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

Александр Адуенко

10е октября 2023

- Формула Байеса: $\mathsf{P}(A|B) = \dfrac{\mathsf{P}(B|A)\mathsf{P}(A)}{\mathsf{P}(B)}$;
- Формула полной вероятности: $\mathsf{P}(B) = \mathsf{P}(B|A)\mathsf{P}(A) + \mathsf{P}(B|\overline{A})\mathsf{P}(\overline{A})$;
- Определение априорных вероятностей и selection bias;
- (Множественное) тестирование гипотез
- Экспоненциальное семейство. Достаточные статистики.
- Naive Bayes. Связь целевой функции и вероятностной модели.
- Линейная регрессия: классический подход, связь МНК и ML-оценки, 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 $p(M_i|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) \propto p(M_i)p_i(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}})$.
- Прогноз для многих моделей: $p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) = \sum_{k=1}^{K} p(M_k|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})p_k(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}})$.

Обоснованность (evidence)

Модель M_i: $p_i(\mathbf{y},\ \mathbf{w}|\mathbf{X}) = p_i(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w})$

Шаг	Наблюдаемые	Скрытые	Результат
Обучение	$(\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})$	$\mathbf{w}$	$p(\mathbf{w}\|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})$
Контроль	$\mathbf{X}_{\text{test}}$	$\mathbf{y}_{\text{test}}$	$p(\mathbf{y}_{\text{test}}\|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})$

$$p(\mathbf{w}|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) = \frac{p(\mathbf{y}_{\text{train}},\ \mathbf{w}|\mathbf{X}_{\text{train}})}{\int p(\mathbf{y}_{\text{train}},\ \mathbf{w}^*|\mathbf{X}_{\text{train}})d\mathbf{w}^*}$$

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

Обоснованность (evidence)

Модель M_i: $p_i(\mathbf{y},\ \mathbf{w}|\mathbf{X}) = p_i(\mathbf{y}|\mathbf{X},\ \mathbf{w})p_i(\mathbf{w})$
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 $\mathbf{w}^*$ из априорного распределения $p_{i^*}(\mathbf{w})$
- Имея i^*, $\mathbf{w}^*$ природа выбирает $\mathbf{X}_{\text{train}}$ и сэмплирует $\mathbf{y}_{\text{train}}$ из $p_{i^*}(\mathbf{y}|\mathbf{X}_{\text{train}},\ \mathbf{w}^*)$
- $(\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})$ даны наблюдателю.
- Природа выбирает $\mathbf{X}_{\text{test}}$ и сэмплирует $\mathbf{y}_{\text{test}}$ из $p_{i^*}(\mathbf{y}|\mathbf{X}_{\text{test}},\ \mathbf{w}^*)$

Модель M_i: $p_i(\mathbf{y},\ \mathbf{w}|\mathbf{X}) = p_i(\mathbf{y}|\mathbf{X},\ \mathbf{w})p_i(\mathbf{w})$
Общая модель M: $p(\mathbf{y},\ \mathbf{w},\ M_i|\mathbf{X}) = p(M_i)p_i(\mathbf{w})p_i(\mathbf{y}|\mathbf{X},\ \mathbf{w})$

$$p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) = \\ \sum_{i=1}^{K} p(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}},\ M_i)p(M_i|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) = \\ \sum_{i=1}^{K} p_i(\mathbf{y}_{\text{test}}|\mathbf{X}_{\text{test}},\ \mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})p(M_i|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}})$$

$$p(M_i|\mathbf{X}_{\text{train}},\ \mathbf{y}_{\text{train}}) = \frac{p(\mathbf{y}_{\text{train}},\ M_i|\mathbf{X}_{\text{train}})}{P(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}})} \propto p(\mathbf{y}_{\text{train}}, M_i|\mathbf{X}_{\text{train}}) = \\ \int p(\mathbf{y}_{\text{train}},\ \mathbf{w},\ M_i|\mathbf{X}_{\text{train}})d\mathbf{w} = p(M_i)p_i(\mathbf{y}_{\text{train}}|\mathbf{X}_{\text{train}})$$

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

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

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

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X},\ w)p(w)dw = \int_0^1 C_9^0(1-w)^9$$

$C_{103}^{97}w^6(1-w)^{97}C_{151}^{132}w^{19}(1-w)^{132}C_{63}^{52}w^{11}(1-w)^{52}dw = 2.8\cdot 10^{-51}CCCC$

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

Поэтому $p(d|a=0) = \mathrm{Be}(w_1)$, $p(d|a=1) = \mathrm{Be}(w_2)$.

