Машинное обучение
в науках о Земле

Михаил Криницкий

к.т.н., н.с.
Институт океанологии РАН им. П.П. Ширшова

Лаборатория взаимодействия океана и атмосферы и мониторинга климатических изменений (ЛВОАМКИ)

ПЛАН ЛЕКЦИИ

- Обобщенные линейные модели (generalized linear models, GLM)
- Обобщенные аддитивные модели (generalized additive models, GAM)
- Искусственная нейронная сеть

Линейная регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\eta_i = \theta \cdot x_i$$

$$\mu(\theta, x_i) = f(\eta_i)$$

previously on ML4ES

Линейная регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\eta_i = \theta \cdot x_i$$

$$\mu(\theta, x_i) = \eta_i$$

Логистическая регрессия

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$

$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$

$$\eta_i^1 = \theta_1 \cdot x_i$$

$$p(\theta_1, x_i) = \mathrm{sigmoid}(\eta_i^1) = \frac{1}{1 + \exp(-\eta_i^1)}$$

previously on ML4ES

Линейная регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\eta_i = \theta \cdot x_i$$

$$\mu(\theta, x_i) = \eta_i$$

Логистическая регрессия

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$

$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$

$$\eta_i^1 = \theta_1 \cdot x_i = \ln \frac{p_i}{1 - p_i}$$

$$p(\theta_1, x_i) = \mathrm{sigmoid}(\eta_i^1) = \frac{1}{1 + \exp(-\eta_i^1)}$$

Линейная регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\eta_i = \theta \cdot x_i$$

$$\mu(\theta, x_i) = \eta_i$$

Логистическая регрессия

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$

$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$

$$\eta_i^1 = \theta_1 \cdot x_i = \ln \frac{p_i}{1 - p_i}$$

$$p(\theta_1, x_i) = \mathrm{sigmoid}(\eta_i^1) = \frac{1}{1 + \exp(-\eta_i^1)}$$

Мультиномиальная логистическая регрессия

$$y_{ik} \sim \mathcal{B}\big(p_k(\theta, x)\big)$$

$$p_k(\theta, x_i) \propto \exp(\theta_k \cdot x_i)$$

$$\eta_i^k = \theta_k \cdot x_i = \ln p_{ik} + C$$

$$p_k(\theta_1, x_i) = \mathrm{softmax}\big(\eta_i^k, \{\eta_i^k\}\big) = \frac{\exp \eta_i^k}{C^*}$$

$$C_i^* = \sum_{j=1}^{K} \exp(\eta_i^{(j)})$$

$$\frac{1}{C_i^*} \sum_{k=1}^{K} \exp \eta_i^{(k)} = 1$$

Линейная регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\color{red}{\eta_i = \theta \cdot x_i}$$

$$\mu(\theta, x_i) = \eta_i$$

$$\eta = g(\mu) = I(\mu)$$

Логистическая регрессия

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$

$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$

$$\color{red}{\eta_i^1 = \theta_1 \cdot x_i} = \ln \frac{p_i}{1 - p_i}$$

$$g(t) = \ln \frac{t}{1-t}$$

$$p(\theta_1, x_i) = \mathrm{sigmoid}(\eta_i^1) = \frac{1}{1 + \exp(-\eta_i^1)}$$

Мультиномиальная логистическая регрессия

$$y_{ik} \sim \mathcal{B}\big(p_k(\theta, x)\big)$$

$$\eta = g(p) =$$

$$p_k(\theta, x_i) \propto \exp(\theta_k \cdot x_i)$$

$$\color{red}{\eta_i^k = \theta_k \cdot x_i} = \ln p_{ik} + C$$

$$= \ln p + C$$

$$p_k(\theta_1, x_i) = \mathrm{softmax}\big(\eta_i^k, \{\eta_i^k\}\big) = \frac{\exp \eta_i^k}{C^*}$$

Обобщенные линейные модели: модели, в которых некоторая функция $g(\cdot)$ мат.ожидания параметра распределения целевой переменной вычисляется как **линейная функция** признакового описания объектов (событий)

$g(\cdot)$ – т.н. функция связи (link function)

Линейная регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\eta_i = \theta \cdot x_i$$

$$\mu(\theta, x_i) = \eta_i$$

Логистическая регрессия

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$

$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$

$$\eta_i^1 = \theta_1 \cdot x_i = \ln \frac{p_i}{1 - p_i}$$

$$p(\theta_1, x_i) = \mathrm{sigmoid}(\eta_i^1) = \frac{1}{1 + \exp(-\eta_i^1)}$$

