Лекция 3
Линейные модели в задачах регрессии и классификации

Макаренко В.А., Габдуллин Р.А.

МГУ им. М.В. Ломоносова

24 января 2023

Обучение с учителем

X – множество объектов,
Y – множество ответов,
$y : X \to Y$ – 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.

Линейная модель регрессии

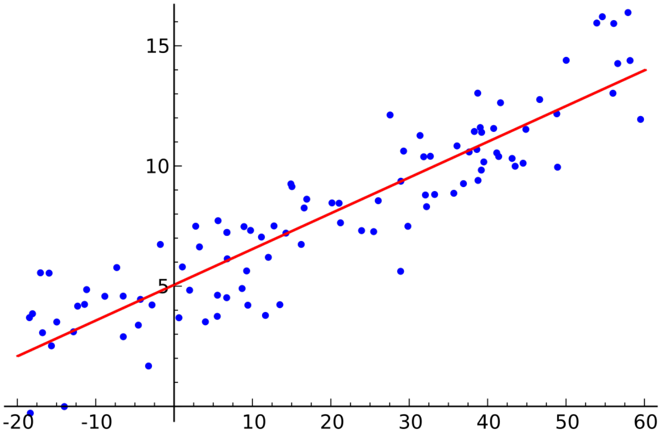


Рис.: Источник: Википедия

Линейная модель регрессии

- Семейство алгоритмов:

$$A = \{a(x,\theta) | \theta \in \mathbb{R}^{n+1}\},$$

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

если положить $f_0(x) \equiv 1$.

- Эмпирический риск:

$$Q(\theta, \mathbb{X}) = \sum_{i=1}^{\ell} w_i \cdot \Big(y_i - a(x_i,\theta)\Big)^2,$$

где w_i – вес, степень важности объекта i-го объекта.

- Метод наименьших квадратов (МНК):

$$\theta^* = \underset{\theta}{\operatorname{argmin}}\, Q(\theta, \mathbb{X}).$$

Метод максимального правдоподобия

- Вероятностная модель:

$$y_i = a(x_i, \theta) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i^2),$$

где $\{\varepsilon_i\}$ – независимые нормальные случайные величины.

- Функция правдоподобия ответов:

$$L(y_1, \dots, y_\ell | \theta) = \prod_{i=1}^{\ell} \frac{1}{\sqrt{2\pi\sigma_i^2}} \cdot \exp\left(-\frac{(y_i - a(x_i, \theta))^2}{2\sigma_i^2} \right).$$

- Метод максимального правдоподобия:

$$L(y_1, \dots, y_\ell | \theta) \to \max_\theta \iff -\ln L(y_1, \dots, y_\ell | \theta) \to \min_\theta,$$

$$\ln L(y_1, \dots, y_\ell | \theta) = \mathrm{const} + \frac{1}{2} \cdot \sum_{i=1}^{\ell} \frac{(y_i - a(x_i, \theta))^2}{\sigma_i^2},$$

$$\sum_{i=1}^{\ell} w_i \cdot (y_i - a(x_i, \theta))^2 \to \min_\theta, \quad w_i = \frac{1}{\sigma_i^2}.$$

Аналитическое решение

- Многомерная линейная регрессия:

$$a(x, \theta) = \sum_{j=0}^{n} \theta_j f_j(x).$$

- Матричная запись:

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

$$y = F\theta,$$

$$Q(\theta, \mathbb{X}) = \|y - F\theta\|^2 \to \min_{\theta}.$$

Какой геометрический смысл у МНК?

Аналитическое решение

- Многомерная линейная регрессия:

$$a(x_i, \theta) = \sum_{j=0}^{n} \theta_j F_{ij}.$$

- Необходимое условие минимума:

$$\frac{\partial Q}{\partial \theta_j} = -2\sum_{i=1}^{\ell}(y_i - a(x_i, \theta)) \cdot \frac{\partial a(x_i, \theta)}{\partial \theta_j} = -2\sum_{i=1}^{\ell}(y_i - a(x_i, \theta)) \cdot F_{ij} = 0,$$

то есть

$$\frac{\partial Q}{\partial \theta} = 2F^T(F\theta - y) = 0.$$

- Нормальная система уравнений:

$$F^T F\theta = F^T y.$$

- Решение нормальной системы уравнений:

$$\theta = (F^T F)^{-1} F^T y.$$

Численное решение

Градиентный спуск:

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

$$\theta(i+1) = \theta(i) - \alpha(i) \cdot \frac{\partial Q}{\partial \theta}\Big|_{\theta=\theta(i)} = \theta(i) - \alpha(i) \cdot 2F^T(F\theta(i) - y).$$

- 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.

