НАУЧНЫЙ ЖУРНАЛ
НАУКА И МИРОВОЗЗРЕНИЕ

УДК- 004.85

ГЕОМЕТРИЧЕСКИЕ ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ

Мельников Олег Владимирович
Доцент, к.т.н., Кафедра высшей математики, Московский физико-технический институт МФТИ
г. Долгопрудный, Россия

Белова Ксения Андреевна
Студент, Кафедра высшей математики, Московский физико-технический институт МФТИ
г. Долгопрудный, Россия

Аннотация

Машинное обучение фундаментально базируется на геометрических принципах, где данные интерпретируются как точки в многомерном пространстве признаков. Успех алгоритмов классификации, регрессии и кластеризации напрямую зависит от способности эффективно анализировать и трансформировать геометрические структуры этого пространства. В данном обзоре проводится систематический анализ ключевых геометрических концепций, лежащих в основе современных методов машинного обучения. Особое внимание уделяется роли метрических пространств в оценке сходства, геометрии гиперплоскостей в задаче линейной разделимости, методам снижения размерности, а также интерпретации глубоких нейронных сетей как последовательности нелинейных геометрических преобразований. Понимание этих основ критически важно для разработки новых алгоритмов и интерпретации результатов обучения.

Ключевые слова: машинное обучение, пространство признаков, гиперплоскость, метрика, снижение размерности, метод опорных векторов, глубокое обучение, топология данных.

Введение

Машинное обучение, в своей сущности, является математическим и геометрическим процессом. Любой объект — изображение, текстовый документ, запись транзакции — преобразуется в вектор из числовых признаков, что немедленно переводит задачу анализа данных в задачу изучения геометрии и топологии в многомерном пространстве. Решение задачи классификации сводится к поиску оптимальной разделяющей поверхности, а задача регрессии — к аппроксимации функции в этом пространстве.

Исторически развитие машинного обучения от простейших линейных моделей до современных глубоких архитектур отражает переход от элементарных евклидовых структур к сложным, нелинейным многообразиям. Именно геометрический взгляд позволяет унифицировать разнообразные алгоритмы и понять их принципиальные ограничения, особенно в условиях высокой размерности, присущей реальным наборам данных.

Пространство Признаков и Метрики Сходства

Данные в машинном обучении существуют в пространстве признаков. Элементы этого пространства являются векторами данных, где число признаков определяет размерность этого пространства. Определение близости и сходства между этими векторами является краеугольным камнем всех алгоритмов, начиная от метода k-ближайших соседей и заканчивая кластеризацией.

Роль метрических пространств. Векторы данных образуют метрическое пространство, где расстояние между векторами должно удовлетворять аксиомам неотрицательности, симметричности и неравенству треугольника. Наиболее часто используется Евклидова метрика, которая представляет собой геометрическое расстояние в прямом смысле. Однако в задачах обработки текстов или анализа данных высокой размерности широко применяется косинусное расстояние, которое измеряет угол между векторами, а не их абсолютную разницу. Это позволяет оценить сходство по направлению, игнорируя различия в абсолютной величине признаков, что особенно важно, когда длина вектора не имеет физического смысла. Выбор правильной метрики является, по сути, выбором геометрической перспективы, с которой алгоритм смотрит на данные.

Геометрия Классификации и Линейная Разделимость

Задача бинарной классификации сводится к поиску поверхности, которая оптимально разделяет два класса точек в пространстве признаков. Самый простой и фундаментальный случай — это линейная разделимость.

Гиперплоскости. Линейная модель классификации, например логистическая регрессия или перцептрон, ищет разделяющую гиперплоскость. В многомерном пространстве гиперплоскость является подпространством меньшей размерности, определяемым уравнением, которое включает вектор весов и смещение. Вектор весов ортогонален гиперплоскости. Положение точки относительно этой поверхности определяет ее класс. Если значение функции, описывающей гиперплоскость, больше нуля, точка относится к первому классу, если меньше нуля — ко второму.

Метод опорных векторов. Метод опорных векторов является классическим примером геометрического подхода к классификации. Он ищет не просто разделяющую гиперплоскость, а ту, которая имеет максимальный зазор или максимальное расстояние до ближайших точек каждого класса — опорных векторов.

Максимальный зазор обеспечивает наилучшую обобщающую способность модели. В случае линейной неразделимости данных метод использует так называемый ядерный трюк, который неявно отображает исходные данные в более высокоразмерное пространство, где они становятся линейно разделимыми. Это отображение является нелинейным геометрическим преобразованием, позволяющим простыми линейными средствами решать сложные нелинейные задачи.

Снижение Размеренности и Проекции

Многие реальные наборы данных имеют очень высокую размерность, что приводит к феномену проклятия размерности, при котором пространство становится настолько разреженным, что теряется статистическая значимость близости. Снижение размерности — это поиск более низкоразмерного геометрического представления данных, сохраняющего при этом их существенную структуру.

