Мультимоделирование как универсальный способ описания выборки общего вида

Роман Алексеевич Логинов

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов), Группа 694, весна 2019
Консультант: А. А. Адуенко

Цель исследования

Задача

Определить неоднородность в данных, построить оптимальную систему моделей (мультимодель), которая корректно опишет выборку. Учесть временную зависимость в выборке

Требования к каждой модели

- Статистически отличима от остальных
- Обучена на соответсвующей ей подвыборке
- Параметры модели изменяются во времени

Исходная задача машинного обучения: двухклассовая классификация

Теоретическая база

Ближайшая работа

Адуенко А.А.
Выбор мультимоделей в задачах классификации, МФТИ, 2017

- Алгоритм построения мультимоделей в частных случаях
- Количественная оценка различимости моделей
- Отсутствие временной структуры

Теоретические сведения

Bishop Pattern recognition and machine learning
Motrenko, Strijov Sample size determination for logistic regression
Ge, Jiang On Consistency of Bayesian Inference with Mixtures of logistic regression

Обозначения: одиночный классификатор

Дано:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, ..., m\}$$

- обучающая выборка
$\mathbf{x}_i \in \mathbb{X} \subset \mathbb{R}^n$ - признаковое описание объекта
$y_i \in \{+1, -1\}$ - корректная метка класса

Вероятностный классификатор:

$$p(y, \mathbf{w}|\mathbf{x}, \mu) = \sigma(y\mathbf{w}^\top \mathbf{x})p(\mathbf{w}|\mu),$$

где $\mathbf{w}$ - параметр модели, получаемый при обучении, μ - гиперпараметры

Временная структура:

$$t(X_i) < t(X_j)\ \forall i < j,\ \ i, j \in \mathcal{I}$$

Мультимодель с временной структурой

Требуется построить

- $\mathbb{M} = \{f_i\}, i \in \mathfrak{M}$ - множество описанных классификаторов
- $\mathbf{w}_i(t), i \in \mathfrak{M}$ - параметры каждой модели в зависимости от времени
- $\mathcal{M} : \mathbb{X} \to \mathfrak{M}$ - правило выбора модели

Итоговый классификатор

Классификатор f_{mult} выбирает модель по правилу и вычисляет предсказание с текущим временем

$$\arg\max AUCROC(\mathfrak{D}, f_{mult}(\mathcal{M}, \mathbf{w}))$$
$$s.t.\ \mathcal{M} : \mathbb{X} \to \mathfrak{M}, \mathbf{w} = \mathbf{w}(t)$$

Данные для эксперимента

Многоуровневая модель

- Заранее заданы правила кластеризации на множестве признаков
- На каждом - определённые веса для генерации

Смесь моделей

- Генерируются k различных весов из $\mathcal{N}(0, diag(\sigma_{true}))$
- Из распределения Дирихле получается вектор π - верятность появления данных в соответствии с каждой моделью
- Из полученного дискретного распределения выбираются необходимые веса

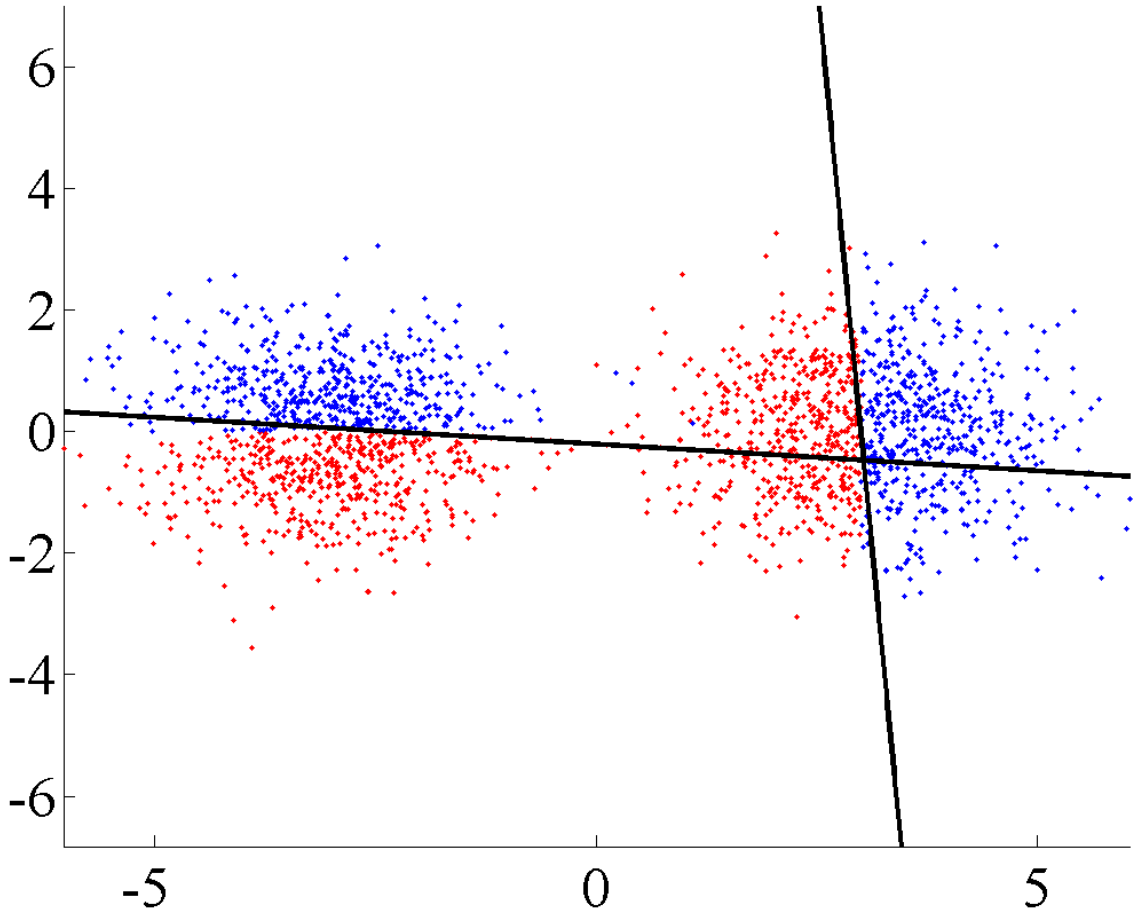
План эксперимента

Для описанных данных выяснить следующее:

- Как меняются предсказания модели во времени
- Какой размер выборки необходим для выделения новой различимой модели
- Какое качество у мультимодели в случае изменения во времени модели генерации
- Получить качество классификации на реальных данных

Базовый алгоритм

Простой случай мультимодели с очевидной неоднородностью



Заключение

Недостатки текущего алгоритма:

- Максимальное количество моделей зафиксировано
- Одинаковому признаковому описанию соответствует одинаковый ответ, без учёта времени