Введение в искусственный интеллект. Машинное обучение

Лекция 2. Непараметрические методы классификации и регрессии

Бабин Д.Н., Иванов И.Е., Петюшко А.А.

МаТИС

22 февраля 2019г.

План лекции

1. Метод ближайших соседей в задаче классификации
2. Непараметрическая регрессия

Параметрические и непараметрические методы машинного обучения

Параметрические методы

- исходят из предположения, что искомая зависимость имеет некоторый специальный вид с точностью до некоторых параметров
- параметры находятся решением оптимизационной задачи

Параметрические и непараметрические методы машинного обучения

Параметрические методы

- исходят из предположения, что искомая зависимость имеет некоторый специальный вид с точностью до некоторых параметров
- параметры находятся решением оптимизационной задачи

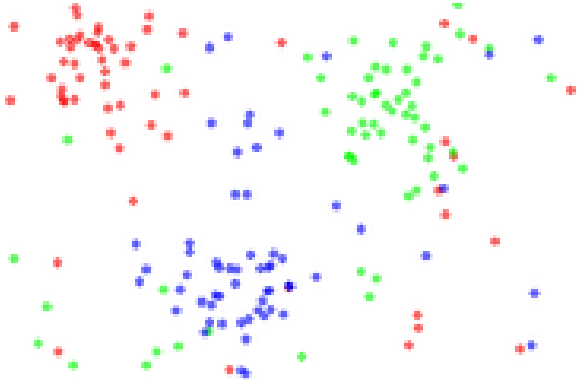
Непараметрические методы

Непараметрические методы – методы не являющиеся параметрическими

- Метрические алгоритмы, ядерные методы

Основное предположение

- "Близкие"объекты лежат в одном классе
- Близость задаётся метрикой
- Типичный пример [1]



[1] https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm

Метод ближайшего соседа

- Параметр метода: метрика
- Алгоритм: по заданной метрике ищем ближайший объект в обучающей выборке и классифицируем объект так же

Метод ближайшего соседа

- Параметр метода: метрика
- Алгоритм: по заданной метрике ищем ближайший объект в обучающей выборке и классифицируем объект так же

Преимущества

- Простота реализации (нет как таковой процедуры обучения в наивной реализации)
- Хорошая интерпретируемость

Метод ближайшего соседа

- Параметр метода: метрика
- Алгоритм: по заданной метрике ищем ближайший объект в обучающей выборке и классифицируем объект так же

Преимущества

- Простота реализации (нет как таковой процедуры обучения в наивной реализации)
- Хорошая интерпретируемость

Недостатки

- Неустойчивость к выбросам
- Неоднозначность классификации при равных расстояниях до двух объектов
- Необходимость хранить всю обучающую выборку
- Алгоритм поиска вычислительно сложен (если обучающая выборка довольно большая)
- Не учитывается значение расстояния

Метод k ближайших соседей

- Параметр метода: метрика, **k**
- Алгоритм: по заданной метрике ищем k ближайших объектов в обучающей выборке и классифицируем объект как большинство из k объектов

Преимущества

- Простота реализации
- Хорошая интерпретируемость
- Параметр k можно оптимизировать по скользящему контролю

Недостатки

- ~~Неустойчивость к выбросам~~
- ~~Неоднозначность классификации при равных расстояниях до двух объектов~~
- Необходимость хранить всю обучающую выборку
- Алгоритм поиска вычислительно сложен (если обучающая выборка довольно большая)
- Не учитывается значение расстояния

Метод k ближайших взвешенных соседей

- Параметры метода: метрика, k, **веса**
- Алгоритм: по заданной метрике ищем k ближайших объектов в обучающей выборке и классифицируем объект взвешенным голосованием

Преимущества

- Простота реализации
- Хорошая интерпретируемость
- Параметр k можно оптимизировать по скользящему контролю

