Байесовский выбор моделей: построение адекватных мультимоделей

Александр Адуенко

17е ноября 2021

Содержание предыдущих лекций

- Формула Байеса и формула полной вероятности;
- 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
- Обоснованность: понимание и связь со статистической значимостью.
- Логистическая регрессия: проблемы ML-оценки $\mathbf{w}$ и связь априорного распределения с отбором признаков.
- ЕМ-алгоритм и отбор признаков в байесовской линейной регрессии.
- Вариационный ЕМ-алгоритм. Смесь моделей лог. регрессии.
- Гауссовские процессы. Учёт эволюции моделей во времени.

Вероятностная модель генерации данных

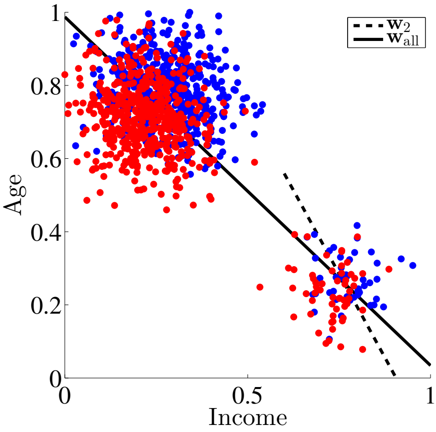
- Веса моделей в смеси $\boldsymbol{\pi}$ получены из априорного распределения $p(\boldsymbol{\pi}|\mu)$;
- Векторы параметров моделей $\mathbf{w}_k$ получены из нормального распределения $p(\mathbf{w}_k|\mathbf{A}_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})$, $k = 1, \ldots, K$;
- Для каждого объекта $\mathbf{x}_i$ выбрана модель f_{k_i}, которой он описывается, причем $p(k_i = k) = \pi_k$;
- Для каждого объекта $\mathbf{x}_i$ класс y_i определен в соответствии с моделью f_{k_i}: $y_i \sim \mathrm{Be}\big(\sigma(\mathbf{w}_{k_i}^{\mathsf{T}}\mathbf{x}_i)\big)$.

Совместное правдоподобие модели

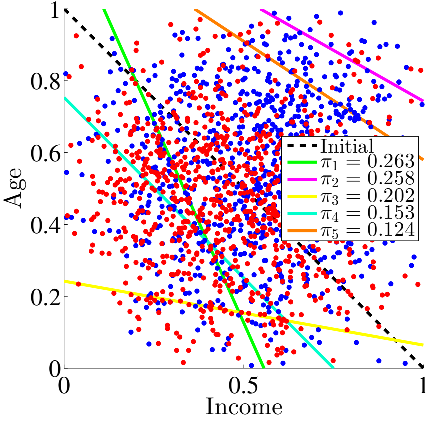
$$p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\mu}) =$$
$$p(\boldsymbol{\pi}|\boldsymbol{\mu})\prod_{k=1}^{K}\mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})\prod_{i=1}^{m}\left(\sum_{l=1}^{K}\pi_l\sigma(y_i\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i)\right).$$

Вопрос: Пусть известна функция $f(\mathbf{x})$, которая по объекту выдает номер модели, которой он описывается. Как изменится совместное правдоподобие?

Проблема: большое число близких или совпадающих моделей ведет к неинтерпретируемости и низкому качеству прогноза.



Неадекватная многоуровневая модель



Неадекватная смесь моделей

Вопрос: почему появление лишних моделей ухудшает качество прогноза?

Постановка задачи сравнения моделей

Определение. Мультимодель с совместным распределением $p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ (\boldsymbol{\pi})|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ (\mu))$ называется $(s,\ \alpha)$-**адекватной**, если модели, ее составляющие, попарно статистически различимы с помощью функции сходства s на уровне значимости α.

Проблема

Несмотря на прореживание мультимодели, она может не являться (s, α) – адекватной, то есть может содержать похожие модели.

Дано

- Две модели f_1 и f_2, векторы параметров моделей $\mathbf{w}_1,\ \mathbf{w}_2$.
- Выборки $(\mathbf{X}_1,\ \mathbf{y}_1)$ и $(\mathbf{X}_2,\ \mathbf{y}_2)$,
  $y_{1,i} = f_1(\mathbf{x}_{1,i},\ \mathbf{w}_1), \quad y_{2,i} = f_2(\mathbf{x}_{2,i},\ \mathbf{w}_2)$.
- Априорные распределения $\mathbf{w}_1 \sim p_1(\mathbf{w}),\ \mathbf{w}_2 \sim p_2(\mathbf{w})$.
- Апостериорные распределения параметров моделей
  $g_1(\mathbf{w}_1) = p(\mathbf{w}_1|\mathbf{X}_1,\ \mathbf{y}_1)$ и $g_2(\mathbf{w}_2) = p(\mathbf{w}_2|\mathbf{X}_2,\ \mathbf{y}_2)$.

