Использование глубинного обучения для классификации музыкальных жанров.

Под глубинным обучением понимается использование моделей машинного обучения, включающих в себя нелинейные преобразования, а также число этих преобразований больше заданного, например, в некоторых источниках к моделям глубинного обучения относят нейронные сети с 2 и более скрытыми слоями. Также признаки, извлекаемые в процессе работы с данными моделями, зачастую являются иерархическими (более подробно в [Википедии](https://ru.wikipedia.org/wiki/%D0%93%D0%BB%D1%83%D0%B1%D0%BE%D0%BA%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5)).

Для предобучения нейронной сети с целью выявить закономерности в неразмеченных данных может использоваться автокодировщик или ограниченная машина Больцмана (RBM) ([тут](http://www.machinelearning.ru/wiki/images/5/59/MMP_Praktikum_317_2015s_5.pdf) можно прочитать про автокодировщик, а [тут](https://habrahabr.ru/post/163819/) – про RBM).

Суть автокодировщика: имеется n входов xi, скрытый слой, и n выходов x`i. Задача – аппроксимировать тождественную функцию h(x) = x, минимизируя функционал ошибки L(x, x`, w, b), выражающий «отличие» вектора x` от вектора x. Существенные ограничения, вводимые при использовании автокодировщика – размер скрытого слоя должен быть строго меньше n, иначе сеть построит прямое отображение. Другой способ заставить сеть «сжимать» входные данные и искать в них закономерности для более достоверного восстановления – разреженный скрытый слой (ограничение на количество одновременно используемых нейронов в скрытом слое).

Суть ограниченной машины Больцмана (RBM): ограниченная машина Больцмана представляет собой двудольный граф, описывающий состояния скрытых и видимых нейронов, а граф двудольный потому, что предполагается независимость состояний нейронов внутри одной группы (скрытые и видимые). Для RBM введено понятие энергии, и процесс обучения (оптимизация правдоподобия) вводится в аналог нахождению равновесия системы. Лень писать, [тут](https://habrahabr.ru/post/159909/) ещё про RBM)

Предобработка данных

[Тут](https://habrahabr.ru/post/263811/) описан процесс возможной предобработки данных. Основная суть предобработки аудиоданных:

1. Для большинства композиций целесообразно брать 30-секундный отрывок (не из начала трека!), а не весь трек, что заметно снижает объём данных
2. Все признаки для аудиоданных можно условно разделить на 2 группы: интерпретируемые человеком и неинтерпретируемые (т.е. различает ли человек музыку по этим признакам)
3. [Есть данные](https://infoscience.epfl.ch/record/87348/files/Scaringella2006_1436.pdf), что гармония и мелодия более надёжно описываются низкоуровневыми атрибутами, нежели нотами или аккордами
4. Один из методов выбора признаков – MMFC (можно почитать [тут](https://habrahabr.ru/post/140828/))

В [одной](https://courses.engr.illinois.edu/ece544na/fa2014/Tao_Feng.pdf) из работ по классификации музыки по жанрам с использованием глубинного обучения (без свёрточных и рекуррентных архитектур) делалось следующее:

1. Используемые данные: GTZAN Genre Collection, всего 1000 треков по 30 секунд каждый (+ метаданные). Представлено 10 жанров, по 100 треков на каждый. Все треки являются 16-битными моно треками с частотой дискретизации 22050Гц в .au формате. Жанры: классическая музыка, джаз, поп, кантри, блюз, диско, метал, рок, регги и хип хоп.
2. Для снижения размерности входных данных (661500-мерный вектор, представляющий один 30-секундный трек) использовали MMFC, в процессе которого использовалось окно размером в 25мс и перекрытием в 10мс для получения последовательных кадров со сглаженным сигналом. Затем с помощью преобразования Фурье были получены частотные составляющие кадров, и отображены на меловую шкалу (шкала, соответствующая восприятию звука человеком – примерно линейна до 1кГц и логарифмическая выше 1кГц). Полученные 20 бинов частотного распределения с помощью дискретного косинусного преобразования (для декорелляции частотных составляющих) были сокращены до 13. Таким образом было получено число признаков в 2600\*13 для каждого трека. Для дальнейшего сокращения размерности в эксперименте, авторы разделили все MMFC признаки на равные 4 группы, и взяли первые 40 из каждой группы. Таким образом, для каждого трека вышло 13\*160 = 2080 признаков. При отрисовке размеченных данных авторы заметили, что при двухклассовой классификации задача проста, однако при 10-классовой классификации задача значительно усложняется
3. Архитектура основной нейросети: 5 слоёв, 3 скрытых слоя. Все слои, кроме последнего, предобучаются с помощью RBM, слой за слоем:
   1. Обучить RBM для первого слоя на входных данных
   2. Обучать итеративно RBM для следующего слоя на основе данных, полученных из последнего слоя сети с предыдущего шага

Обучение нейросети:

1. Начальные веса в нейросети – соответствующие каждому слою обученные RBM
2. Использовать прямое и обратное распространение (или [метод сопряжённых градиентов](http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B5%D1%82%D0%BE%D0%B4_%D1%81%D0%BE%D0%BF%D1%80%D1%8F%D0%B6%D1%91%D0%BD%D0%BD%D1%8B%D1%85_%D0%B3%D1%80%D0%B0%D0%B4%D0%B8%D0%B5%D0%BD%D1%82%D0%BE%D0%B2)) для обучения сети
3. Обучение нейросети: для тренировочного набора было использовано 60% признаков MMFC, для тестового – 40%. Все 10 жанров имели одинаковый вес. RBM для каждого слоя обучался с помощью 5 итераций алгоритма Contrastive Divergence. При классификации на исходных данных, для 2 и 3-класовой классификации DBN (Deep Belief Network, нейронная сеть с предобученными с помощью RBM слоями) и NN показали схожие результаты. На 4-классовой классификации NN превзошла DBN, в частности, из-за более явного переобучения DBN. Однако на более большом наборе данных (были сгенерированы треки с использованием признаков MMFC для каждого из жанров), где для каждого жанра было теперь не 100, а 1500 треков, DBN превзошла NN по точности классификации.