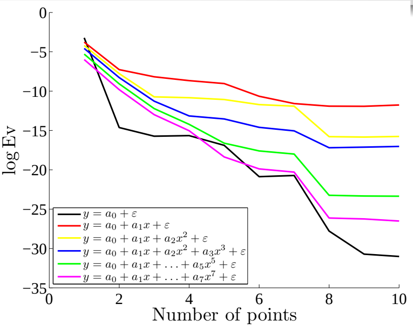
$y_i = \sin x_i + \varepsilon_i$, x_i равномерно выбрано на $[-\pi/2,\ \pi/2]$, $\varepsilon_i \sim N(0,\ \sigma^2)$

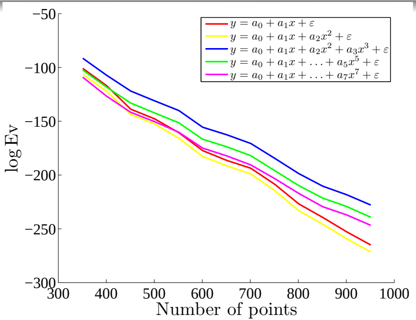
$\mathbf{w} \sim N(\mathbf{w}|\mathbf{0},\ \alpha^{-1}\mathbf{I})$

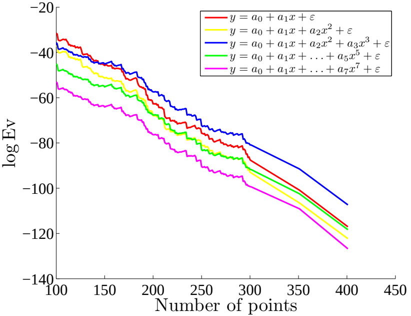
Значения параметров: $\alpha = 0.01$, $\sigma^2 = 0.1$.

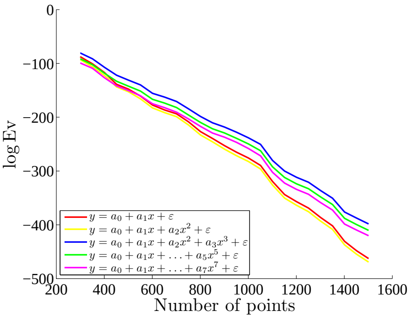
Признаки: $1,\ x_i,\ x_i^2,\ \ldots,\ x_i^k,\ \ldots$.

Пример: сравнение моделей





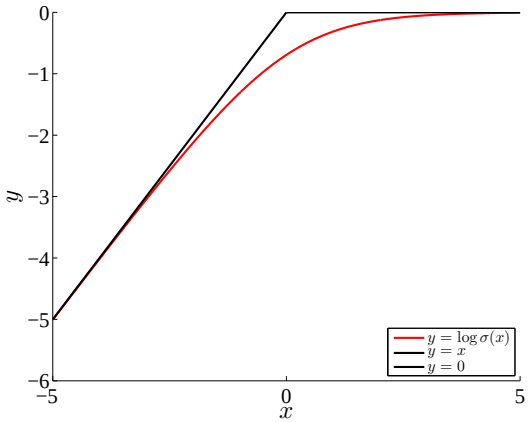




Байесовская логистическая регрессия

Пусть $\mathbf{X} \in \mathbb{R}^{m \times n}$ – признаковая матрица, а $\mathbf{y} \in \{\pm 1\}^m$ – метки класса.