Требуется: построить функцию сходства, определенную на паре распределений $g_1(\mathbf{w})$ и $g_2(\mathbf{w})$, удовлетворяющую ряду требований.

Требования к функции сходства s

Корректная функция сходства s должна быть

1. определена в случае несовпадения носителей,
2. $s(g_1,\ g_2) \leq s(g_1,\ g_1)$,
3. $s \in [0,\ 1]$,
4. $s(g_1,\ g_1) = 1$,
5. близка к 1, если $g_2(\mathbf{w})$ — малоинформативное распределение,
6. симметрична, $s(g_1,\ g_2) = s(g_2,\ g_1)$.

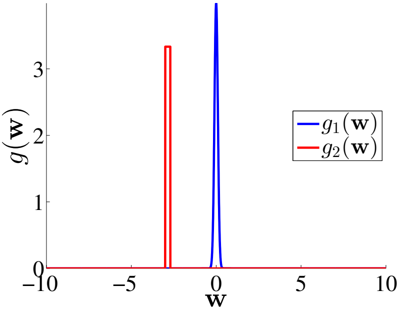
Теорема 1 (Адуенко, 2014)

Функции сходства, порожденные расстояниями Кульбака-Лейблера, Дженсона-Шеннона, Хеллингера, Бхаттачарайа, не являются корректными.

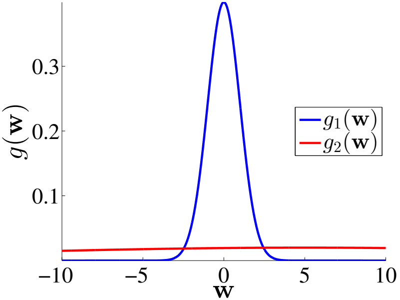
Иллюстрация требований к функции сходства

Важно, чтобы значение функции s

было близко к 1, если $g_2(\mathbf{w})$ — малоинформативное распределение.



$$g_1(\mathbf{w}) = \mathcal{N}(0,\ 0.1^2),$$
$$g_2(\mathbf{w}) = U[-3,\ -2.7].$$



$$g_1(\mathbf{w}) = \mathcal{N}(0,\ 1),$$
$$g_2(\mathbf{w}) = \mathcal{N}(5, 20^2).$$

Теорема 2 (Адуенко, 2014)

Функции сходства, порожденные дивергенциями Брегмана, симметризованными дивергенциями Брегмана и f-дивергенциями, не являются корректными.

В качестве меры сходства распределения предлагается мера сходства s-score:

$$s(g_1,\ g_2) = \frac{\int_{\mathbf{w}} g_1(\mathbf{w}) g_2(\mathbf{w}) d\mathbf{w}}{\max_{\mathbf{b}} \int_{\mathbf{w}} g_1(\mathbf{w}-\mathbf{b}) g_2(\mathbf{w}) d\mathbf{w}}.$$

Теорема 3 (Адуенко, 2014). Предлагаемая функция сходства является корректной.

Примеры:

$g_1(\mathbf{w})$	$g_2(\mathbf{w})$	$s(g_1,\ g_2)$
$U[0,\ 1]$	$U[0.5,\ 1.5]$	0.5
$U[0,\ 1]$	$U[0,\ 1]$	1
$\mathcal{N}(0,\ 1)$	$\mathcal{N}(10, 10^{10})$	1

Выражение для $s(g_1,\ g_2)$ для пары нормальных распределений

Определение. *Обобщенно-линейной моделью с натуральной функцией связи* и априорным распределением на вектор параметров $p(\mathbf{w}|\mathbf{A})$ называется вероятностная модель с совместным правдоподобием
$p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A}) = p(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A})$, где $p(\mathbf{y}|\mathbf{X},\ \mathbf{w}) = \prod\limits_{i=1}^{m} p(y_i|\mathbf{x}_i,\ \mathbf{w})$,
$p(y_i|\mathbf{x}_i,\ \mathbf{w}) = c(y_i)\exp\big(\theta_i y_i - b(\theta_i)\big)$, где $\theta_i = \mathbf{w}^{\mathsf{T}}\mathbf{x}_i$.