$\textbf{Prior}:\ p(w_1) = U[0,\ 1]$, $p(w_2) = U[0,\ 1]$

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, w_1, w_2)p(w_1)p(w_2)dw_1dw_2 = \int_0^1\int_0^1 C_9^0(1-w_1)^9C_{103}^{97}$$

$w_1^6(1-w_1)^{97}C_{151}^{132}w_2^{19}(1-w_2)^{132}C_{63}^{52}w_2^{11}(1-w_2)^{52}dw_1dw_2 = 4.7\cdot 10^{-51}CCCC$

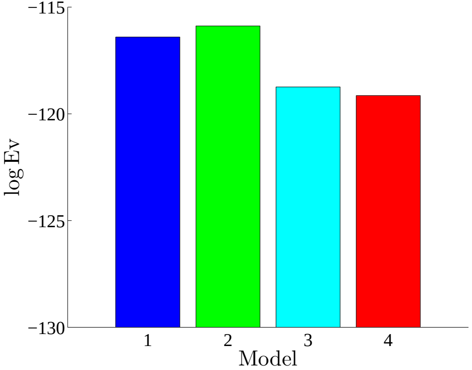
3)$p(d|a,\, r) = p(d|r)$
Поэтому $p(d|r=0) = \mathrm{Be}(w_1)$, $p(d|r=1) = \mathrm{Be}(w_2)$.
$\mathbf{Prior}:\ p(w_1) = U[0,\,1]$, $p(w_2) = U[0,\,1]$

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

4) $p(d|a,\, r) = p(d|a,\, r)$
Поэтому $p(d|a=0,\, r=0) = \mathrm{Be}(w_1)$, $p(d|a=0,\, r=1) = \mathrm{Be}(w_2)$,
$p(d|a=1,\, r=0) = \mathrm{Be}(w_3)$, $p(d|a=1,\, r=1) = \mathrm{Be}(w_4)$.
$\mathbf{Prior}:\ p(w_1) = U[0,\,1]$, $p(w_2) = U[0,\,1]$,
$p(w_3) = U[0,\,1]$, $p(w_4) = U[0,\,1]$

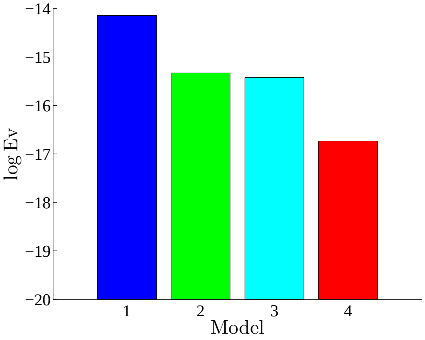
$$p(\mathbf{y}|\mathbf{X}) = 0.18 \cdot 10^{-51} CCCC$$

Пример выбора модели

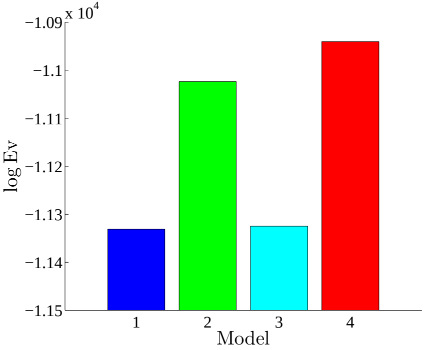


Сравнение обоснованностей, 326 объектов в выборке

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

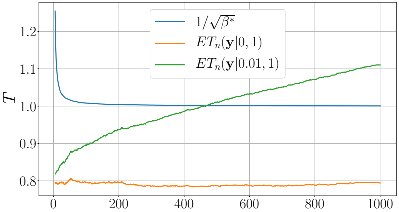
Для M-мерного $\mathbf{w}$: $\log p_i(\mathbf{y}|\mathbf{X}) \approx \log p_i(\mathbf{y}|\mathbf{X},\ \mathbf{w}_{MP}) + M \log \left(\frac{\Delta w_{\text{post}}}{\Delta w_{\text{prior}}}\right).$

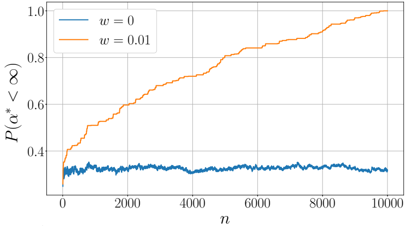
$$y_i = w + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(\varepsilon|0,\ \beta^{-1})$$

$$y_1|w,\ \dots,\ y_n|w \sim \mathcal{N}(y_i|w,\ \beta^{-1}),\ w \sim \mathcal{N}(w|0,\ \alpha^{-1}).$$

