



# Машинное обучение в науках о Земле

Михаил Криницкий

К.Т.Н.

Зав. лабораторией машинного обучения в науках о Земле МФТИ  
с.н.с. Института океанологии РАН им. П.П. Ширшова

# ПЛАН ЛЕКЦИИ

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

## Линейная регрессия

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

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

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

## 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

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

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

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

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

## 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

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

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

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

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

## 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

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

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

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

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

## Мультиномиальная логистическая регрессия

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

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

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

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

## 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

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

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

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

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

## Мультиномиальная логистическая регрессия

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

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

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

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

**Обобщенные линейные модели:** модели, в которых некоторая функция  $g(\cdot)$  математического 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

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

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

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

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

## Мультиномиальная логистическая регрессия

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

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

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

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

**Обобщенные линейные модели:** модели, в которых некоторая функция  $g(\cdot)$  математического 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

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

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

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

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

## Мультиномиальная логистическая регрессия

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

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

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

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

Альтернативно (некорректно, но легче понять):

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

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

**Как зависит вид функции потерь от вида функции связи?**

# Как зависит вид функции потерь от вида функции связи?

Линейная регрессия

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

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

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

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

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

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

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

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

# Как зависит вид функции потерь от вида функции связи?

Логистическая регрессия

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

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

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

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

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

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

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

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

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

# Как зависит вид функции потерь от вида функции связи?

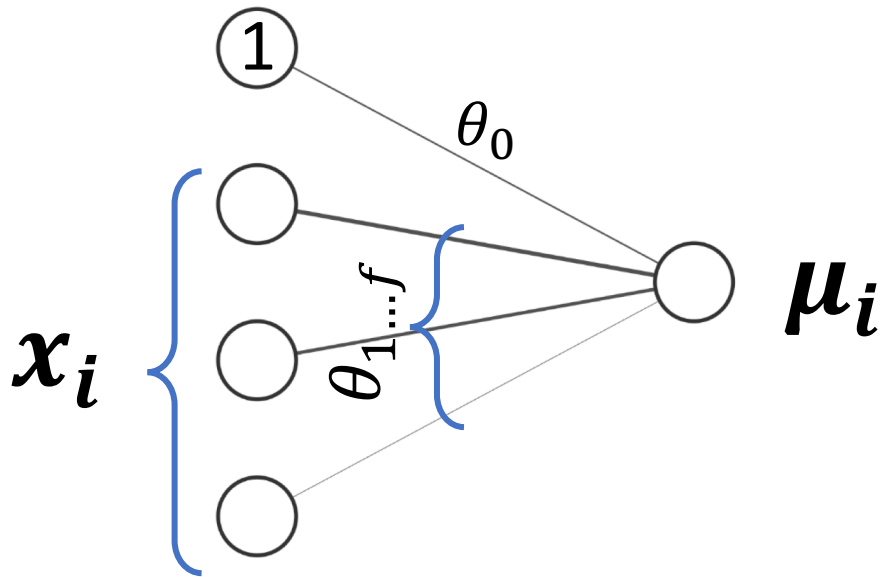
**НИКАК**

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

- обратная функция связи  $g^{-1}(\cdot)$  должна отображать  $R^1$  (множество значений произвольной линейной функции  $\theta \cdot x_i$ ) на множество параметров распределения переменной  $y_i$
- вычисление параметра (параметров) распределения переменной  $y_i$  в модели производится согласно принципу GLM
- при таких условиях вычисление правдоподобия выборки  $\mathcal{T}$  производится независимо от вида функции связи  $g(\cdot) \Rightarrow$  вычисление функции потерь в подходе максимизации правдоподобия также производится независимо от вида функции связи  $g(\cdot)$