Теорема 4 (Адуенко, 2014).
Пусть $g_1 = \mathcal{N}(\mathbf{v}_1,\ \boldsymbol{\Sigma}_1)$, $g_2 = \mathcal{N}(\mathbf{v}_2,\ \boldsymbol{\Sigma}_2)$. Тогда выражение для $s(g_1,\ g_2)$ имеет вид

$$s(g_1,\ g_2) = \exp\left(-\frac{1}{2}(\mathbf{v}_1 - \mathbf{v}_2)^{\mathsf{T}}(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\mathbf{v}_1 - \mathbf{v}_2)\right).$$

Следствие. В случае $\boldsymbol{\Sigma}_2 = \mathbf{0}$ выражение для s-score

$$s(g_1,\ g_2) = \exp\left(-\frac{1}{2}(\mathbf{v}_2 - \mathbf{v}_1)^{\mathsf{T}}\boldsymbol{\Sigma}_1^{-1}(\mathbf{v}_2 - \mathbf{v}_1)\right).$$

Распределение s-score в условии истинности гипотезы о совпадении моделей

Рассматриваем пару обобщенно-линейных моделей с натуральной функцией связи. Введем $O_m^\delta(\mathbf{w}) = \{\mathbf{v} : \|\mathbf{H}_m^{T/2}(\mathbf{v}-\mathbf{w})\| \leq \delta\}$.

Теорема 5 (Адуенко, 2016). Пусть

- Модели f_1 и f_2 совпадают, то есть $\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{w}$;
- Априорное распределение: $\mathbf{w}_k \sim \mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_{m^k}^{-1}),\ k = 1,\ 2$;
- $\sum\limits_{i=1}^{m^k} \mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}$ имеет полный ранг для $m^k \geq m_0,\ k = 1,\ 2$;
- $\lambda_{\min}\bigl(\mathbf{H}_{m^k}(\mathbf{w})\bigr) \to \infty$ при $m^k \to \infty,\ k = 1,\ 2$;
- $\forall\, \delta > 0 \max\limits_{\mathbf{v}\in O^\delta_{m^k}(\mathbf{w})} \Bigl\|\mathbf{H}_{m^k}^{-\frac{1}{2}}\mathbf{H}_{m^k}(\mathbf{v})\mathbf{H}_{m^k}^{-\frac{\mathsf{T}}{2}} - \mathbf{I}\Bigr\| \to 0\ m^k \to \infty,\ k = 1,\ 2$;
- $\|\tilde{\mathbf{H}}_{m^1}(\hat{\mathbf{w}}_1)\|\|\tilde{\mathbf{H}}_{m^2}^{-1}(\hat{\mathbf{w}}_2)\| \xrightarrow{\mathsf{P}} 0$ при $m = \min(m^1,\ m^2) \to \infty$.

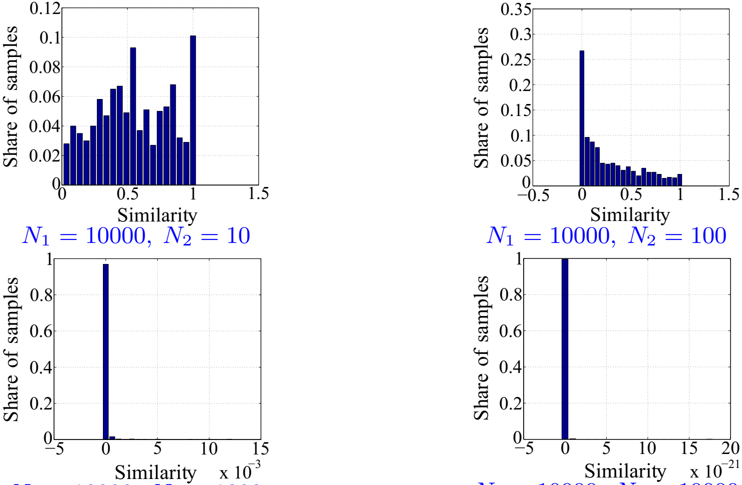
Тогда

$$-2\log \text{s-score} = (\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1)^{\mathsf{T}}\left(\tilde{\mathbf{H}}_{m^1}^{-1}(\hat{\mathbf{w}}_1) + \tilde{\mathbf{H}}_{m^2}^{-1}(\hat{\mathbf{w}}_2)\right)^{-1}(\hat{\mathbf{w}}_2 - \hat{\mathbf{w}}_1) \xrightarrow{d} \chi^2(n).$$

Следствие. Для случая $n = 2$ s-score имеет асимптотически равномерное распределение на отрезке $[0,\ 1]$.

Иллюстрация применения s-score для сравнения двух моделей, $\rho = 0.9$

Рассмотрим две близкие в терминах $\|\mathbf{w}_1 - \mathbf{w}_2\|$ модели, $\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$, $\mathbf{w}_1^{\mathsf{T}}\mathbf{w}_2 = \rho$.