Мультиномиальная логистическая регрессия

$$y_{ik} \sim \mathcal{B}\big(p_k(\theta, x)\big)$$

$$p_k(\theta, x_i) \propto \exp(\theta_k \cdot x_i)$$

$$\eta_i^k = \theta_k \cdot x_i = \ln p_{ik} + C$$

$$p_k(\theta_1, x_i) = \mathrm{softmax}\big(\eta_i^k, \{\eta_i^k\}\big) = \frac{\exp \eta_i^k}{C^*}$$

Обобщенные линейные модели: модели, в которых некоторая функция $g(\cdot)$ мат.ожидания параметра распределения целевой переменной вычисляется как линейная функция признакового описания объектов (событий)

$g(\cdot)$ – т.н. функция связи (link function)

Линейная регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\eta_i = \theta \cdot x_i$$

$$\mu(\theta, x_i) = \boldsymbol{\eta_i}$$

$\mu = I(\eta)$

Логистическая регрессия

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$

$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$

$$\eta_i^1 = \theta_1 \cdot x_i = \ln \frac{p_i}{1 - p_i}$$

$$p(\theta_1, x_i) = \mathbf{sigmoid}(\boldsymbol{\eta_i^1}) = \frac{1}{1 + \exp(-\eta_i^1)}$$

Мультиномиальная логистическая регрессия

$$y_{ik} \sim \mathcal{B}\big(p_k(\theta, x)\big)$$

$$p_k(\theta, x_i) \propto \exp(\theta_k \cdot x_i)$$

$$\eta_i^k = \theta_k \cdot x_i = \ln p_{ik} + C$$

$$p_k(\theta_1, x_i) = \mathbf{softmax}\big(\boldsymbol{\eta_i^k}, \{\boldsymbol{\eta_i^k}\}\big) = \frac{\exp \eta_i^k}{C^*}$$

Альтернативно (некорректно, но легче понять):

Обобщенные линейные модели: модели, в которых мат.ожидание параметра распределения целевой переменной вычисляется как некоторая **функция** $\boldsymbol{g^{-1}}(\cdot)$ от линейной функции η признакового описания объектов (событий) x_i

$\boldsymbol{g}(\cdot)$ – т.н. функция связи (link function)

$$y \sim \mathcal{N}(\mu, \sigma^2) \qquad \mu = \theta x$$

Как зависит вид функции потерь от вида функции связи?

$$\sigma \quad g^{-1}(\cdot)$$

$$\mu = \exp(\theta x)$$

$$\mu = g^{-1}(\theta x)$$

Как зависит вид функции потерь от вида функции связи?

Линейная регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$
$$\eta_i = \theta \cdot x_i$$
$$\mu(\theta, x_i) = \boldsymbol{\eta_i}$$

$$p(y_i, x_i, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)$$

$$L(\mathcal{T}, \theta) = \prod_{\mathcal{T}} p(y_i, \mu_i)$$

$$\ell(\mathcal{T}, \theta) = \ln L(\mathcal{T}, \theta) = \sum_{\mathcal{T}} \ln p(y_i, \mu_i) = \ln \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{\mathcal{T}} \left(-\frac{(y_i - \mu_i)^2}{2\sigma^2}\right)$$

$$\theta = \underset{\Theta}{\operatorname{argmax}}\, \ell(\mathcal{T}, \theta) = \underset{\Theta}{\operatorname{argmin}} \left(\ln \frac{1}{2\sigma^2\sqrt{2\pi\sigma^2}} * \sum_{\mathcal{T}} (y_i - \mu_i)^2\right)$$

$$\mathcal{L}(\mathcal{T}, \theta) = C \sum_{\mathcal{T}} (y_i - \mu_i)^2$$

Как зависит вид функции потерь от вида функции связи?

Логистическая регрессия

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$
$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$
$$\eta_i^1 = \theta_1 \cdot x_i = \ln \frac{p_i}{1 - p_i}$$
$$\color{red}{p(\theta_1, x_i)} = \color{blue}{\mathbf{sigmoid}(\boldsymbol{\eta}_i^1)} = \frac{1}{1 + \exp(-\eta_i^1)}$$

$$L(\mathcal{T}, \theta) = \prod_{\mathcal{T}} p(y_i, p_i) = \prod_{i=1}^{N} \left(p_i^{y_i} * (1 - p_i)^{(1-y_i)}\right)$$

