Дмитриев Егор Андреевич

студент

ФГАОУ ВО «Самарский национальный исследовательский университет им. академика С.П. Королева»

г. Самара, Самарская область

ЛИНЕЙНЫЕ КЛАССИФИКАТОРЫ

Аннотация: *в данной исследовательской работе автором рассматривается один из методов машинного обучения – использование метрических классификаторов.*

Ключевые слова: *линейные классификаторы, алгоритм классификации, машинное обучение.*

Основные определения

Определение 1. Линейный классификатор – алгоритм классификации, который основывается на применении разделяющей поверхности. Разделяющая поверхность называется гиперплоскостью, подпространство размерности n-1 пространства признаков.

Определение 2. Алгоритм классификации – отображение $a: X \to Y$, где X – метрическое пространство признаков классифицируемого объекта, а Y – множество классов. Классифицировать объект – однозначно определить, к какому классу относится объект.

Определение 3. Машинное обучение – нахождение отображения, в частности алгоритма классификации, который строится по множеству, называемому обучаемой выборкой, а качество обучения проверяется по множеству, называемому тестовой выборкой.

Введение

В последнее время к задачам машинного обучения, в частности к классификации объектов, привлекают все больше и больше внимания. В большинстве компаний, приходиться иметь дело с многочисленными данными, такие как: пер-

сональные данные сотрудников и клиентов, информация на крупных веб ресурсах, научные источники, наименования болезней, растений, животных и т. д. Для того, чтобы автоматизировать различные процессы, требуется использовать обучение системы для выполнения определенных задач.

Постановка проблемы

Необходимо ознакомиться с принципами работы линейных классификаторов объектов.

Описание линейных классификаторов

Суть задачи классификации заключается в построении алгоритма классификации по обучаемой выборке. Идея линейных классификаторов заключается в построении линейных функций, являющимися разделяющими поверхности, по обе стороны которой лежат объекты разных классов.

Формальное определение линейного классификатора:

Линейный классификатор – отображение $a: X \to Y$ вида:

$$a(x, w) = sign(\sum_{j=1}^{n} w_j x_j(x) - w_0) = sign(x, w),$$

где w_j – веса признаков, w_0 – порог.

Введем функционал, который называется эмпирическим риском:

$$Q(w, X^l) = \sum_{i=1}^{n} \varphi((w, x_i) y_i) \to min,$$

где величина $(w, x_i) y_i$ – величина отступа $M(x_i)$, а φ – функция, зависящая от отступа. Существует несколько видов таких функций:

1) $V(M) = (1-M)_+$-кусочно-линейная функция;

2) $L(M) = log_2(1 + e^{-M})_+$-логарифмическая функция;

3) $S(M) = 2(1 + e^M)^{-1}$-сигмоидная функция.

Градиентный спуск

Для минимизации эмпирического риска, чаще всего используется метод под названием градиентный спуск. Суть метода заключается в движении по направлению антиградиента, то есть направление, где происходит максимальное уменьшении целевой функции.

Градиентный метод состоит из нескольких шагов:

1)$w^{(0)} :=$ начальное приближение;

2)$w^{(t+1)} := w^{(t)} - \eta * \nabla Q(w^{(t)})$,

где $\nabla Q(w) = (\frac{\partial Q(w)}{\partial w_j})_{j=0}^{n}$ – градиент функции, а η – коэффициент скорости обучения, если подставить формула для функционала мы получим:

$$w^{(t+1)} := w^{(t)} - \eta * \sum_{i=1}^{n} \varphi'((w^{(t)}, x_i)y_i)x_i y_i.$$

В формуле для изменения весов видим, что для того, чтобы один раз изменить веса линейного классификатора, необходимо перебрать все элементы обучающей выборки. Для сходимости может понадобиться большое количество шагов и с учетом большого размера выборки, нахождение оптимального решения занимает достаточно большого количества времени. На практике используется метод стохастического градиента, которые позволяет достигать оптимального значения целевой функции без использования всей выборки, а лишь ее часть.

Достоинства стохастического градиента:

1) легко реализуется;

2) использование любого дифференцируемого функционала;

3) возможно использовать только часть выборки.

Недостатки:

1) возможна медленная сходимость или расходимость;

2) переобучение;

3) застревание в локальных минимумах.

Заключение

В данной работе был рассмотрен принцип работы логических классификаторов.

Список литературы

1. Айвазян С.А. Прикладная статистика: классификация и снижение размерности / С.А. Айвазян, В.М. Бухштабер, И.С. Енюков, Л.Д. Мешалкин. – М.: Финансы и статистика, 1989.

2. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: ИМ СО РАН, 1999.