Недостатки

- ~~Неустойчивость к выбросам~~
- ~~Неоднозначность классификации при равных расстояниях до двух объектов~~
- Необходимость хранить всю обучающую выборку
- Алгоритм поиска вычислительно сложен (если обучающая выборка довольно большая)
- ~~Не учитывается значение расстояния~~

Метод k ближайших взвешенных соседей: выбор весов

- Веса в зависимости от порядкового номера

Метод k ближайших взвешенных соседей: выбор весов

- Веса в зависимости от порядкового номера
  - Линейно убывающие веса

Метод k ближайших взвешенных соседей: выбор весов

- Веса в зависимости от порядкового номера
  - Линейно убывающие веса
  - Экспоненциально убывающие веса

Метод k ближайших взвешенных соседей: выбор весов

- Веса в зависимости от порядкового номера
  - Линейно убывающие веса
  - Экспоненциально убывающие веса
  - Любая невозрастающая функция от порядкового номера

Метод k ближайших взвешенных соседей: выбор весов

- Веса в зависимости от порядкового номера
  - Линейно убывающие веса
  - Экспоненциально убывающие веса
  - Любая невозрастающая функция от порядкового номера
- Веса в зависимости от расстояния

Метод k ближайших взвешенных соседей: выбор весов

- Веса в зависимости от порядкового номера
  - Линейно убывающие веса
  - Экспоненциально убывающие веса
  - Любая невозрастающая функция от порядкового номера
- Веса в зависимости от расстояния
  - Любая невозрастающая функция от расстояния

Метод k ближайших взвешенных соседей: выбор весов

- Веса в зависимости от порядкового номера
  - Линейно убывающие веса
  - Экспоненциально убывающие веса
  - Любая невозрастающая функция от порядкового номера
- Веса в зависимости от расстояния
  - Любая невозрастающая функция от расстояния
- Фиксированные веса объектов

Метод k ближайших взвешенных соседей среди набора эталонов

- Параметры метода: метрика, k, веса, **метод выбора эталонов**
- Алгоритм: по заданной метрике ищем k ближайших объектов среди эталонов выбранных из обучающей выборки и классифицируем объект взвешенным голосованием

Преимущества

- Простота реализации
- Хорошая интерпретируемость
- Параметр k можно оптимизировать по скользящему контролю

Недостатки

- ~~Неустойчивость к выбросам~~
- ~~Неоднозначность классификации при равных расстояниях до двух объектов~~
- ~~Необходимость хранить всю обучающую выборку~~
- ~~Алгоритм поиска вычислительно сложен~~
- ~~Не учитывается значение расстояния~~

Выбор эталонов

Задача

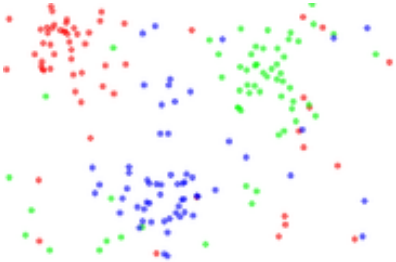
Получить примерно такое же качество работы алгоритма при меньшем количестве хранимых данных.
Возможно получить улучшение качества, так как в процессе выбора эталонов будут удалены выбросы.

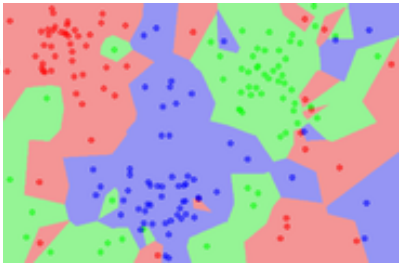
Идеи

- Кластеризация объектов
- Жадный алгоритм
- Все элементы обучающего множества можно ранжировать по количеству вхождений в k ближайших соседей

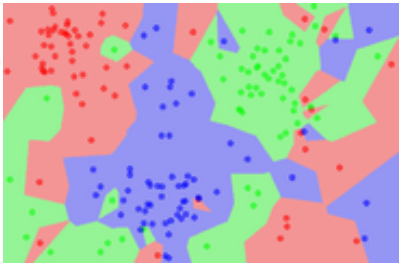
Примеры 1-nn, 5-nn, 1-nn с выбором эталонов