$$\ell(\mathcal{T}, \theta) = \ln(L(\mathcal{T}, \theta)) = \sum_i \log\left(p_i^{y_i}\right) + \sum_i \log\left((1 - p_i)^{(1-y_i)}\right) =$$
$$= \sum_{\mathcal{T}} (y_i * \log p_i + (1 - y_i) * \log(1 - p_i))$$

$$\theta = \operatorname*{argmax}_{\Theta} \ell(\mathcal{T}, \theta) = \operatorname*{argmin}_{\Theta} \left(-\sum_{\mathcal{T}} (y_i * \log p_i + (1 - y_i) * \log(1 - p_i))\right)$$

$$\mathcal{L}(\mathcal{T}, \theta) = -\sum_{\mathcal{T}} (y_i * \log p_i + (1 - y_i) * \log(1 - p_i))$$

Как зависит вид функции потерь от вида функции связи?

НИКАК

вид функции потерь $\mathcal{L}(\mathcal{T},\theta)$ зависит от вида распределения целевой переменной y_i, но не от вида функции связи $g(\cdot)$

$\frac{\partial g^{-1}(\theta x)}{\partial \theta} = x \quad \mathcal{L} = (y - g^{-1}(\theta x))^2$

- обратная функция связи $g^{-1}(\cdot)$ должна отображать R^1 (множество значений произвольной линейной функции $\theta \cdot x_i$) на множество параметров распределения переменной y_i

  $p, \mu, p_k = g^{-1}(\theta x)$

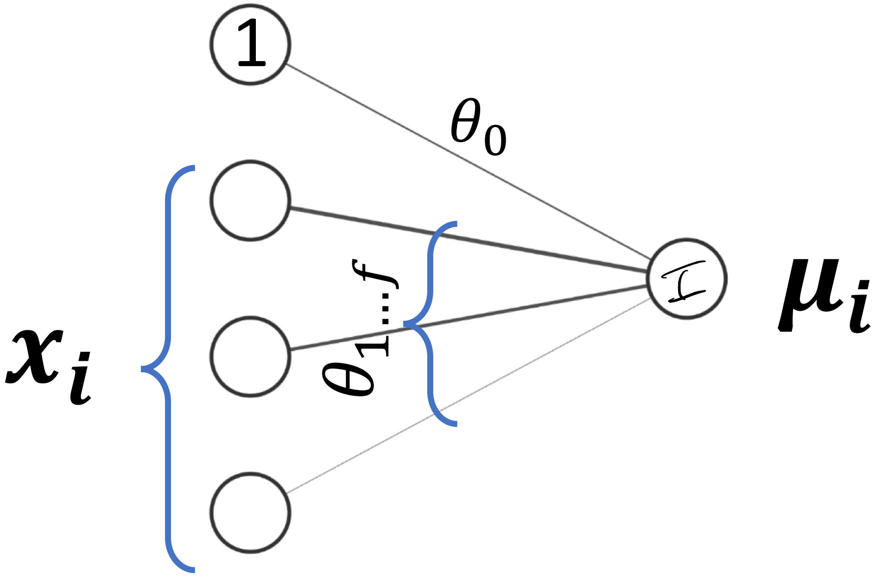
- вычисление параметра (параметров) распределения переменной y_i в модели производится согласно принципу GLM

  $I \frac{\partial g^{-1}(\theta x)}{\partial \theta} \quad \mu = \theta x \quad g^{-1}(\cdot) = I$

- при таких условиях вычисление правдоподобия выборки $\mathcal{T}$ производится независимо от вида функции связи $g(\cdot)$ => вычисление функции потерь в подходе максимизации правдоподобия также производится независимо от вида функции связи $g(\cdot)$

Диаграммы GLM

Линейная регрессия



$$\boldsymbol{y} \sim \boldsymbol{\mathcal{N}}(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$$