# Диаграммы GLM

Линейная регрессия



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

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

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

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

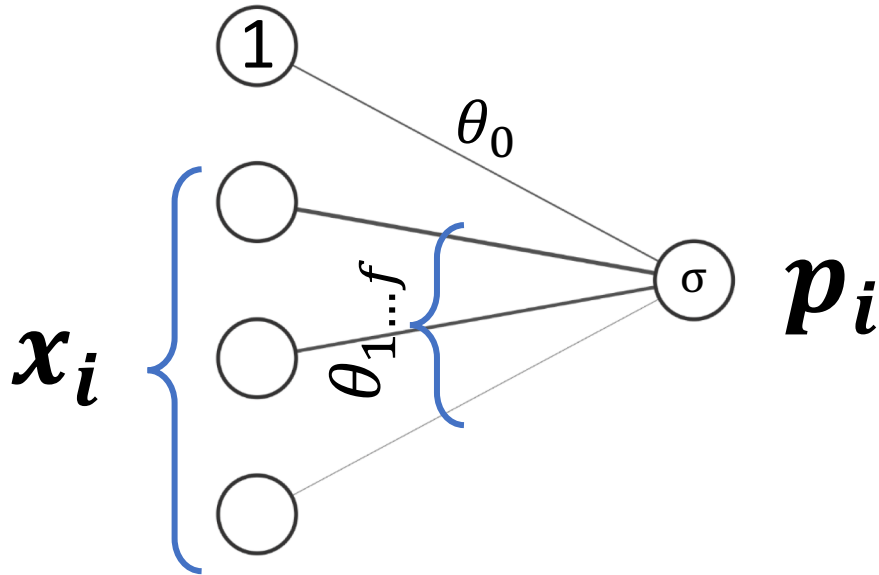
$$\theta = \underset{\Theta}{\operatorname{argmin}}(\mathcal{L}(\mathcal{T}, \theta))$$

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

=> градиентная оптимизация

# Диаграммы GLM

Логистическая регрессия



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

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

$$p(\theta_1, x_i) = \sigma(\eta_i^1)$$

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

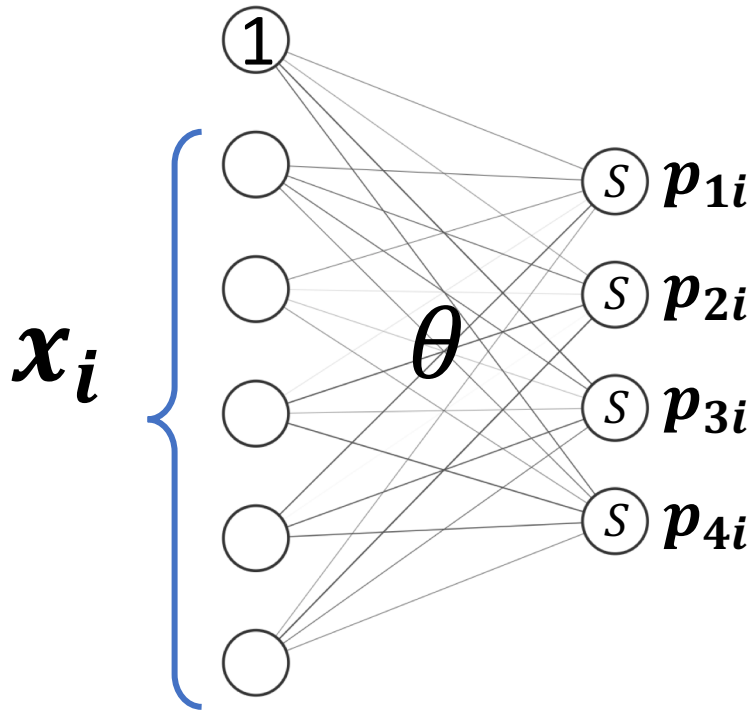
$$\theta = \underset{\Theta}{\operatorname{argmin}}(\mathcal{L}(\mathcal{T}, \theta))$$

