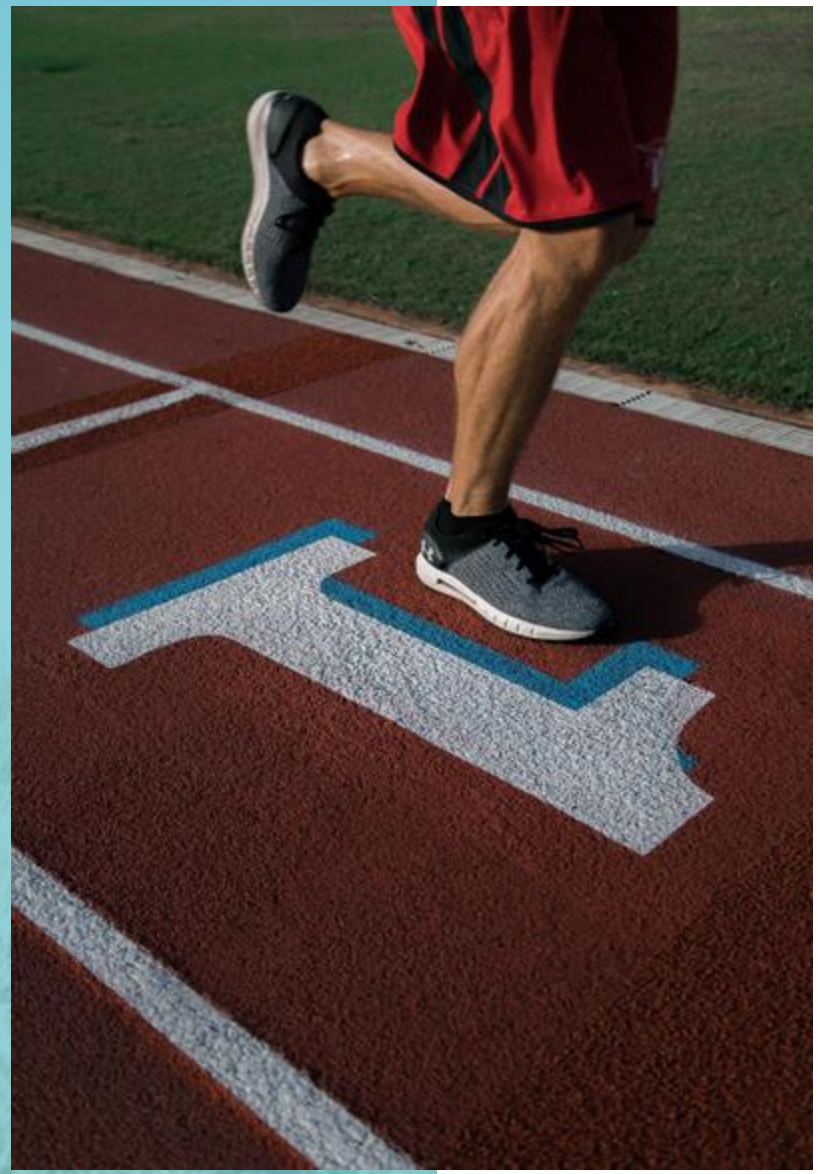
PAMAP2 Physical Activity Monitoring Data Set



Общая информация

Данные были взяты с репозитория машинного обучения UC Irvine (ссылка на данные будет предоставлена в конце презентации).

Набор данных мониторинга физической активности PAMAP2 содержит данные о 18 различных физических нагрузках, выполняемых 9 субъектами с 3 инерционными измерительными устройствами и монитором сердечного ритма.

Характеристики dataset’a

Характеристики набора данных: многомерный, временной ряд.

Характеристики атрибута: real

Связанные задачи: классификация

Количество экземпляров: 3850505

Количество атрибутов: 52

Тип данных: float

Дополнительно

Некоторые данные* отсутствуют по следующим причинам:

1. Потеря данных из-за беспроводных датчиков
2. Проблемы с настройкой оборудование (могут вызвать, например, потерю соединения или сбой оборудования)

*Такие данные помечены как Nan

Задача

Провести классификацию по столбцу activityID (2-й из 54 столбцов в списке данных).

Практическое применение

Пример: определение вида нагрузки на умном браслете (бег, подъем по лестнице, плавание) и вывод сжигаемых калорий исходя из определенного ранее типа.