$$\eta_i = \theta \cdot x_i$$

$$\mu(\theta, x_i) = \boldsymbol{\eta_i}$$

$$\mathcal{L}(\mathcal{T}, \theta) = C \sum_{\mathcal{T}} (y_i - \mu_i)^2$$

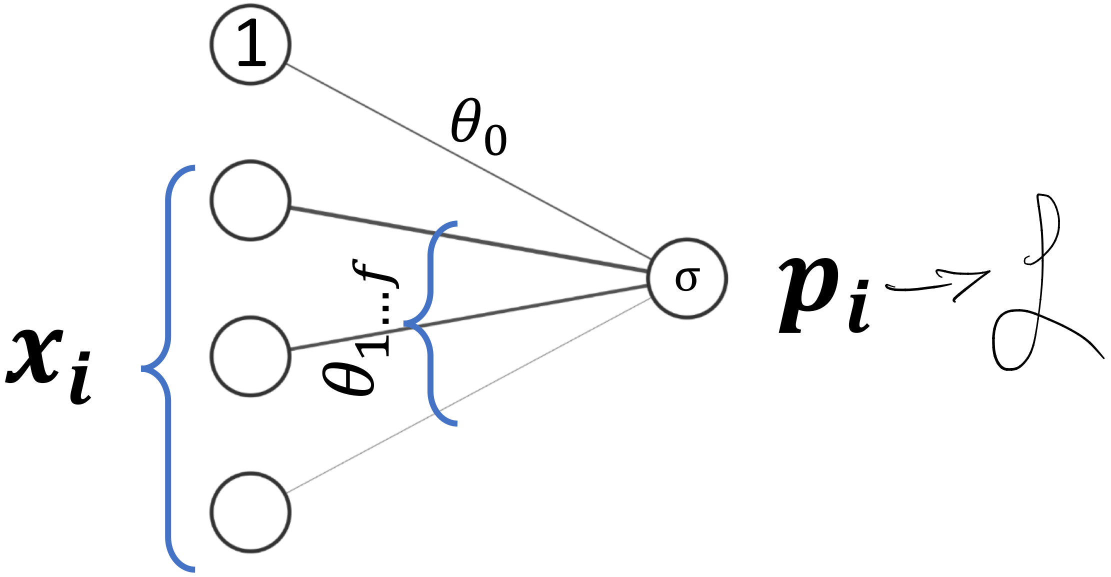
$$\theta = \underset{\Theta}{\operatorname{argmin}}\big(\mathcal{L}(\mathcal{T}, \theta)\big)$$

$$\frac{\partial \mathcal{L}(\mathcal{T}, \theta)}{\partial \theta} = -2X^T Y + 2X^T X\theta$$

=> градиентная оптимизация

Диаграммы GLM

Логистическая регрессия



$$\boldsymbol{y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)}$$

$$\eta_i^1 = \theta_1 \cdot x_i$$

$$p(\theta_1, x_i) = \boldsymbol{\sigma\left(\eta_i^1\right)}$$

$$\mathcal{L}(\mathcal{T}, \theta) = -\sum_{\mathcal{T}} (y_i * \log p_i + (1 - y_i) * \log(1 - p_i))$$

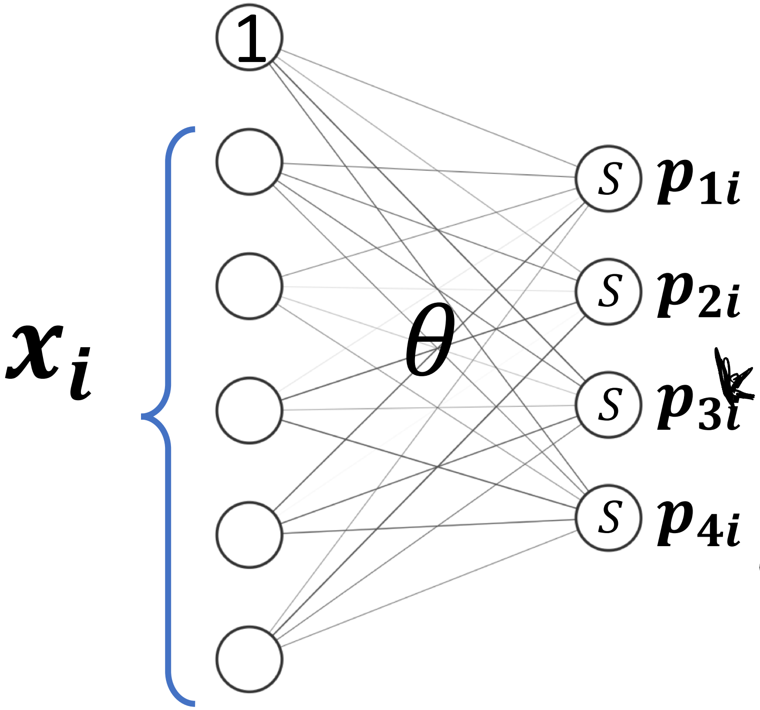
$$\theta = \underset{\Theta}{\operatorname{argmin}}\big(\mathcal{L}(\mathcal{T}, \theta)\big)$$