Численное решение

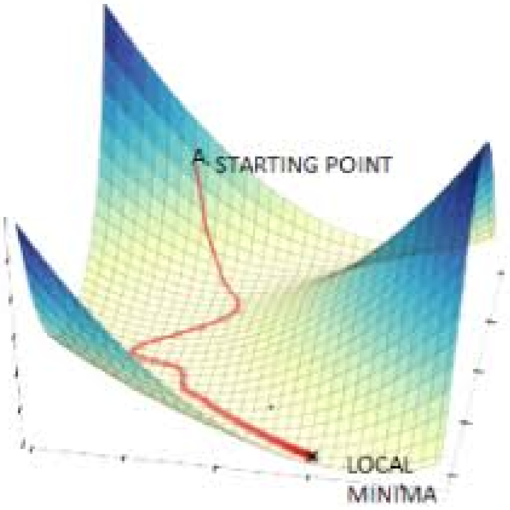


Рис.: Источник: datasciencecentral.com

Свойства оценок по МНК

Теорема (Гаусса-Маркова)

Пусть выполнены следующие условия:

- $y_i = a(x_i, \theta) + \varepsilon_i$.
- $\mathrm{rank}(F) = n + 1$.
- $\mathbb{E}\varepsilon_i = 0$.
- $\mathrm{cov}(\varepsilon) = \sigma^2 I$.

Тогда

$$\theta^* = \underset{\theta}{\mathrm{argmin}} \sum_{i=1}^{\ell} \Big(y_i - a(x_i, \theta)\Big)^2$$

являеся оптимальной оценкой в классе линейных оценок.

Проблема мультиколлинеарности

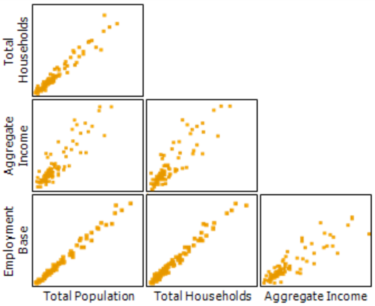


Рис.: Источник: medium.com

- Два или более признаков почти линейно зависимы.
- Решение получается неустойчивым.
- Неустойчивое решение ведет к переобучению.

Гребневая (Ridge) регрессия (L_2-регуляризация)

- Эмпирический риск:

$$Q(\theta, \mathbb{X}) = \|y - F\theta\|^2 + \lambda\|\theta\|^2, \quad \lambda > 0.$$

- Необходимое условие минимума:

$$\frac{\partial Q}{\partial \theta} = 2F^T(F\theta - y) + 2\lambda\theta = 2((F^TF + \lambda I)\theta - F^Ty) = 0,$$

  где I – единичная матрица.

- Решение в явном виде:

$$\theta = (F^TF + \lambda I)^{-1}F^Ty.$$

- Собственные числа матрицы F^TF становятся больше на λ, при этом собственные векторы – те же самые.

Вероятностная интерпретация гребневой регрессии

- Вероятностная модель:

$$\theta \sim \mathcal{N}(0, \tau^2 I),$$

$$y_i = a(x_i, \theta) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2),$$

где $\{\varepsilon_i\}$ – независимые нормальные случайные величины.

- Апостериорное распределение весов:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}.$$

- Максимум апостериорного распределения:

$$p(\theta|y) \to \max_{\theta} \iff -\ln p(y|\theta)p(\theta) \to \min_{\theta},$$

$$-\ln p(y|\theta)p(\theta) = \text{const} + \frac{1}{2} \cdot \left(\sum_{i=1}^{\ell} \frac{(y_i - a(x_i, \theta))^2}{\sigma^2} + \sum_{j=0}^{n} \frac{\theta_j^2}{\tau^2} \right),$$

то есть $\lambda = \frac{\sigma^2}{\tau^2}$.

Lasso-регрессия (L_1-регуляризация)

- Эмпирический риск:

$$Q(\theta, \mathbb{X}) = \|y - F\theta\|^2 + \lambda \sum_{j=0}^{n} |\theta_j|, \quad \lambda > 0.$$

- Вероятностная интерпретация: веса независимы и имеют одно и то же распределение Лапласа.

Ridge vs. Lasso

Эквивалентные задачи поиска условного минимума.