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

=> градиентная оптимизация

# Диаграммы GLM

Мультиномиальная  
логистическая регрессия



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

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

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

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

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

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

=> градиентная оптимизация



## Регрессия

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

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

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

## Бинарная классификация

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

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

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

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

## Мультиномиальная классификация

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

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

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

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

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

## Регрессия

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

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

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

## Бинарная классификация

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

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

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

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

## Мультиномиальная классификация

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

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

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

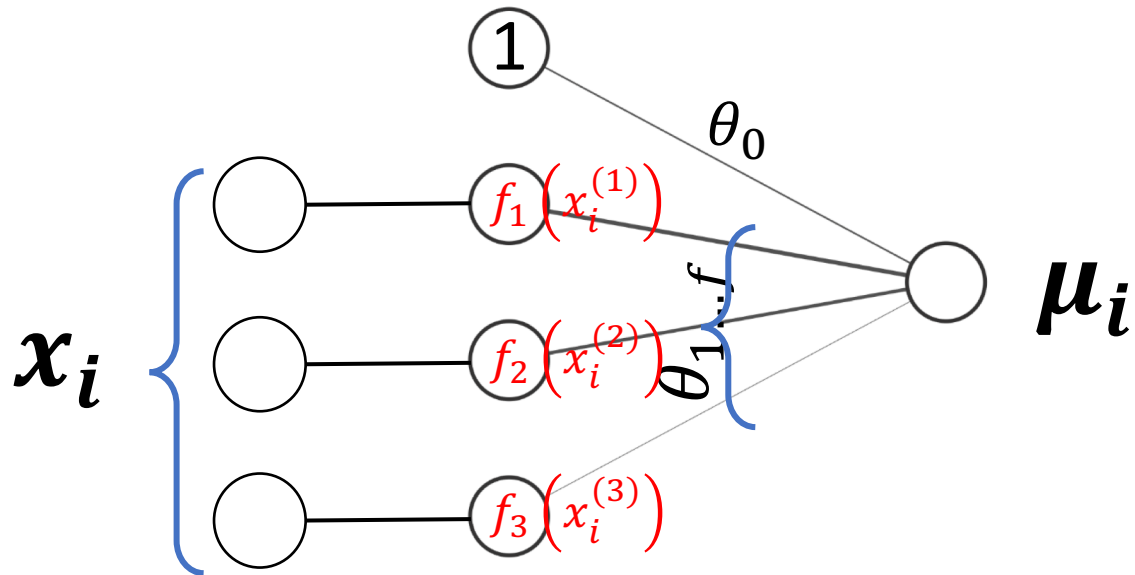
$$p(\theta_1, x_i) = \text{softmax}(\eta_i^k, \{\eta_i^k\}) = \frac{\exp \eta_i^k}{C^*}$$

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



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

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

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

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

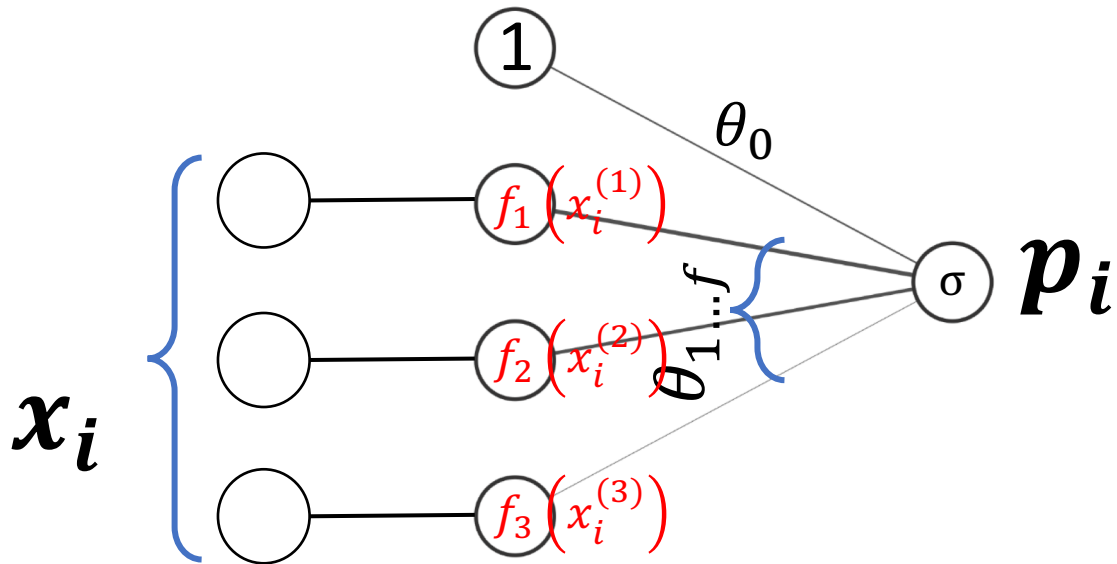
$$\theta = \underset{\Theta}{\operatorname{argmin}} (\mathcal{L}(\mathcal{T}, \theta))$$

