Лекция 4
Линейные модели в задаче классификации

Габдуллин Р.А., Макаренко В.А.

МГУ им. М.В. Ломоносова

9 февраля 2021

Задача классификации

X – множество объектов,
Y – множество ответов:

- $|Y| = 2$ – двухклассовая (binary) классификация.
- $|Y| = K$ – множественная (multiclass) классификация.

$y : X \to Y$ – неизвестная зависимость.

Дано:
$\{x_1, x_2, \dots, x_\ell\} \subset X$ – обучающая выборка,
$y_i = y(x_i), \; i = 1, \dots, \ell$ – известные ответы.

Найти:
$a : X \to Y$ – решающая функция, приближающая y на всём X.

Описание объектов. Признаки

X – множество объектов,
$f_j : X \to F_j, \quad j = 1, \dots, n$ – признаки объектов (features),

Типы признаков:

Бинарные	Binary	$F_j = \{\text{true, false}\}$
Номинальные	Categorical	F_j – конечное мн-во
Порядковые	Ordinal	F_j – конечное упорядоченное мн-во
Количественные	Numerical	$F_j = \mathbb{R}$

$(f_1(x), f_2(x), \dots, f_n(x))$ – признаковое описание объекта $x \in X$.
Матрица «объекты-признаки» (feature data)

$$F = \|f_j(x_i)\|_{\ell \times n} = \begin{pmatrix} f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots \\ f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}$$

Типы задач классификации

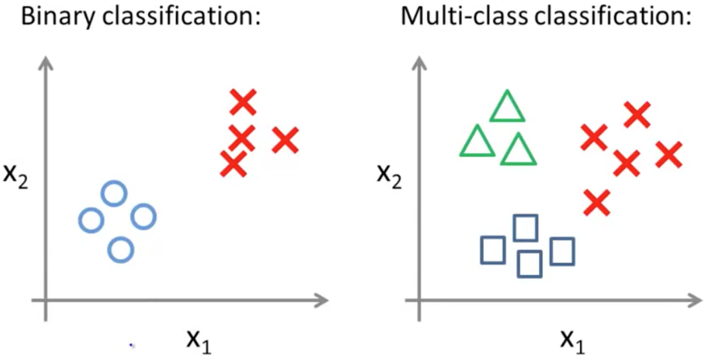


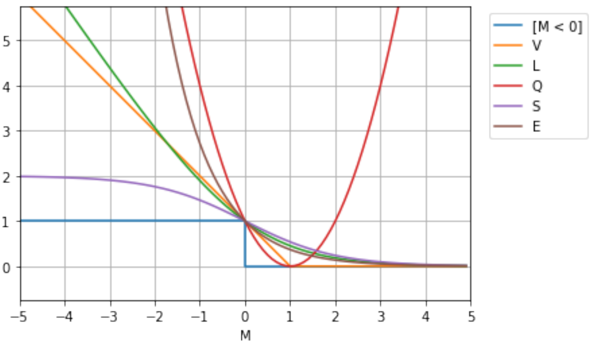
Рис.: Источник: medium.com

- Множество ответов конечно: $|Y| = K$.

Модель бинарной классификации

- Множество ответов:
$$Y = \{-1, 1\}.$$
- Семейство вещественных дискриминантных функций:
$$S = \{s(x,\theta) | \theta \in \Theta\}.$$
- Семейство алгоритмов:
$$a(x,\theta) = \operatorname{sign} s(x,\theta).$$
- Эмпирический риск:
$$Q(\theta, \mathbb{X}) = \sum_{i=1}^{\ell} [M(x_i,\theta) < 0] \equiv \sum_{i=1}^{\ell} [y_i \cdot s(x_i,\theta) < 0].$$
- Минимизация мажоранты эмпирического риска:
$$Q(\theta, \mathbb{X}) = \sum_{i=1}^{\ell} [M(x_i,\theta) < 0] \leqslant \sum_{i=1}^{\ell} \mathscr{L}(M(x_i,\theta)) \to \min_{\theta}.$$

Мажоранты эмпирического риска



Часто используемые функции потерь $\mathscr{L}$:

- $V(M) = (1 - M)_+$
- $L(M) = \log_2(1 + e^{-M})$
- $Q(M) = (1 - M)^2$
- $S(M) = 2(1 + e^{M})^{-1}$
- $E(M) = e^{-M}$

Вероятностная модель бинарной классификации

Постановка задачи.