- Для *Ridge*-регрессии (L_2):

$$\|y - F\theta\|^2 \to \min_{\theta}, \quad \sum_{j=0}^{n} \theta_j^2 \leqslant \varkappa_1.$$

- Для *Lasso*-регрессси (L_1):

$$\|y - F\theta\|^2 \to \min_{\theta}, \quad \sum_{j=0}^{n} |\theta_j| \leqslant \varkappa_2.$$

Ridge vs. Lasso

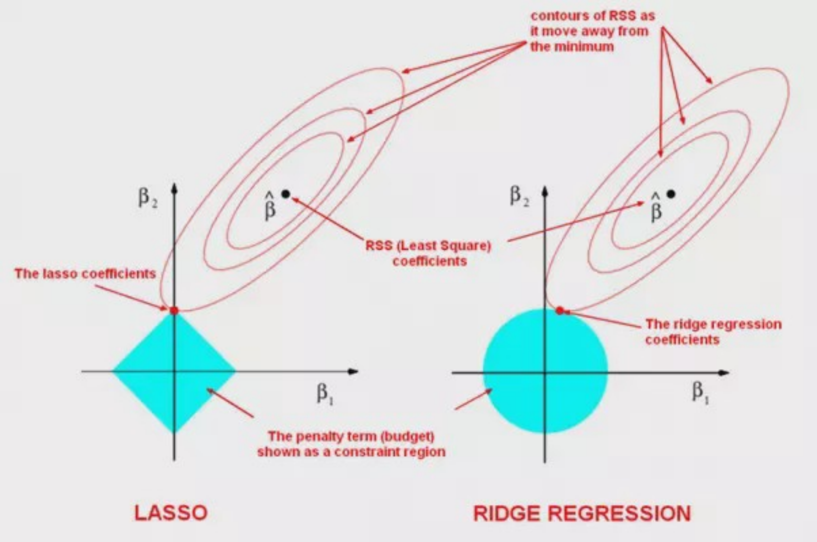


Рис.: Источник: medium.com

Ridge vs. Lasso

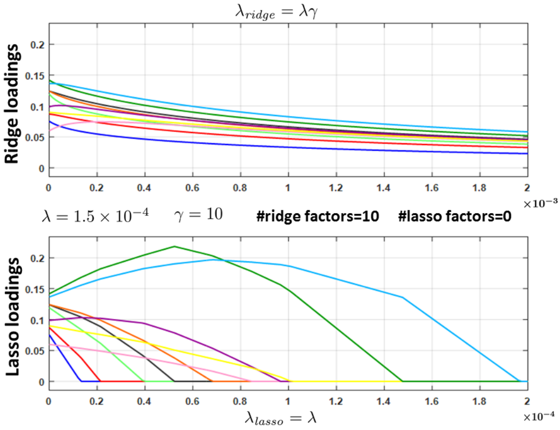


Рис.: Источник: arpm.co

Обязательно ли штрафовать за отклонение коэффициентов именно от нуля?

Масштабирование вещественных признаков

Веса модели чувствительны к сдвиг-масштабным преобразованиям признаков:

$$\widetilde{f}_j(x) = \alpha_j + \beta_j f_j(x),$$

$$\theta_0 + \sum_{j=1}^{n} \theta_j \cdot \frac{\widetilde{f}_j(x) - \alpha_j}{\beta_j} = \left(\theta_0 - \sum_{j=1}^{n} \frac{\alpha_j \theta_j}{\beta_j} \right) + \sum_{j=1}^{n} \frac{\theta_j}{\beta_j} \cdot \widetilde{f}_j(x) =$$

$$= \widetilde{\theta}_0 + \sum_{j=1}^{n} \widetilde{\theta}_j . \widetilde{f}_j(x).$$

Часто признаки преобразовывают, приводя их к единой шкале:

$$\widetilde{f}_j(x) = \frac{f_j(x) - \mathbb{E} f_j(x)}{\sqrt{\mathbb{D} f_j(x)}} \quad \text{или} \quad \widetilde{f}_j(x) = \frac{f_j(x) - \min\limits_i f_j(x_i)}{\max\limits_i f_j(x_i) - \min\limits_i f_j(x_i)}.$$

Преобразование категориальных признаков

Пусть признак f_j может принимать одно из K возможных значений:

$$f_j(x) \in \{1, \ldots, K\}.$$

One-hot кодирование. Признак f_j «разбивается» на $K - 1$ признаков:

$$\widetilde{f}_{j,k}(x) = [f_j(x) = k], \quad k = 1, \ldots, K - 1.$$

Усложнение модели

Конструирование новых признаков на основе имеющихся:

- Применение функций к признакам (степени, логарифм, экспонента, ...).
- Добавление взаимодействий между признаками (перемножение, деление, ...).