$N_1 = 10000, \; N_2 = 10$

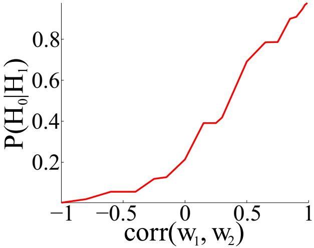
$N_1 = 10000, \; N_2 = 100$

$N_1 = 10000, \; N_2 = 1000$

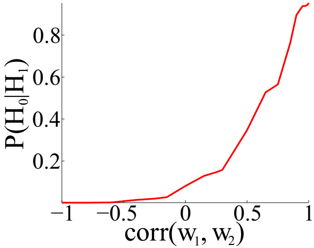
$N_1 = 10000, \; N_2 = 10000$

Зависимость $\mathsf{P}(H_0|H_1)$ от корреляции между истинными параметрами двух моделей.

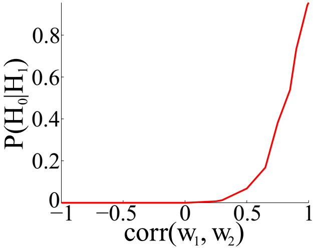
Рассмотрим две близкие в терминах $\|\mathbf{w}_1 - \mathbf{w}_2\|$ модели,
$\|\mathbf{w}_1\| = \|\mathbf{w}_2\| = 1$, $\cos(\mathbf{w}_1, \mathbf{w}_2) = \rho$.



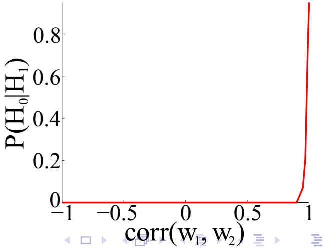
$N_1 = 10000, \ N_2 = 30$



$N_1 = 10000, \ N_2 = 50$



$N_1 = 10000, \ N_2 = 100$



$N_1 = 10000, \ N_2 = 1000$

Обозначим матрицу парных сходств $\mathbf{S} = \bigl[s_{kl}(g_k(\mathbf{w}_k),\ g_l(\mathbf{w}_l))\bigr]$,
а матрицу достигаемых уровней значимости
$\mathbf{T} = \bigl[\mathsf{P}(s(g_k(\mathbf{w}_k),\ g_l(\mathbf{w}_l)) < s_{kl} | \mathbf{w}_k = \mathbf{w}_l)\bigr],\ k,\ l = 1, \dots, K.$

1. Находим $[k^*,\ l^*] = \arg\max\limits_{k<l} t_{kl}$.
2. Если $t_{k^*l^*} < \alpha$, останавливаемся. Иначе на шаг 3.
3.
   - *Для многоуровневых моделей:*
     Объединяем модели k^*, l^* и пересчитываем $g_{k^*}(\mathbf{w}_{k^*})$.
     $\mathcal{I}_{k^*} \sqcup \mathcal{I}_{l^*} \to \mathcal{I}_{k^*},\ \mathbf{A}^*_{k^*} = \arg\max\limits_{\mathbf{A}_{k^*}} p(\mathbf{y}_{\mathcal{I}_{k^*}} | \mathbf{X}_{\mathcal{I}_{k^*}},\ \mathbf{A}_{k^*});$
     $g_{k^*}(\mathbf{w}_{k^*}) = \mathcal{N}(\mathbf{w}_{k^*} | \mathbf{w}^*_{k^*},\ \boldsymbol{\Sigma}^*_{k^*})$
   - *Для смесей моделей:*
     Объединяем модели k^*, l^* и перенастраиваем смесь моделей.
     Начальное приближение:
     $$\pi_{k^*} + \pi_{l^*} \to \pi_{k^*},\ 0 \to \pi_{l^*},\ \frac{\mathbf{w}_{k^*} + \mathbf{w}_{l^*}}{2} \to \mathbf{w}_{k^*},\ \mathbf{w}_k \to \mathbf{w}_k,\ k \neq k^*,\ l^*.$$
4. Удаляем l^*-й столбец матриц $\mathbf{S}$ и $\mathbf{T}$ и пересчитываем s_{k^*l} и t_{k^*l} для $l \neq k^*$. Переходим на шаг 1.

Литература

1. Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 48-58, 203-213, 653-674.
2. Адуенко, А. А. "Выбор мультимоделей в задачах классификации". Москва, 2017. URL: http://frccsc.ru/sites/default/files/docs/ds/002-073-05/diss/11-aduenko/11-Aduenko_main.pdf?626
3. Baghishani, Hossein, and Mohsen Mohammadzadeh. "Asymptotic normality of posterior distributions for generalized linear mixed models."Journal of Multivariate Analysis 111 (2012): 66-77.
4. MacKay, David JC. Bayesian methods for adaptive models. Diss. California Institute of Technology, 1992.
5. MacKay, David JC. "The evidence framework applied to classification networks."Neural computation 4.5 (1992): 720-736.
6. Gelman, Andrew, et al. Bayesian data analysis, 3rd edition. Chapman and Hall/CRC, 2013.
7. Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.