- Объекты: $\{x_i\}_{i=1}^{\ell}$.
- Ответы:

$$
\begin{gathered}
y_1, y_2, \ldots, y_\ell \text{ — н.с.в.}, \\
y_i \sim \mathrm{Be}(p(x_i, \theta)), \quad i = 1, \ldots, \ell, \quad \theta \in \Theta, \\
\mathbb{P}(y_i = y | \theta) = p(x_i, \theta)^y \cdot (1 - p(x_i, \theta))^{1-y}
\end{gathered}
$$

- Оценить параметр $\theta \in \Theta$.

Логарифм функции правдоподобия ответов:

$$
\ln L(y_1, \ldots, y_\ell | \mathbb{X}, \theta) = \sum_{i=1}^{\ell} \Big(y_i \ln p(x_i, \theta) + (1 - y_i) \ln(1 - p(x_i, \theta)) \Big).
$$

Задача оптимизации:

$$
\sum_{i=1}^{\ell} \mathscr{L}(y_i, p(x_i, \theta)) \to \min_{\theta},
$$

где $\mathscr{L}(y, a) = -\Big(y \ln a + (1 - y) \ln(1 - a) \Big)$ – функция потерь Log Loss.

Log Loss и оценка вероятностей

Функция потерь Log Loss:

$$\mathscr{L}(y, a) = -\Big(y \ln a + (1 - y) \ln(1 - a)\Big), \quad y \in \{0, 1\}, \quad a \in [0, 1].$$

Пусть $Y \sim Be(p)$, тогда

$$a^* = \underset{a \in [0,1]}{\operatorname{argmin}}\, \mathbb{E} \mathscr{L}(Y, a) = \underset{a \in [0,1]}{\operatorname{argmax}} (p \ln a + (1 - p) \ln(1 - a)).$$

Имеем:

$$\frac{\partial (p \ln a + (1 - p) \ln(1 - a))}{\partial a} = \frac{p}{a} - \frac{1 - p}{1 - a},$$

$$\frac{p}{a^*} - \frac{1 - p}{1 - a^*} = 0 \iff a^* = p.$$

Таким образом, $p(x, \theta^*)$ – оценка $\mathbb{P}(y(x) = 1)$.

Пороговая модель бинарной классификации

- Модель ответов:

$$y_i = [s(x_i, \theta) + \varepsilon_i > 0],$$

где $\{\varepsilon_i\}$ - н.о.р.с.в. с абсолютно непрерывной симметричной ф.р. F_ε,

$$p(x_i, \theta) = \mathbb{P}(s(x_i, \theta) + \varepsilon_i > 0) = \mathbb{P}(\varepsilon_i < s(x_i, \theta)) = F_\varepsilon(s(x_i, \theta)).$$

- Модель инвариантна относительно масштабирования:

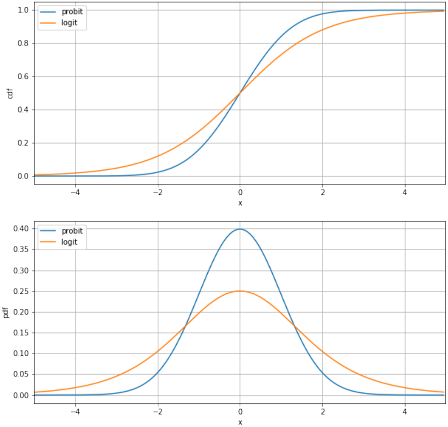
$$y_i = [\alpha \cdot (s(x_i, \theta) + \varepsilon_i) > 0] = [s(x_i, \theta) + \varepsilon_i > 0], \quad \alpha > 0,$$

поэтому можно зафиксировать любой удобный масштаб для $\{\varepsilon_i\}$.