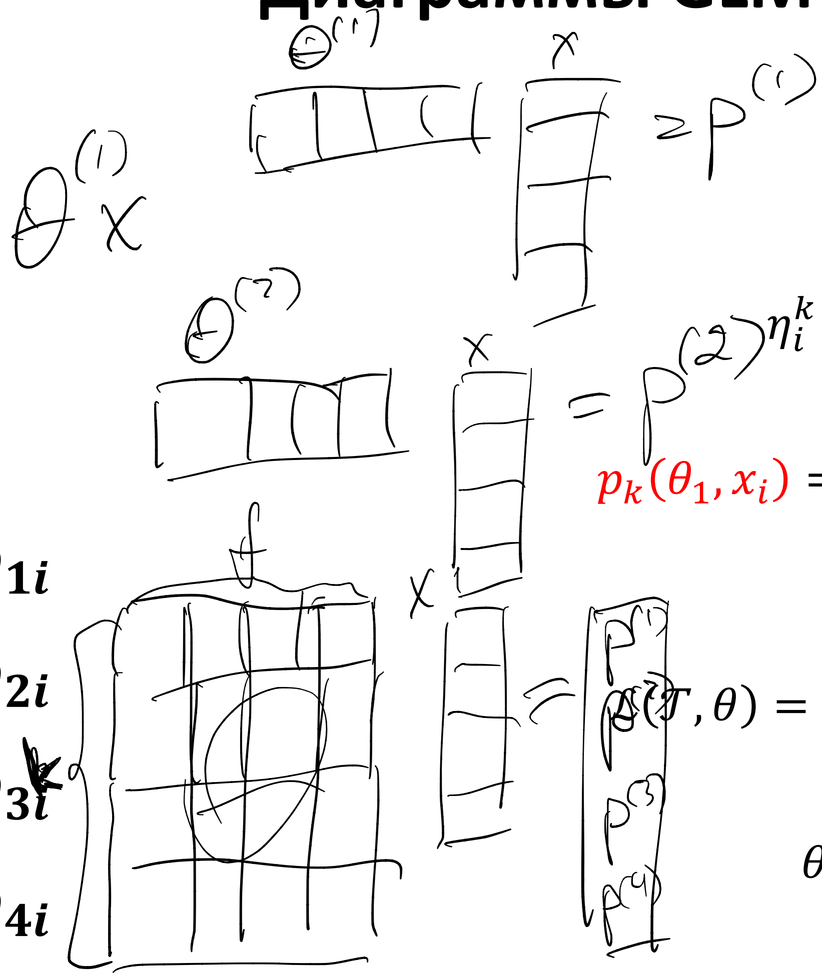
$$\frac{\partial \mathcal{L}(\mathcal{T}, \theta)}{\partial \theta} = -\sum_i \big(y_i - p(\theta, x_i)\big) x_i$$

=> градиентная оптимизация

Диаграммы GLM

Мультиномиальная логистическая регрессия





$$y_{ik} \sim \mathcal{B}(p_k(\theta, x))$$

$$\eta_i^k = \theta_k \cdot x_i = \ln p_{ik} + C$$

$$p_k(\theta_1, x_i) = \mathbf{softmax}(\boldsymbol{\eta}_i^k, \{\boldsymbol{\eta}_i^k\}) = \frac{\exp \eta_i^k}{C^*}$$

$$\mathcal{L}(\mathcal{T}, \theta) = -\sum_{\mathcal{T}} \sum_{k=1}^{K} ([y_i == k] * \log p_{ik})$$

$$\theta = \underset{\Theta}{\operatorname{argmin}}(\mathcal{L}(\mathcal{T}, \theta))$$

$$\nabla_{\theta_k} \mathcal{L}(\mathcal{T}, \theta) = -\sum_{\mathcal{T}} ([y_i == k] - p_{ik}) x_i$$

=> градиентная оптимизация

Регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\eta_i = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \dots f_f(x_i^{(f)})\right]$$

$$\mu(\theta, x_i) = \eta_i = g(x)$$

Мультиномиальная классификация

$$y_{ik} \sim \mathcal{B}\big(p_k(\theta, x)\big)$$

$$p_k(\theta, x_i) \propto \exp(\theta_k \cdot x_i) = g(x)$$

$$\eta_i^k = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \dots f_f(x_i^{(f)})\right] = \ln p_{ik} + C$$

$$p(\theta_1, x_i) = \mathrm{softmax}\big(\eta_i^k, \{\eta_i^k\}\big) = \frac{\exp \eta_i^k}{C^*}$$

Бинарная классификация

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$

$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$

$$\eta_i^1 = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \dots f_f(x_i^{(f)})\right] = \ln \frac{p_i}{1 - p_i}$$

$$p(\theta_1, x_i) = \mathrm{sigmoid}(\eta_i^1) = \frac{1}{1 + \exp(-\eta_i^1)}$$