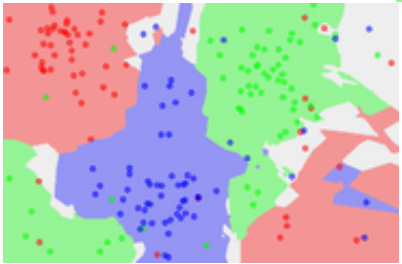
Примеры 1-nn, 5-nn, 1-nn с выбором эталонов



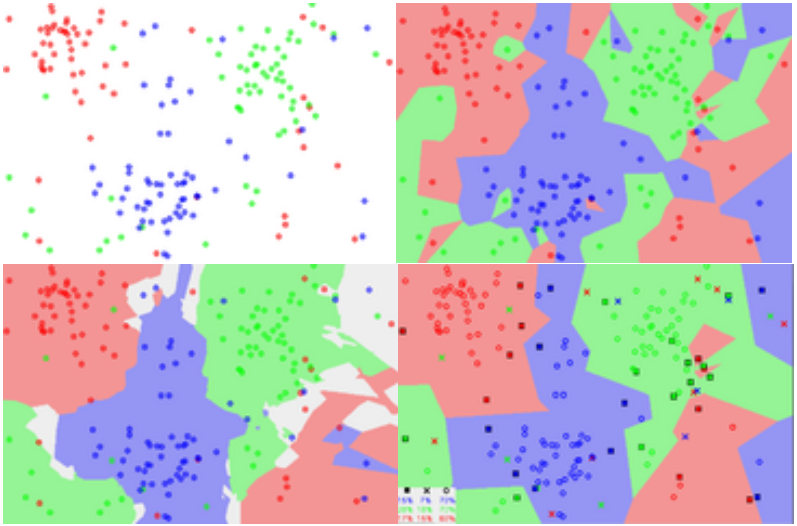


Примеры 1-nn, 5-nn, 1-nn с выбором эталонов





Примеры 1-nn, 5-nn, 1-nn с выбором эталонов



Реализация метода в scikit-learn

class sklearn.neighbors.KNeighborsClassifier(n_neighbors=5, weights='uniform', algorithm='auto', leaf_size=30, p=2, metric='minkowski', metric_params=None, n_jobs=None, **kwargs)

Основные параметры

- **n_neighbors** : int, optional (default = 5)
  Number of neighbors to use by default for kneighbors queries.
- **weights** : str or callable, optional (default = 'uniform')
  weight function used in prediction. Possible values:
  'uniform' : uniform weights 'distance' : weight points by the inverse of their distance.
  [callable] : a user-defined function which accepts an array of distances, and returns an array of the same shape containing the weights.
- **metric** : string or callable, default 'minkowski'
- **n_jobs** : int or None, optional (default=None)
  The number of parallel jobs to run for neighbors search

Заключение

- Метод ближайших соседей – простой и хорошо интерпретируемый метод классификации

Заключение

- Метод ближайших соседей – простой и хорошо интерпретируемый метод классификации
- Метод имеет большое число вариаций для настройки

Заключение

- Метод ближайших соседей – простой и хорошо интерпретируемый метод классификации
- Метод имеет большое число вариаций для настройки
  - Подбор метрики (metric learning)

Заключение

- Метод ближайших соседей – простой и хорошо интерпретируемый метод классификации
- Метод имеет большое число вариаций для настройки
  - Подбор метрики (metric learning)
  - Число ближайших соседей

Заключение

- Метод ближайших соседей – простой и хорошо интерпретируемый метод классификации
- Метод имеет большое число вариаций для настройки
  - Подбор метрики (metric learning)
  - Число ближайших соседей
  - Веса во взвешенном варианте метода