- Задача оптимизации (минимизация Log Loss):

$$-\sum_{i=1}^{\ell} \Big(y_i \ln F_\varepsilon(s(x_i, \theta)) + (1 - y_i) \ln(1 - F_\varepsilon(s(x_i, \theta))) \Big) \to \min_\theta .$$

Модели остатков



Примеры моделей остатков:

- Probit-модель: $F_\varepsilon(x) = \Phi(x)$.
- Logit-модель: $F_\varepsilon(x) = \sigma(x) = (1 + e^{-x})^{-1}$.

Logit-модель и эмпирический риск

Минимизация Log Loss:

$$\sum_{i=1}^{\ell}\left(y_i\log\left(\frac{1}{1+\exp\{-s(x_i,\theta)\}}\right)+(1-y_i)\log\left(\frac{1}{1+\exp\{s(x_i,\theta)\}}\right)\right)\to\max_{\theta}.$$

Потери на одном наблюдении:

$$\mathscr{L}(x_i,\theta)=\begin{cases}\log\Big(1+\exp\{-s(x_i,\theta)\}\Big), y=1\\ \log\Big(1+\exp\{s(x_i,\theta)\}\Big), y=0\end{cases}=\log(1+\exp\{-M(x_i,\theta)\}).$$

Эмпирический риск:

$$Q(\mathbb{X},\theta)=\sum_{i=1}^{\ell}\log(1+\exp\{-M(x_i,\theta)\}).$$

Логистическая регрессия

- Семейство дискриминантных функций:

$$s(x,\theta) = \theta_0 + \sum_{j=1}^{n} \theta_j \cdot f_j(x) = \sum_{j=0}^{n} \theta_j \cdot f_j(x),$$

если положить $f_0(x) \equiv 1$.

- Семейство алгоритмов:

$$a(x,\theta) = [s(x,\theta) > 0].$$

- Оценка вероятности принадлежности позитивному классу:

$$\mathbb{P}(y(x) = 1) \approx \sigma(s(x,\theta)).$$

- Задача оптимизации (минимизация Log Loss):

$$-\sum_{i=1}^{\ell} \Big(y_i \ln \sigma(s(x_i,\theta)) + (1-y_i)\ln(1-\sigma(s(x_i,\theta))) \Big) \to \min_{\theta}.$$

Обучение модели логистической регрессии

Производная сигмоиды:

$$\sigma'(x) = \sigma(x)(1 - \sigma(x)).$$

Производные дискриминантной функции по параметрам:

$$\frac{\partial s(x, \theta)}{\partial \theta_j} = f_j(x).$$

Производная Log Loss по предсказанной вероятности:

$$\frac{\partial \mathscr{L}(y, a)}{\partial a} = \frac{1 - y}{1 - a} - \frac{y}{a}.$$

Обучение модели логистической регрессии

$$\sigma'(x) = \sigma(x)(1-\sigma(x)), \quad \frac{\partial s(x,\theta)}{\partial \theta_j} = f_j(x), \quad \frac{\partial \mathscr{L}(y,a)}{\partial a} = \frac{1-y}{1-a} - \frac{y}{a}.$$

Объединяем:

$$\frac{\partial \mathscr{L}(y, \sigma(s(x,\theta)))}{\partial \theta_j} = \frac{\partial \mathscr{L}}{\partial \sigma} \cdot \frac{\partial \sigma}{\partial s} \cdot \frac{\partial s}{\partial \theta_j} =$$

$$= \left(\frac{1-y}{1-\sigma(s(x,\theta))} - \frac{y}{\sigma(s(x,\theta))} \right) \cdot \sigma(s(x,\theta)) \cdot (1-\sigma(s(x,\theta))) \cdot f_j(x) =$$

$$= \Big(\sigma(s(x,\theta))(1-y) - (1-\sigma(s(x,\theta)))y \Big) \cdot f_j(x) = \Big(\sigma(s(x,\theta)) - y \Big) \cdot f_j(x).$$

Обучение модели логистической регрессии

Обозначим:

$$F = \begin{pmatrix} f_0(x_1) & f_1(x_1) & \dots & f_n(x_1) \\ \dots & \dots & \dots & \dots \\ f_0(x_\ell) & f_1(x_\ell) & \dots & f_n(x_\ell) \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \dots \\ \theta_n \end{pmatrix}, \quad y = \begin{pmatrix} y_1 \\ \dots \\ y_\ell \end{pmatrix}$$