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



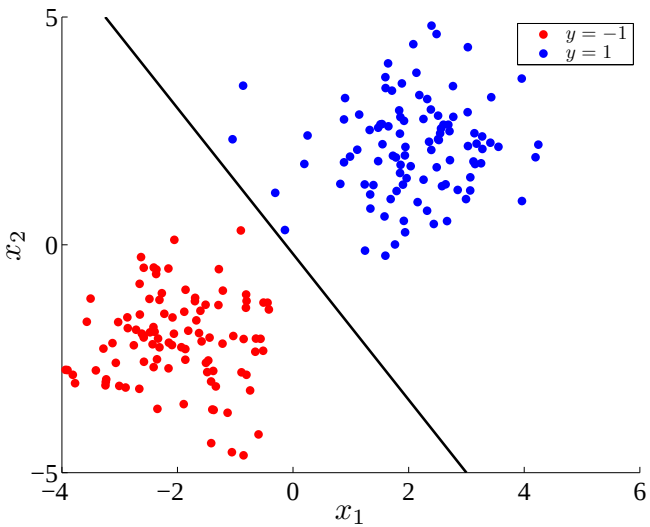
$$p(y_j|\mathbf{w}, \mathbf{x}_j) = \sigma(y_j \mathbf{w}^\mathsf{T} \mathbf{x}_j) = \frac{1}{1+\exp(-y_j \mathbf{w}^\mathsf{T} \mathbf{x}_j)}.$$

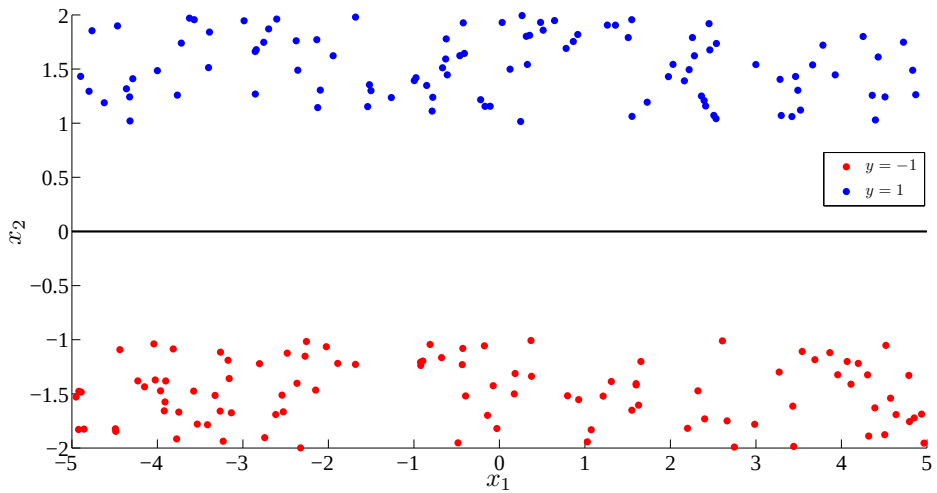
Вопрос 1: как выбрать $p(\mathbf{w}|\mathbf{A})$?

Вопрос 2: Пусть $p(\mathbf{w}|\mathbf{A}) = N(\mathbf{0}, \mathbf{A}^{-1})$, $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$

Что происходит, когда $\alpha_i \to \infty$?

Примеры неоднозначности выбора разделительной прямой.





Вопрос 3: чему равна $\mathbf{w}_{ML}$ для выборок на рис. выше?

Обоснованность для логистической регрессии

Пусть $\mathbf{X} \in \mathbb{R}^{m \times n}$ – признаковая матрица, а $\mathbf{y} \in \{\pm 1\}^m$ – метки класса.

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

Идея: выбрать модель с максимальной обоснованностью.

Вопрос 1: чем отличаются разные модели байесовской логистической регрессии, описанные выше?

Вычисление обоснованности.

Пусть далее $p(\mathbf{w}|\mathbf{A}) = N(\mathbf{0}, \mathbf{A}^{-1})$, $\mathbf{A} = \operatorname{diag}(\boldsymbol{\alpha})$.

Тогда $\mathbf{A}^* = \arg\max_{\mathbf{A}} p(\mathbf{y}|\mathbf{X}, \mathbf{A}) = \arg\max_{\mathbf{A}} \int \underbrace{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\mathbf{A})}_{Q(\mathbf{w})} d\mathbf{w}$.

Проблема: интеграл аналитически не вычисляется.