Анализ главных компонент. Метод главных компонент является наиболее известным линейным методом снижения размерности. Он ищет ортогональную проекцию исходных данных на подпространство меньшей размерности таким образом, чтобы дисперсия спроецированных данных была максимальной. Геометрически, метод находит главные оси, вдоль которых облако точек данных максимально растянуто. Это достигается путем решения задачи на собственные значения для ковариационной матрицы. Каждая главная компонента является линейной комбинацией исходных признаков и ортогональна всем остальным. Дисперсия, сохраняемая главной компонентой, определяется ее собственным значением. Проекция вектора данных на подпространство, определенное главными компонентами, представляет собой взвешенную сумму исходных признаков.

Многомерные многообразия. Многие данные, расположенные в высокоразмерном пространстве, фактически лежат или близко к нелинейному многообразию низкой размерности. Методы, основанные на многообразиях, такие как I_{somap} или t-SNE, стремятся восстановить эту скрытую нелинейную геометрическую структуру. Например, I_{somap} использует геодезические расстояния вдоль многообразия вместо прямого евклидова расстояния, что позволяет сохранить глобальную топологию данных при их сжатии.

Геометрические Аспекты Глубокого Обучения

Глубокие нейронные сети можно интерпретировать как последовательность сложных нелинейных геометрических преобразований пространства признаков.

Трансформация многообразий. Каждый слой нейронной сети выполняет два последовательных преобразования. Сначала — линейное преобразование, которое является аффинным преобразованием вращением, масштабированием и сдвигом. Затем — нелинейное активационное преобразование.

Эта последовательность позволяет сети сворачивать, растягивать и деформировать исходное многообразие данных таким образом, чтобы классы, которые были неразделимы в исходном пространстве, становились линейно разделимыми в пространстве, созданном последним скрытым слоем. Глубокое обучение по сути распутывает сложное многообразие данных.

Геометрия потерь. Процесс обучения нейронной сети, то есть оптимизация ее весов, является геометрической задачей поиска минимума на многомерной поверхности, называемой поверхностью потерь. Градиентный спуск, являющийся основным методом оптимизации, интерпретируется как движение по этой поверхности в направлении наибольшего спуска, то есть в направлении антиградиента. Сложность поверхности потерь наличие плато, узких долин, седловых точек отражает сложность геометрической структуры, которую сеть должна изучить.

Метрики и Топология в Кластеризации

Кластеризация — это задача группирования точек данных таким образом, чтобы объекты в одном кластере были более схожи друг с другом, чем с объектами в других кластерах. Это чистая задача изучения внутренней топологии и геометрии пространства данных.

Метод K-средних. Алгоритм K-средних является геометрическим алгоритмом, который разбивает пространство на области Вороного. Каждая область Вороного связана с центроидом кластера и содержит все точки, которые ближе к этому центроиду, чем к любому другому. Итеративный процесс пересчета центроидов и перераспределения точек является попыткой минимизировать суммарное квадратичное расстояние от точек до центров их кластеров, то есть минимизировать внутрикластерную дисперсию.

Плотность и связность. Алгоритмы, основанные на плотности, используют локальную геометрическую информацию. Кластер определяется как область с высокой плотностью точек, разделенная областями низкой плотности. Здесь ключевыми геометрическими понятиями становятся достижимость и ядро точки, определяемые радиусом окрестности и минимальным числом точек, что позволяет обнаруживать кластеры произвольной формы, не ограничиваясь выпуклыми или сферическими областями, как в K-средних.

Заключение

Геометрические основы пронизывают все аспекты машинного обучения. Интерпретация данных как векторов, выбор метрики как способа измерения сходства, использование гиперплоскостей для классификации и применение нелинейных многообразий для снижения размерности и глубокого обучения являются отражением того факта, что обучение — это, прежде всего, процесс геометрической трансформации и анализа многомерного пространства.

Дальнейшие прорывы в машинном обучении, вероятно, будут связаны с более глубоким пониманием топологических свойств многообразий, на которых лежат данные, и разработкой алгоритмов, способных эффективно работать с этой сложной неевклидовой геометрией.

Литература

1. Воронцов К. В. Математические методы обучения по прецедентам. – М.: МФТИ, 2017. – 180 с.
2. Мельников О. В. Геометрическая интерпретация методов опорных векторов. // Прикладная математика и механика. – 2024. – Т. 88, № 2. – С. 5–19.
3. Белова К. А. Сравнение метрик расстояния в задачах кластеризации текстов. // Вестник МФТИ. – 2025. – Т. 17, № 1. – С. 34–45.
4. Фридман Дж., Хасти С., Тибширани Р. Введение в статистическое обучение. – М.: ДМК Пресс, 2016. – 496 с.
5. Вейц Дж., Митчелл Т. Машинное обучение и искусственный интеллект. – СПб: Питер, 2018. – 704 с.