Байесовский выбор моделей: обоснованность и отбор признаков в логистической регрессии

Александр Адуенко

17е октября 2018

Содержание предыдущих лекций

- 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 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.

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

Идея: выбрать модель с максимальной обоснованностью.

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

Вычисление обоснованности.

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

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

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 $\mathbf{w}_{\mathsf{MAP}}$?

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

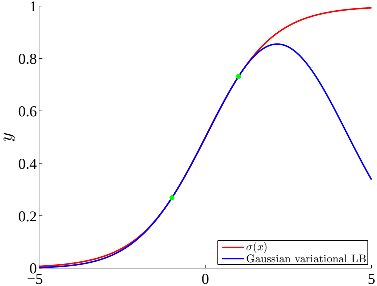
1. $f(x) \geq g(x,\, \xi)\; \forall\, x,\, \xi$
2. $f(\xi) = g(\xi,\, \xi)$.

Вместо $f(x) \to \max\limits_x$ 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?

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

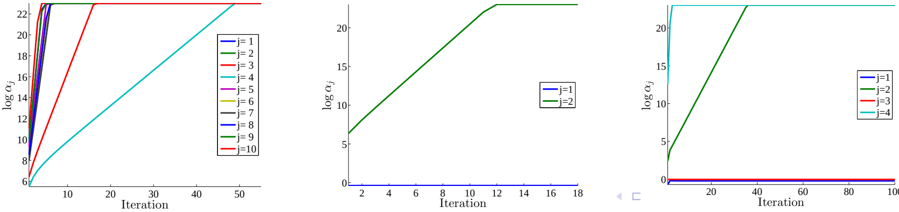
$$\frac{\sqrt{\det \mathbf{A}}}{(2\pi)^{n/2}} e^{-\frac{1}{2}\mathbf{w}^\mathsf{T}\mathbf{A}\mathbf{w}} \prod_{j=1}^{m} \sigma(\xi_j) \exp\left(-\frac{2\sigma(\xi_j)-1}{4\xi_j}(\mathbf{w}^\mathsf{T}\mathbf{x}_j\mathbf{x}_j^\mathsf{T}\mathbf{w} - \xi_j^2) + \frac{y_j\mathbf{w}^\mathsf{T}\mathbf{x}_j -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

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

$p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = p(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A})$, 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 $h \to \infty$?

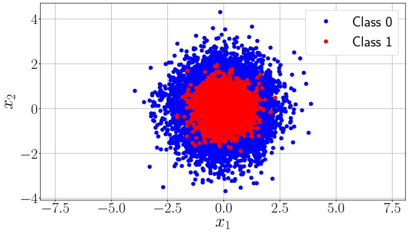
Нелинейная разделяющая поверхность

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

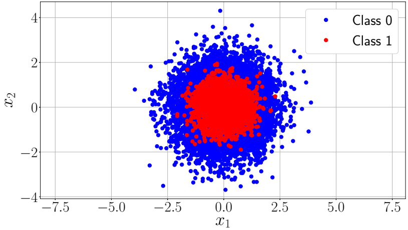
Прогноз вероятности класса 1 в зависимости от неопределенности h^{-1}

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