



Машинное обучение в науках о Земле

Михаил Криницкий

krinitsky.ma@phystech.edu

К.Т.Н., С.Н.С.

Институт океанологии РАН им. П.П. Ширшова

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

Общий принцип обучения по прецедентам (оптимизация функции ошибки)

$x \in \mathbb{X}$ — объекты, objects

$y \in \mathbb{Y}$ — ответы, labels

$\mathcal{F}: \mathbb{X} \rightarrow \mathbb{Y}$ — искомая закономерность

$\mathcal{T}: \{x_i; y_i\}$ — «обучающая выборка»
(прецеденты), train dataset

Найти: $\hat{\mathcal{F}}: \{x_i\} \rightarrow \{y_i\}$

один из способов решения:

$\mathcal{L}(\hat{\mathcal{F}}(x))$ — функционал ошибки
(эмпирического риска, потерь), Loss function

$\hat{y}_i = \hat{\mathcal{F}}(x_i) = f(\vec{p}, x_i)$ — функционально
задаваемая зависимость. **Предположение**
исследователя о виде закономерности. Иногда
задается параметрически, \vec{p} — вектор параметров.

$\mathcal{L} = L(\vec{p}, \mathcal{T})$ — функция ошибки

$$\hat{p} = \underset{\mathbb{P}}{\operatorname{argmin}}(L(\vec{p}, \mathcal{T}))$$

$$\hat{\mathcal{F}} = f(\hat{p}, x)$$

Обучение по прецедентам: вероятностная постановка

принцип максимального правдоподобия maximum likelihood estimation

\mathbf{x}_i - признаковое описание объектов
 \mathbf{y}_i - признаковое описание ответов
 $p(\mathbf{x}, \mathbf{y})$ – (искомая, аппроксимируемая)
совместная плотность распределения
событий на множестве $X \times Y$
 $\phi(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$ - модель плотности
распределения, предлагаемая
исследователем

$\mathcal{T}: \{\mathbf{x}_i; \mathbf{y}_i\}$ — «обучающая выборка»
(прецеденты), train dataset

Предположение!

$(\mathbf{x}_i, \mathbf{y}_i)$ – выбираются из $p(\mathbf{x}, \mathbf{y})$
независимо и случайно

MLE

$\phi(\mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\theta})$ - правдоподобие для одного экземпляра выборки

$L(\{\mathbf{x}_i\}, \{\mathbf{y}_i\}, \boldsymbol{\theta}) = \prod_{i=1}^N \phi(\mathbf{x}_i, \mathbf{y}_i, \boldsymbol{\theta})$ - правдоподобие выборки

$$\boldsymbol{\theta}^* = \underset{\Theta}{\operatorname{argmax}} L(\{\mathbf{x}_i\}, \{\mathbf{y}_i\}, \boldsymbol{\theta})$$

Функция потерь определяется видом модели плотности
распределения $\phi(\mathbf{x}, \mathbf{y}, \boldsymbol{\theta})$, предложенной исследователем!

Правдоподобие выборки $L(\boldsymbol{\theta}, \mathcal{T})$ – **максимизировать** (в
пространстве параметров Θ)

Функцию потерь $\mathcal{L}(\boldsymbol{\theta}, \mathcal{T})$ – **минимизировать** (в пространстве
параметров Θ)

Обучение по прецедентам. Вероятностная постановка, MLE

Примеры

Линейная регрессия

MSE

$$\phi(x, y, \theta) = \theta x + \epsilon,$$

$$p(\epsilon) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{\epsilon^2}{2\sigma^2}}$$

$$\Rightarrow \mathcal{L}(\theta, \mathcal{T}) = \frac{1}{N} \sum_{i=1}^N (y_i - \theta x_i)^2$$