Аппроксимация Лапласа

$$\log Q(\mathbf{w}) \approx \log Q(\mathbf{w}_{\text{MAP}}) + \frac{1}{2}(\mathbf{w} - \mathbf{w}_{\text{MAP}})^\mathsf{T} \overbrace{\nabla\nabla \log Q(\mathbf{w}_{\text{MAP}})}^{-\mathbf{H}^{-1}} (\mathbf{w} - \mathbf{w}_{\text{MAP}}).$$

$$\mathbf{A}^* = \arg\max_{\mathbf{A}} \left(Q(\mathbf{w}_{\text{MAP}}) \int e^{-\frac{1}{2}(\mathbf{w}-\mathbf{w}_{\text{MAP}})^\mathsf{T} \mathbf{H}^{-1} (\mathbf{w}-\mathbf{w}_{\text{MAP}})} d\mathbf{w} \right).$$

Вопрос 2: Как определяется $\mathbf{w}_{\text{MAP}}$?

Вариационные нижние оценки

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

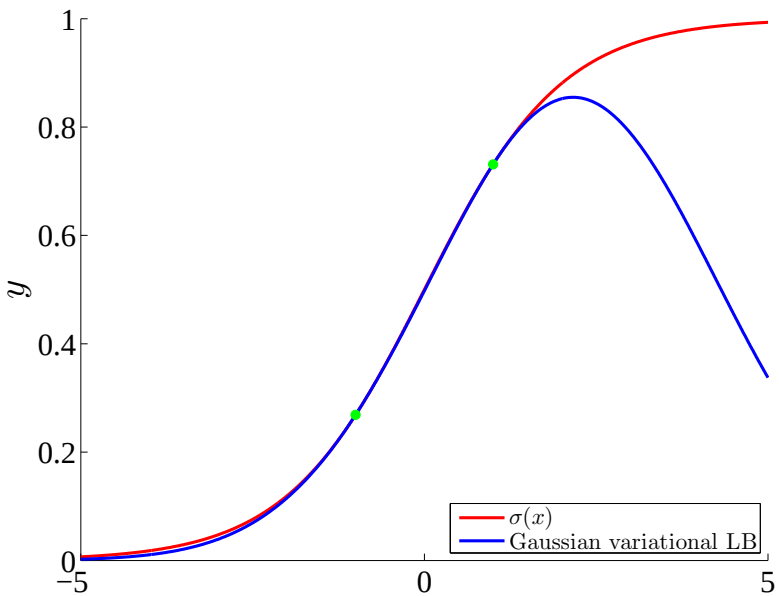
1. $f(x) \geq g(x,\,\xi)\;\forall\, x,\,\xi$
2. $f(\xi) = g(\xi,\,\xi)$.

Вместо $f(x) \to \max\limits_{x}$ рассмотрим $g(x,\,\xi) \to \max\limits_{x,\,\xi}$

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

VLB для сигмоидной функции

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



Вопрос: в чем преимущество использования VLB при максимизаци обоснованности в логистической регрессии?

LB для обоснованности в логистической регрессии

$$p(\mathbf{y},\,\mathbf{w}|\mathbf{X},\,\mathbf{A}) = \prod_{j=1}^{m} \sigma(y_j \mathbf{w}^\mathsf{T}\mathbf{x}_j)\frac{\sqrt{\det \mathbf{A}}}{(2\pi)^{n/2}} e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}} \geq \mathrm{VLB}(\mathbf{w},\,\boldsymbol{\xi},\,\mathbf{A}) =$$