Заключение

- Метод ближайших соседей – простой и хорошо интерпретируемый метод классификации
- Метод имеет большое число вариаций для настройки
  - Подбор метрики (metric learning)
  - Число ближайших соседей
  - Веса во взвешенном варианте метода
  - Алгоритм подбора эталонов

Непараметрическая регрессия

- Главный минус параметрических моделей, что для описания зависимости необходимо иметь параметрическую модель

Непараметрическая регрессия

- Главный минус параметрических моделей, что для описания зависимости необходимо иметь параметрическую модель
- В случае невозможности подбора адекватной модели имеет смысл пользоваться непараметрическими регрессионными методами

Непараметрическая регрессия

- Главный минус параметрических моделей, что для описания зависимости необходимо иметь параметрическую модель
- В случае невозможности подбора адекватной модели имеет смысл пользоваться непараметрическими регрессионными методами

Предположение

Близким объектам соответствуют близкие ответы

Непараметрическая регрессия

Простейшая модель

Приближаем искомую зависимость константой в некоторой окрестности

Формула Надарая-Ватсона

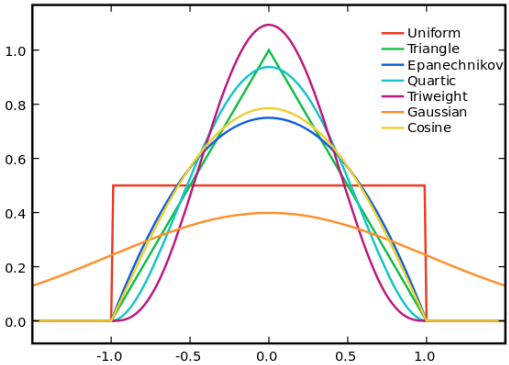
Если в окрестности точки несколько объектов из обучающей выборку, то разумно использовать взвешенное среднее в качестве предсказания алгоритма

$$a(x) = \frac{\sum\limits_i y_i \omega_i(x)}{\sum\limits_i \omega_i(x)},$$

где $\omega_i(x) = K_h(x, x_i)$, а функция K_h называется ядром с шириной окна сглаживания h.

Примеры ядер

- $K_h(x, x_i) = K(\frac{||x-x_i||}{h})$
- Типичные примеры [2]



[2] https://ru.wikipedia.org/wiki/Ядро_(статистика)

Bias-variance разложение в простейшем случае

$$E(a(x) - f(x))^2 = \left(f(x) - \frac{1}{k}\sum_{i=1}^{k} f(x_{(i)})\right)^2 + \frac{\sigma^2}{k} + \sigma^2$$

- С ростом k разброс уменьшается
- А сдвиг увеличивается
- С ростом n сдвиг уменьшается

Заключение

- Главное преимущество непараметрической регрессии — это отсутствие предположений о виде модели зависимости

Заключение

- Главное преимущество непараметрической регрессии — это отсутствие предположений о виде модели зависимости
- Метод имеет большое число вариаций для настройки

Заключение

- Главное преимущество непараметрической регрессии — это отсутствие предположений о виде модели зависимости
- Метод имеет большое число вариаций для настройки
  - Подбор метрики (metric learning)

Заключение

- Главное преимущество непараметрической регрессии — это отсутствие предположений о виде модели зависимости
- Метод имеет большое число вариаций для настройки
  - Подбор метрики (metric learning)
  - Число ближайших соседей

Заключение

- Главное преимущество непараметрической регрессии — это отсутствие предположений о виде модели зависимости
- Метод имеет большое число вариаций для настройки
  - Подбор метрики (metric learning)
  - Число ближайших соседей
  - Веса во взвешенном варианте метода

Заключение

- Главное преимущество непараметрической регрессии — это отсутствие предположений о виде модели зависимости
- Метод имеет большое число вариаций для настройки
  - Подбор метрики (metric learning)
  - Число ближайших соседей
  - Веса во взвешенном варианте метода
  - Ширину окна сглаживания

Спасибо за внимание!