Обобщенные аддитивные модели – это обобщенные **линейные** модели, в которых некоторая функция $g(\cdot)$ мат.ожидания параметра распределения целевой переменной вычисляется как **линейная функция** некоторых других гладких функций (часто нелинейных, но необязательно) признакового описания объектов (событий)

Регрессия

$$y_i \sim \mathcal{N}(\mu(\theta, x_i), \sigma^2)$$

$$\eta_i = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \dots f_f\left(x_i^{(f)}\right)\right]$$

$$\mu(\theta, x_i) = \boldsymbol{\eta_i}$$

Бинарная классификация

$$y_i \sim \mathcal{B}\big(p(\theta, x_i)\big)$$

$$p(\theta, x_i) \propto \exp(\theta_1 \cdot x_i)$$

$$\eta_i^1 = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \dots f_f\left(x_i^{(f)}\right)\right] = \ln \frac{p_i}{1 - p_i}$$

$$p(\theta_1, x_i) = \mathbf{sigmoid}(\boldsymbol{\eta_i^1}) = \frac{1}{1 + \exp(-\eta_i^1)}$$

Мультиномиальная классификация

$$y_{ik} \sim \mathcal{B}\big(p_k(\theta, x)\big)$$

$$p_k(\theta, x_i) \propto \exp(\theta_k \cdot x_i)$$

$$\eta_i^k = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \dots f_f\left(x_i^{(f)}\right)\right] = \ln p_{ik} + C$$

$$p(\theta_1, x_i) = \mathbf{softmax}\big(\boldsymbol{\eta_i^k}, \{\boldsymbol{\eta_i^k}\}\big) = \frac{\exp \eta_i^k}{C^*}$$

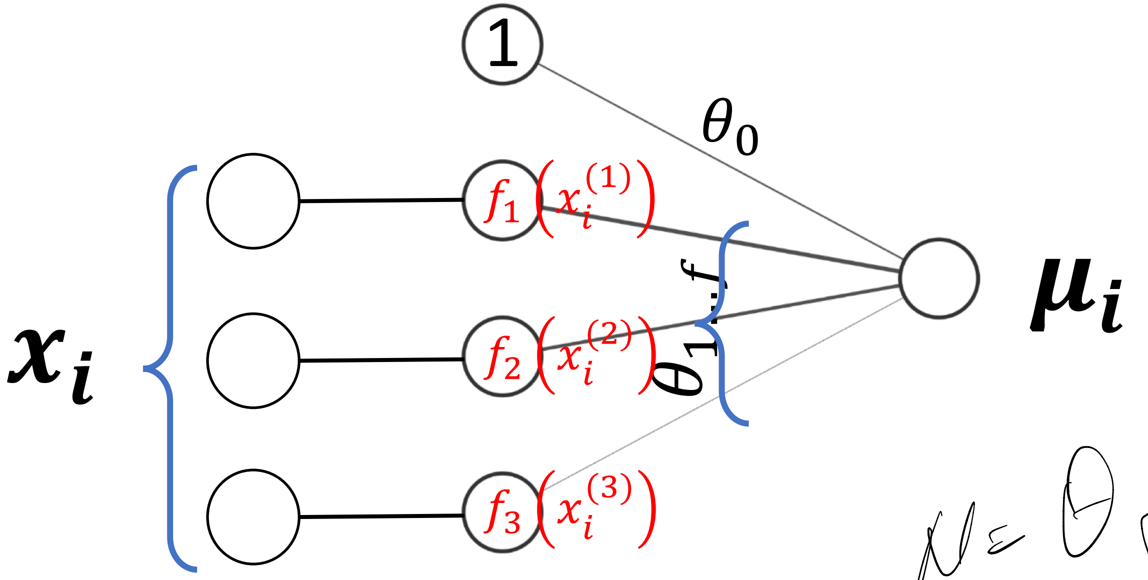
Альтернативно (некорректно, но легче понять):

Обобщенные аддитивные модели – это обобщенные линейные модели, в которых в качестве признаков объектов x_i выступают некоторые гладкие (часто нелинейные) функции этих признаков