Задача классификации

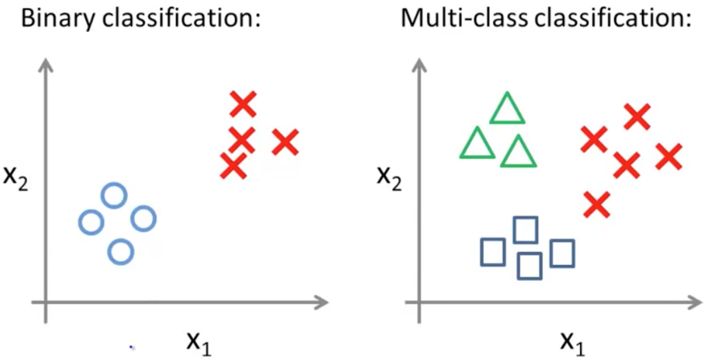
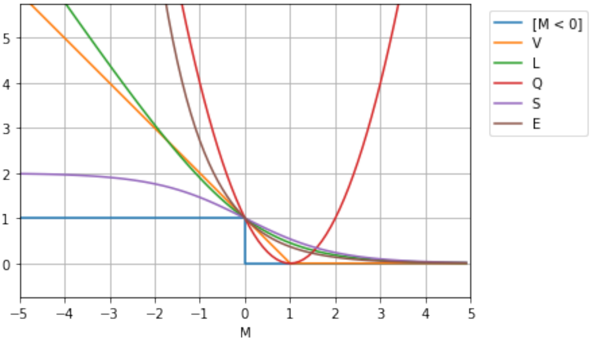


Рис.: Источник: medium.com

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:
$$Q(\theta, \mathbb{X}) = \sum_{i=1}^{\ell} [M(x_i, \theta) < 0] \leqslant \sum_{i=1}^{\ell} L(M(x_i, \theta)) \to \min_{\theta}.$$

Мажоранты эмпирического риска



Часто используемые функции потерь 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:

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

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

Log Loss и оценка вероятностей

Функция потерь Log Loss:

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

Пусть $Y \sim Be(p)$, тогда

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

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,

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

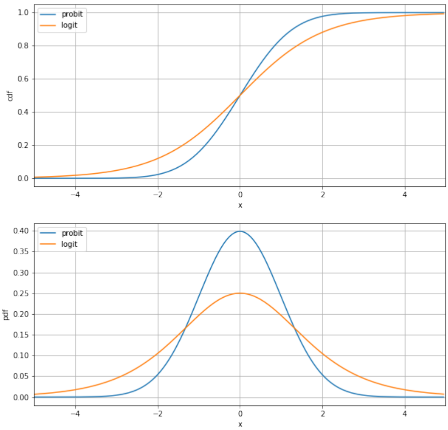
- Модель инвариантна относительно масштабирования:

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

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

- Задача оптимизации (минимизация Log Loss):

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



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:

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

Потери на одном наблюдении:

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

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):

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

Почему лог. регрессию нужно обязательно учить с регуляризатором?

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:

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

Обучение модели логистической регрессии

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

Объединяем:

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

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

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

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:

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

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:

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

- 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

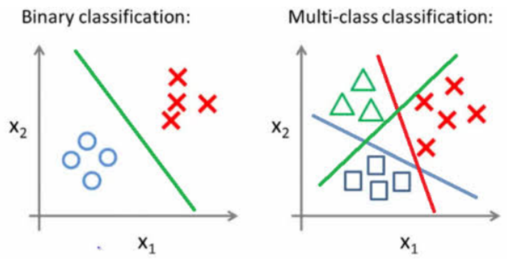


Рис.: Источник: towardsdatascience.com

- 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:

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

Как изменится Softmax, если ко всем компонентам добавить одно и то же число? Как связаны между собой бинарный случай и множественный с двумя классами?

Множественная логистическая регрессия

Обучение методом максимального правдоподобия ответов:

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

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

Что делать, если у нас классы не являются взаимоисключающими (например, предсказываем хэштеги для для статей) ?

Резюме лекции

- Линейная модель регрессии
  - МНК и ММП, их связь
  - Аналитическое решение
  - Численное решение. Градиентный спуск
  - Проблема мультиколлинеарности
  - L_1 и L_2-регуляризации
  - Вероятностный смысл регуляризации
  - Преобразование и конструирование признаков
- 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