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

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





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

Обоснованность для линейной регрессии

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

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:

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

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

Поэтому:

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

Пример

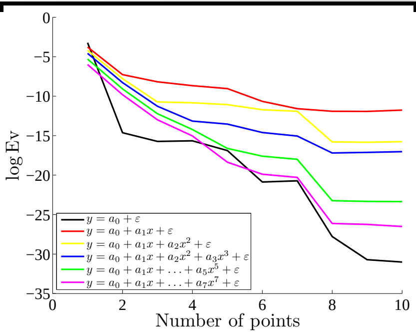
$y_i = \sin x_i + \varepsilon_i$, x_i равномерно выбрано на $[-\pi/2,\ \pi/2]$, $\varepsilon_i \sim \mathcal{N}(0,\ \sigma^2)$

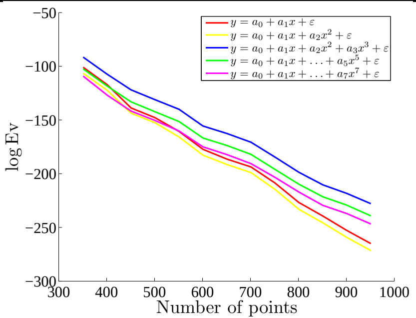
$\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{0},\ \alpha^{-1}\mathbf{I})$

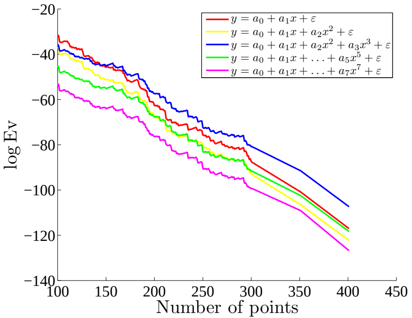
Значения параметров: $\alpha = 0.01,\ \sigma^2 = 0.1$.

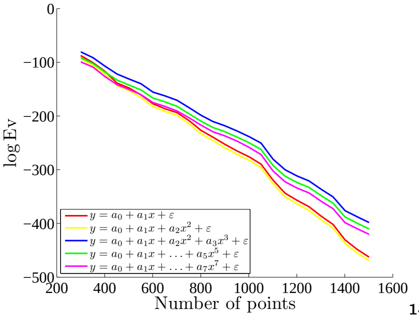
Признаки: $1,\ x_i,\ x_i^2,\ \dots,\ x_i^k,\ \dots$.

Пример: сравнение моделей





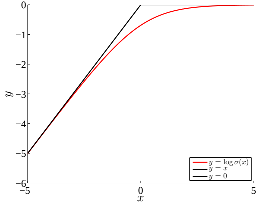




Байесовская логистическая регрессия

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

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



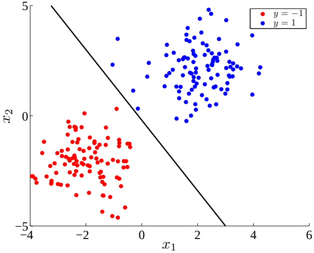
$$p(y_j|\mathbf{w},\ \mathbf{x}_j) = \sigma(y_j \mathbf{w}^{\mathsf{T}} \mathbf{x}_j) = \frac{1}{1+\exp(-y_j \mathbf{w}^{\mathsf{T}} \mathbf{x}_j)}.$$

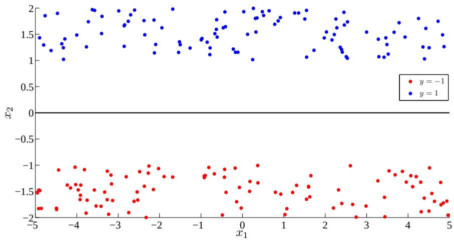
Вопрос 1: как выбрать $p(\mathbf{w}|\mathbf{A})$?

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

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

Примеры неоднозначности выбора разделительной прямой.





Вопрос 3: чему равна $\mathbf{w}_{\mathrm{ML}}$ 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 $p(\mathbf{w}|\mathbf{A}) = \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$, $\mathbf{A} = \operatorname{diag}(\boldsymbol{\alpha})$.