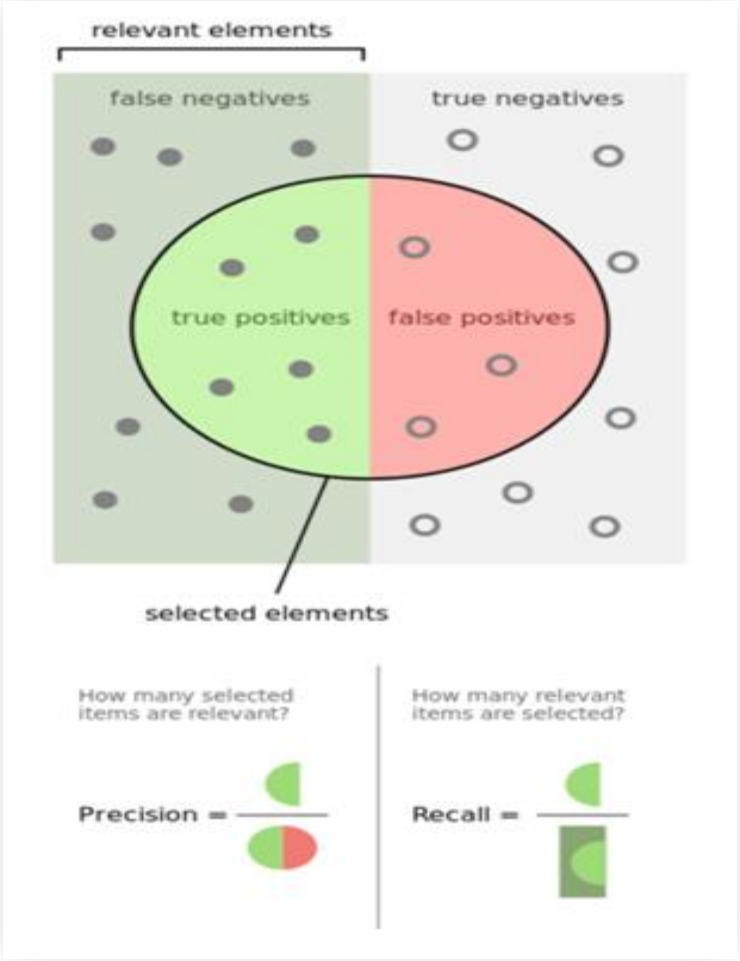
Предпочитаемый алгоритм

Деревья решений - один из методов обучения с учителем, используемый для классификации и регрессии. Задача классификации при данном алгоритме состоит в том, чтобы объекты одного класса были сгруппированы вместе

- Простой пример из жизни

Метрики оценивания качества модели

- Accuracy
- Precision (точность)
- Recall (полнота)
- F-мера - среднее гармоническое precision и recall



Преимущества и недостатки метода

1. Формируют четкие и понятные правила классификации
2. Быстро обучаются и прогнозируют
3. Не требуется много параметров модели
4. Способны генерировать правила в областях, где специалисту трудно формализовать свои знания

1. Возможно переобучение дерева решений, из-за чего приходится прибегать к методу «отсечения ветвей» и пр.
2. Деревья решений чувствительны к шумам во входных данных
3. Сложный поиск оптимального дерева решений

Обоснование выбора метода и примерно ожидаемый результат

Метод часто является предпочтительным при решении задач классификации, и этот случай не исключение.

Благодаря нему, после обучения на основе имеющихся данных, возможно предсказывать, каким видом спорта на данный момент занимается человек (по классам, установленным в activityID), на котором надет умный браслет/иное устройство, способное собирать подобные данные, позволяющие, например, предсказать оптимальное время для занятий в течение дня.

Способы решения проблемы пропущенных значений

Существуют множество разных способов борьбы с пропущенными значениями (пз) в классах. Приведем в пример некоторые из тех, которые не удаляют/игнорируют пз:

- Заполнение нулями/медианой/средним арифметическим значением (ad-hoc методы)
- Метод HOT DECK (представляет собой подстановку вместо пропуска значения по данной переменной у наиболее близкого объекта с полной информацией)
- RESAMPLING (строки, содержащие пропущенные данные заменяют случайно подобранными строками из матрицы полных наблюдений)
- Одни из самых часто используемых в практике - EM-оценивание и регрессионное моделирование пропусков

Основные этапы построения дерева

1. Выбор атрибута, по которому будет производиться разбиение в данном узле.
2. Выбор критерия остановки обучения
3. Выбор метода отсечения ветвей (упрощение)
4. Оценка точности построенного дерева

Критерии разбиения атрибута

Теоретико-информационный критерий: как следует из названия, критерий основан на понятиях теории информации, а именно — информационной энтропии.

Показатель энтропии может быть от 0 до 1, где 1 - наибольшая неопределенность классификации, 0 - наименьшая

Статистический подход: в его основе лежит использование индекса Джини (назван в честь итальянского статистика и экономиста Коррадо Джини). Статистический смысл данного показателя в том, что он показывает — насколько часто случайно выбранный пример обучающего множества будет распознан неправильно, при условии, что целевые значения в этом множестве были взяты из определенного статистического распределения. Данный показатель также имеет значения от 0 до 1, где 1 - все примеры относятся к одному классу, 0 - классы равновероятны

Полезные ссылки

- https://archive.ics.uci.edu/ml/datasets/PAMAP2+Physical+Activity+Monitoring - ссылка на архив данных
- https://habr.com/ru/company/ods/blog/328372/ - источник данных по метрикам