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

=> градиентная оптимизация

# Диаграммы GAM

Бинарная классификация



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

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

$$p(\theta_1, x_i) = \sigma(\eta_i^1)$$

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

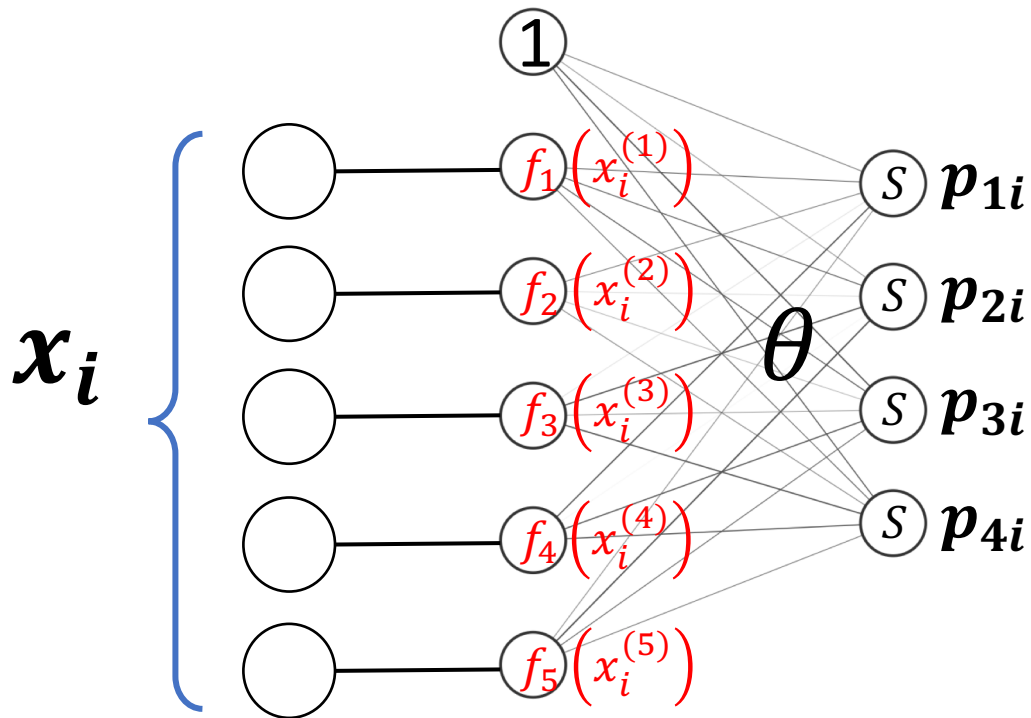
$$\theta = \underset{\Theta}{\operatorname{argmin}}(\mathcal{L}(\mathcal{T}, \theta))$$

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

=> градиентная оптимизация

# Диаграммы GAM

Мультиномиальная  
классификация



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

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

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

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

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

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

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

=> градиентная оптимизация

# Обобщенные аддитивные модели: за и против

## ЗА

- Довольно простые, но при этом предоставляют достаточно свободы в выборе нелинейных преобразований  $f_j(x_i^{(j)})$  исходных признаков объектов (событий);
- Подбирать функции  $f_j$  - иногда проще, чем подбирать степени полинома;
- Позволяют применять несколько разных нелинейных функций  $f_j(x_i^{(j)})$  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?