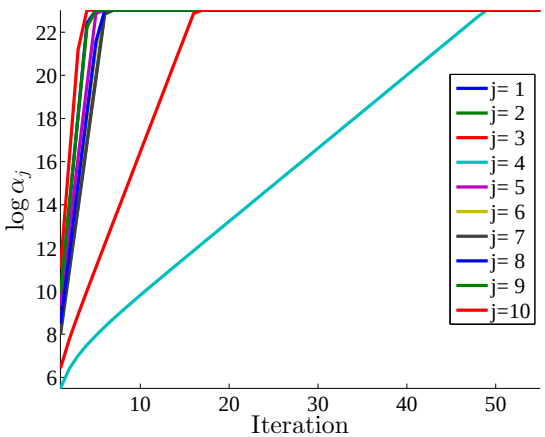
$$\frac{\sqrt{\det \mathbf{A}}}{(2\pi)^{n/2}} e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}} \prod_{j=1}^{m} \sigma(\xi_j) \exp\left(-\frac{2\sigma(\xi_j)-1}{4\xi_j}(\mathbf{w}^\mathsf{T}\mathbf{x}_j\mathbf{x}_j^\mathsf{T}\mathbf{w} - \xi_j^2) + \frac{y_j \mathbf{w}^\mathsf{T}\mathbf{x}_j - \xi_j}{2} \right) =$$

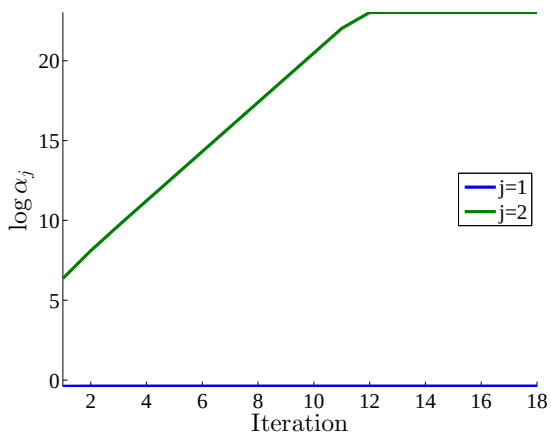
$$\frac{\sqrt{\det \mathbf{A}}}{(2\pi)^{n/2}} \prod_{j=1}^{m} \sigma(\xi_j) e^{\frac{2\sigma(\xi_j)-1}{4\xi_j}\xi_j^2 - \frac{\xi_j}{2}} e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}'\mathbf{w} + \mathbf{w}^\mathsf{T}\mathbf{v}},\ \text{где}$$

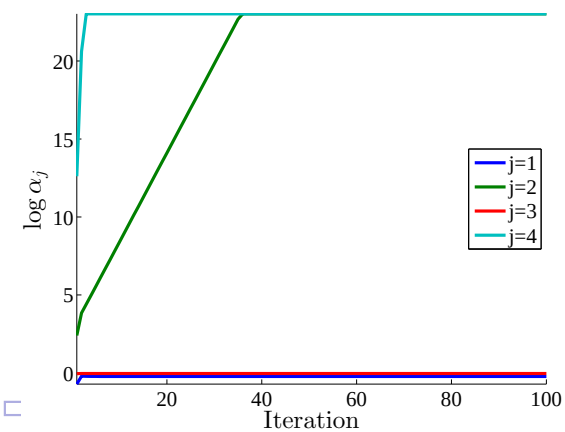
$$\mathbf{A}' = \mathbf{A} + \sum_{j=1}^{m} \frac{2\sigma(\xi_j)-1}{2\xi_j}\mathbf{x}_j\mathbf{x}_j^\mathsf{T},\ \mathbf{v} = \frac{1}{2}\sum_{j=1}^{m} y_j \mathbf{x}_j.$$

Тогда $p(\mathbf{y}|\mathbf{X},\,\mathbf{A}) \geq \mathrm{LB}(\mathbf{A},\,\boldsymbol{\xi}) = \int \mathrm{VLB}(\mathbf{w},\,\boldsymbol{\xi},\,\mathbf{A})d\mathbf{w} \to \max_{\mathbf{A},\,\boldsymbol{\xi}}$

Иллюстрация отбора признаков в логистической регрессии







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. Agresti, Alan. Analysis of ordinal categorical data. Vol. 656. John Wiley & Sons, 2010.
6. Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.
7. Chen, Ming-Hui, and Joseph G. Ibrahim. "Conjugate priors for generalized linear models."Statistica Sinica (2003): 461-476.