MAE

$$\phi(x, y, \theta) = \theta x + \epsilon,$$

$$p(\epsilon) = \frac{1}{2b} e^{-\frac{|\epsilon|}{b}}$$

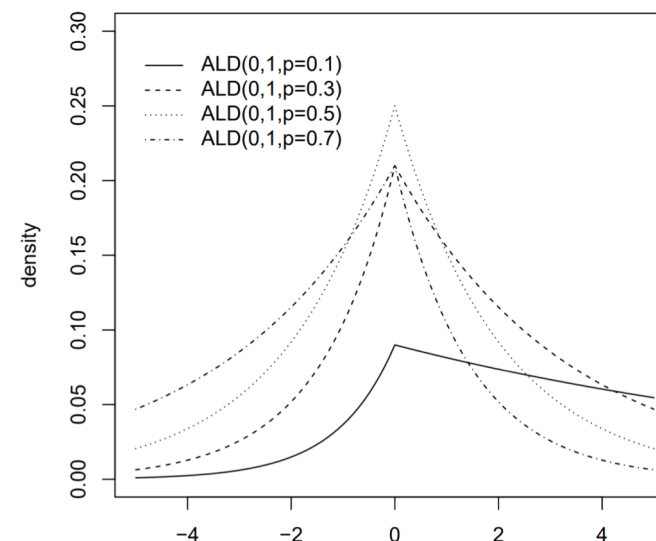
$$\Rightarrow \mathcal{L}(\theta, \mathcal{T}) = \frac{1}{N} \sum_{i=1}^N |y_i - \theta x_i|$$

previously on ML4ES

Квантильная регрессия

Bera, Anil & Galvao, Antonio & Montes-Rojas, Gabriel & Park, Sung Y.. (2015). **Asymmetric Laplace Regression: Maximum Likelihood, Maximum Entropy and Quantile Regression**. Journal of Econometric Methods. 10.1515/jem-2014-0018.

Sánchez, B. L., Lachos, H. V., & Labra, V. F. (2013). **Likelihood based inference for quantile regression using the asymmetric Laplace distribution**. Journal of Statistical Computation and Simulation, 81, 1565-1578.



Asymmetric Laplace density. From Sánchez et al.



Решение задач типа ОБУЧЕНИЕ С УЧИТЕЛЕМ

Михаил Криницкий

krinitsky.ma@phystech.edu

К.Т.Н., Н.С.

Институт океанологии РАН им. П.П. Ширшова

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

ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

1. формулировка задачи:

- какой тип (классификация, регрессия, другой)? Или переформулировать в легко решаемый тип!
- определиться, что есть объекты (события)
- определиться, что есть целевая переменная
- определить признаковое описание объектов (событий)
- определить критерии качества решения задачи (MSE, MAE, pattern correlation, etc.)

2. сформулировать модель:

- вид модели (линейная регрессия, дерево решений, композиционный алгоритм, нейронная сеть, etc.)
- определиться с функцией потерь (MSE, MAE, BCE, CCE, etc., комбинации)
- сложность модели (задается гиперпараметрами – настройками модели)

ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

3. подготовить данные или генератор данных:
 - стандартизировать данные (если нужно)
 - обработка пропусков, категориальных значений, кодирование текста, понижение размерности данных
 - оставить часть данных для проверки качества (train-validation-test split)
 - подготовить генератор данных с учетом стратегии скользящего контроля (cross-validation quality estimation)
4. оптимизировать модель на обучающей выборке:
 - $\hat{p} = \underset{\mathbb{P}}{\operatorname{argmin}}(L(\vec{p}, \mathcal{T}))$
5. оптимизация гиперпараметров модели и отбор моделей. Провизодится по значениям метрик качества на контрольной(контрольных) выборке(выборках)
6. оценка модели:
 - оценить качество по метрикам, определенным на этапе 1. на тестовой выборке
 - оценить неопределенность параметров модели (если возможно)
 - оценить неопределенность оценок целевой переменной

ОБЩАЯ СХЕМА РЕШЕНИЯ ЗАДАЧ ОБУЧЕНИЯ С УЧИТЕЛЕМ

6. применение модели на вновь получаемых данных:
 - оценка распределения вновь получаемых данных: генерируются ли они из того же распределения, что и обучающая выборка?
 - предобработка новых данных идентично п.3 с точностью до коэффициентов стандартизации и деталей способов предобработки
 - применение модели к предобработанным новым данным для получения значений целевой переменной
 - построение научных выводов, описание их в виде статей, получение наград в виде Нобелевских премий, etc.