Для $u \in \mathbb{R}^m$ обозначим:

$$\sigma(u) = \begin{pmatrix} \sigma(u_1) \\ \sigma(u_2) \\ \dots \\ \sigma(u_m) \end{pmatrix}.$$

Обучение модели логистической регрессии

Эмпирический риск:

$$Q(\mathbb{X}, \theta) = \sum_{i=1}^{\ell} \mathscr{L}(y_i, s(x_i, \theta))$$

Производные по параметрам:

$$\frac{\partial Q}{\partial \theta_j} = \sum_{i=1}^{\ell} \Big(\sigma(s(x_i, \theta)) - y_i\Big) \cdot f_j(x_i)$$

В матричных обозначениях:

$$\frac{\partial Q}{\partial \theta} = F^T(\sigma(F\theta) - y).$$

Ср. с линейной регрессией:

$$\frac{\partial Q}{\partial \theta} = 2F^T(F\theta - y).$$

Обучение модели логистической регрессии

Градиентный спуск:

- Выбрать начальное приближение $\theta^{(0)}$.
- Шаг в сторону антиградиента:

$$\theta^{(i+1)} = \theta^{(i)} - \alpha^{(i)} \cdot \frac{\partial Q}{\partial \theta}\Big|_{\theta=\theta^{(i)}} = \theta^{(i)} - \alpha^{(i)} \cdot F^T(\sigma(F\theta^{(i)}) - y).$$

- Повторять до сходимости.

Варианты:

- Классический градиентный спуск: на каждой итерации делаем шаг в сторону антиградиента эмпирического риска по всей выборке.
- Стохастический градиентный спуск: на каждой итерации делаем шаг в сторону антиградиента эмпирического риска по одному наблюдению.
- Mini-batch градиентный спуск: на каждой итерации делаем шаг в сторону антиградиента эмпирического риска по части выборки.

Множественная классификация. One-vs-all

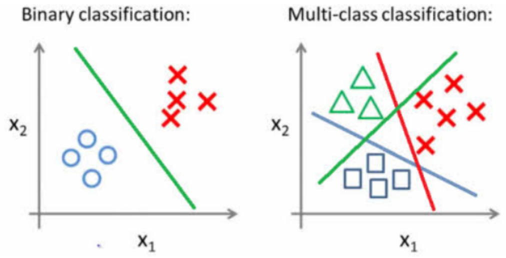


Рис.: Источник: towardsdatascience.com

- Для k-го класса обучаем свою дискриминантную функцию $s(x, \theta_k)$, отделяющую объекты этого класса от всех остальных.
- Итоговый алгоритм: $a(x) = \underset{k}{\operatorname{argmax}}\, s(x, \theta_k)$.

Множественная логистическая регрессия

- Совместно учим K дискриминантных функций. У каждого свой набор весов:

$$s(x, \theta_k) = \sum_{j=0}^{n} \theta_{k,j} \cdot f_j(x).$$

- Итоговый алгоритм:

$$a(x) = \operatorname*{argmax}_k s(x, \theta_k).$$

- Распределение моделируется с помощью Softmax:

$$\mathrm{Softmax}(s_1, \ldots, s_K) = \left(\frac{\exp(s_1)}{\sum_k \exp(s_k)}, \ldots, \frac{\exp(s_K)}{\sum_k \exp(s_k)} \right).$$

Множественная логистическая регрессия

Обучение методом максимального правдоподобия ответов:

$$\ln L(y_1, \dots, y_\ell | \mathbb{X}, \theta_1, \dots, \theta_K) =$$

$$= \sum_{i=1}^{\ell} \left(s(x_i, \theta_{y_i}) - \ln \left(\sum_k \exp\{s(x_i, \theta_k)\} \right) \right) \to \max_{\theta_1, \dots, \theta_K} .$$

Резюме лекции

- Задача классификации
  - Типы задач классификации
  - Постановка задачи. Эмпирический риск
  - Мажоранты эмпирического риска
  - Вероятностная модель бинарной классификации
  - Log Loss
  - Пороговая модель бинарной классификации
  - Probit и Logit
  - Логистическая регрессия
  - Обучение модели логистической регрессии
  - One-vs-all
  - Множественная логистическая регрессия