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

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

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

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

Вариационные нижние оценки

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

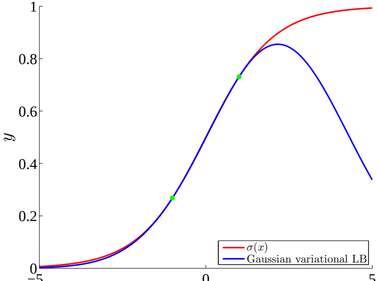
1. $f(x) \geq g(x,\, \xi)\, \forall\, x,\, \xi$
2. $f(\xi) = g(\xi,\, \xi)$.

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

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

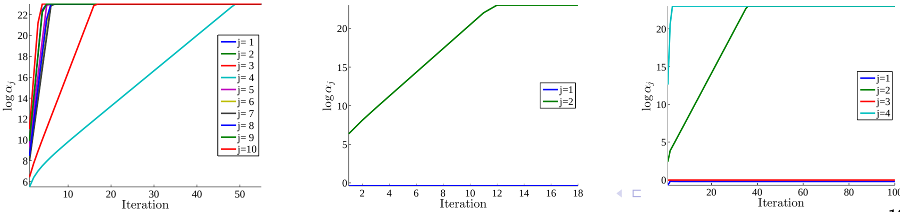
$$\frac{\sqrt{\det \mathbf{A}}}{(2\pi)^{n/2}} e^{-\frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w}} \prod_{j=1}^{m} \sigma(\xi_j) \exp\left(-\frac{2\sigma(\xi_j)-1}{4\xi_j}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_j\mathbf{x}_j^{\mathsf{T}}\mathbf{w} - \xi_j^2) + \frac{y_j \mathbf{w}^{\mathsf{T}}\mathbf{x}_j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

$$\frac{\sqrt{\det \mathbf{A}}}{(2\pi)^{n/2}} \prod_{j=1}^{m} \sigma(\xi_j) e^{\frac{2\sigma(\xi_j)-1}{4\xi_j}\xi_j^2 - \frac{\xi_j}{2}} e^{-\frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}'\mathbf{w} + \mathbf{w}^{\mathsf{T}}\mathbf{v}},$$ 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 $p(\mathbf{y}|\mathbf{X},\ \mathbf{A}) \geq \mathrm{LB}(\mathbf{A},\ \boldsymbol{\xi}) = \int \mathrm{VLB}(\mathbf{w},\ \boldsymbol{\xi},\ \mathbf{A})d\mathbf{w} \to \max\limits_{\mathbf{A},\ \boldsymbol{\xi}}.$

Иллюстрация отбора признаков в логистической регрессии



Апостериорное распределение в логистической регрессии

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

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

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

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 $\mathbf{R} = \operatorname{diag}(\sigma(\mathbf{w}_\text{MAP}^\mathsf{T}\mathbf{x}_j)\sigma(-\mathbf{w}_\text{MAP}^\mathsf{T}\mathbf{x}_j))$.
Нормальная аппроксимация: $p(\mathbf{w}|\mathbf{X},\ \mathbf{y},\ \mathbf{A}) \approx \mathcal{N}(\mathbf{w}|\mathbf{w}_\text{MAP},\ \mathbf{H}^{-1})$.

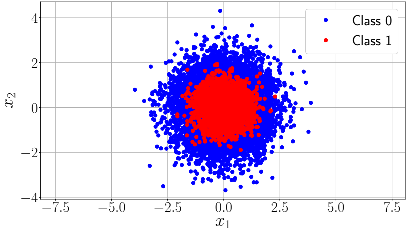
Пример. Пусть $n = 1$, $\mathbf{w}_\text{MAP} = 1$.
Вопрос 2: Что можно сказать про принадлежность объектов с $x = 0;\ 1;\ -1;\ 5;\ -5$ к классу 1?
Вопрос 3: Как результат зависит от неопределенности h^{-1}? Что происходит при $h \to 0$ и при $h \to \infty$?

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

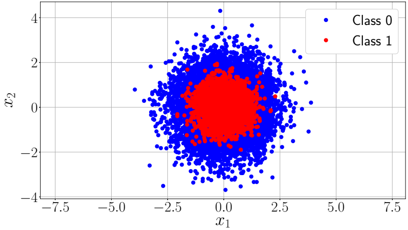
Прогноз вероятности класса 1 в зависимости от неопределенности h^{-1}

	$x = 5$	$x = 1$	$x = 0$	$x = -1$	$x = -5$
$h = \infty$	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?



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

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

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. Chen, Ming-Hui, and Joseph G. Ibrahim. "Conjugate priors for generalized linear models."Statistica Sinica (2003): 461-476.
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.