Диаграммы GAM

Регрессия

$$\frac{\partial \mathcal{L}}{\partial \theta} = 2(\mu - y)\cdot\begin{pmatrix} f_1(x_1) \\ f_2(x_2) \\ f_3(x_3) \\ \vdots \\ f_f(x_f) \end{pmatrix}$$



$$\mu = \theta f^{(i)}\left(x^{(i)}\right)$$

$$\frac{\partial \mu}{\partial \theta} = f^{(i)}\left(x^{(i)}\right)$$

$$\boldsymbol{y \sim \mathcal{N}(\mu, \sigma^2)}$$

$$\eta_i = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \ldots f_f\left(x_i^{(f)}\right)\right]$$

$$\mu(\theta, x_i) = \boldsymbol{\eta_i}$$

$$\mathcal{L}(\mathcal{T}, \theta) = C \sum_{\mathcal{T}} (y_i - \mu_i)^2$$

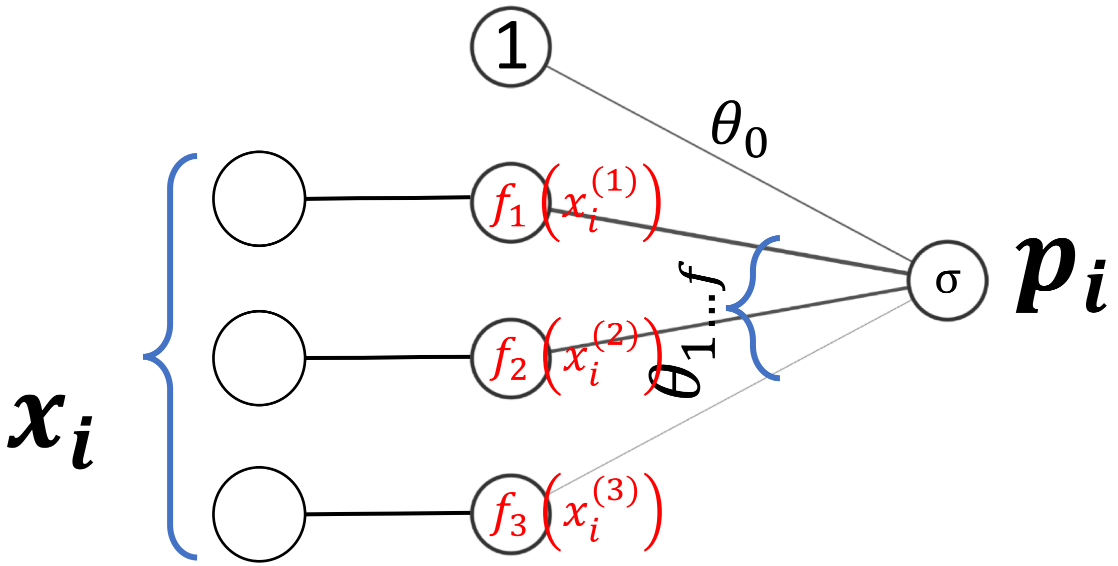
$$\theta = \underset{\Theta}{\operatorname{argmin}}\left(\mathcal{L}(\mathcal{T}, \theta)\right)$$

$$\frac{\partial \mathcal{L}(\mathcal{T}, \theta)}{\partial \theta} = ?$$

=> градиентная оптимизация

Диаграммы GAM

Бинарная классификация



$$\boldsymbol{y_i} \sim \boldsymbol{\mathcal{B}}\left(\boldsymbol{p}(\boldsymbol{\theta}, \boldsymbol{x_i})\right)$$

$$\eta_i^1 = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \dots f_f\left(x_i^{(f)}\right)\right]$$

$$p(\theta_1, x_i) = \boldsymbol{\sigma}\left(\boldsymbol{\eta_i^1}\right)$$

$$\mathcal{L}(\mathcal{T}, \theta) = -\sum_{\mathcal{T}} \left(y_i * \log p_i + (1 - y_i) * \log(1 - p_i)\right)$$

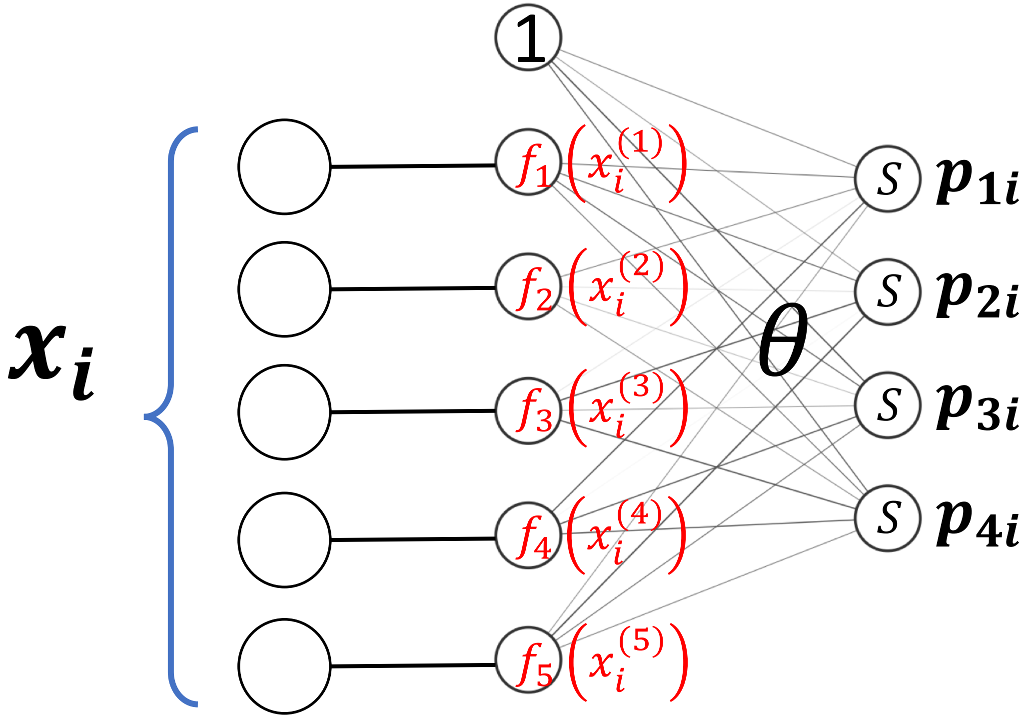
$$\theta = \underset{\Theta}{\operatorname{argmin}}\left(\mathcal{L}(\mathcal{T}, \theta)\right)$$

$$\frac{\partial \mathcal{L}(\mathcal{T}, \theta)}{\partial \theta} = ?$$

=> градиентная оптимизация

Диаграммы GAM

Мультиномиальная классификация



$$y_{ik} \sim \mathcal{B}\big(p_k(\theta, x)\big)$$

$$\eta_i^k = \theta \cdot \left[f_1\left(x_i^{(1)}\right), f_2\left(x_i^{(2)}\right), f_3\left(x_i^{(3)}\right) \dots f_f\left(x_i^{(f)}\right)\right]$$

$$\ln p_{ik} + C = \eta_i^k$$

$$p_k(\theta_1, x_i) = \mathbf{softmax}\left(\boldsymbol{\eta}_i^k, \{\boldsymbol{\eta}_i^k\}\right) = \frac{\exp \eta_i^k}{C^*}$$

$$\mathcal{L}(\mathcal{T}, \theta) = -\sum_{\mathcal{T}} \sum_{k=1}^{K} \left([y_i == k] * \log p_{ik}\right)$$

$$\theta = \underset{\Theta}{\operatorname{argmin}}\big(\mathcal{L}(\mathcal{T}, \theta)\big)$$

$$\nabla_{\theta_k} \mathcal{L}(\mathcal{T}, \theta) = ?$$

=> градиентная оптимизация

Обобщенные аддитивные модели: за и против

ЗА

- Довольно простые, но при этом предоставляют достаточно свободы в выборе нелинейных преобразований $f_j\left(x_i^{(j)}\right)$ исходных признаков объектов (событий);
- Подбирать функции f_j - иногда проще, чем подбирать степени полинома;
- Позволяют применять несколько разных нелинейных функций $f_j\left(x_i^{(j)}\right)$ к разным признакам;
- Модели аддитивные – т.е. можно изучать чувствительность ответа к отдельным входным признакам, просто зафиксировав все остальные.

ПРОТИВ

- Это аддитивные модели: учет взаимодействия между признаками ведется только на уровне вычисления линейной функции $g^{-1}(\cdot)$;
- Можно предоставить новые признаки, учитывающие взаимодействие имеющихся, но это не делает сама модель GAM => такие признаки не «обучаемые».

Есть ли способ еще увеличить выразительную способность функциональных параметрических моделей, оставаясь в рамках подхода моделей, обучаемых градиентными методами?

$$\mu = f(\theta, x) \qquad \frac{\partial \mathcal{L}}{\partial \theta} = \frac{\partial \mathcal{L}}{\partial \mu} \frac{\partial \mu}{\partial \theta}$$

$$p = f(\theta, x) \qquad \frac{\partial p}{\partial \theta}$$

$$p^{(k)} = f(\theta, x) \qquad \frac{\partial p^{(k)}}{\partial \theta}$$