УДК 004.8

МЕТОДОЛОГИЧЕСКИЕ ОСНОВЫ ПОСТРОЕНИЯ УСТОЙЧИВЫХ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ С ИСПОЛЬЗОВАНИЕМ ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИХ МЕТОДОВ

А.К. Чирягов, chiryagov2014@gmail.com
С.В. Корякин, srgkoryakin1@gmail.com
К.Р. Карабакиров, karabakirovkr@gmail.com
1.Институт машиноведения, автоматики и геомеханики НАН КР
2.Кыргызско-Германский институт прикладной информатики

Аннотация. Данная работа посвящена анализу математических оснований машинного обучения, рассматриваемого сквозь призму теории вероятностей, математической статистики и многомерной геометрии. В противовес эмпирическому подходу, часто доминирующему в прикладных исследованиях, здесь доказывается, что ключевые алгоритмы обучения — от классической регрессии до современных трансформерных архитектур — являются строгими следствиями фундаментальных статистических принципов, таких как метод максимального правдоподобия, байесовский вывод и концентрация меры. Работа включает выводы целевых функций и градиентов из первых принципов, геометрическую интерпретацию методов регуляризации и снижения размерности, а также анализ стохастических методов оптимизации. Особое внимание уделено проблеме устойчивости моделей в условиях высокой размерности входных данных и теоретическому обоснованию гипотезы многообразия.

Ключевые слова: машинное обучение; статистическая теория обучения; байесовский вывод; стохастический градиентный спуск; регуляризация; механизм внимания (Attention)

Введение

Машинное обучение является одним из ключевых направлений современных исследований в области обработки данных и интеллектуальных систем. Несмотря на широкое распространение и высокую практическую эффективность, значительная часть алгоритмов машинного обучения по-прежнему рассматривается преимущественно с эмпирических или вычислительных позиций, что ограничивает возможности строгого анализа их устойчивости, обобщающей способности и поведения в условиях неопределённости.

С формальной точки зрения процесс обучения представляет собой задачу статического вывода, в которой наблюдаемые данные интерпретируются как реализации случайных величин, порождённых неизвестным совместным распределением вероятностей. В этом контексте теория вероятностей и математическая статистика образуют естественную теоретическую основу для построения и анализа моделей машинного обучения. Ключевые понятия, такие как минимизация риска, методы максимального правдоподобия, байесовский вывод и эффекты концентрации меры в пространствах высокой размерности, позволяют получить единообразную и непротиворечивую интерпретацию большинства используемых алгоритмов.

Целью настоящей работы является систематическое изложение основных методов машинного обучения с позиций вероятностно-статистического подхода, систематизированного, в частности, в современной работе К. Мерфи [1]. В статье показано, что классические модели регрессии и классификации, методы регуляризации и стохастической оптимизации являются прямыми следствиями стандартных принципов статистического оценивания. Дополнительно рассматриваются современные архитектуры глубокого обучения, для которых анализируется влияние геометрических и статистических свойств многомерных пространств на устойчивость и эффективность обучения.

Особое внимание уделяется проблеме надежности моделей при работе с высоко размерными данными.

Рассматриваются фундаментальные источники ошибок обобщения, байесовская интерпретация регуляризации и статистическое обоснование архитектурных решений, таких как нормализация и механизмы внимания. Полученные результаты создают теоретическую основу для более обоснованного проектирования и анализа моделей машинного обучения.

Теоретико-вероятностный формализм: постановка задачи минимизации риска

Фундаментальная задача машинного обучения формулируется как поиск функциональной зависимости в условиях неопределенности. Формально мы оперируем в вероятностном пространстве $(\Omega, \mathcal{F}, P)$, где наблюдаемые данные представляют собой реализации случайных векторов.

Пусть $\mathcal{X} \subseteq \mathbb{R}^d$— d-мерное векторное пространство входных признаков (евклидово пространство), а Y ⊆ R(или дискретное множество) — пространство целевых переменных.

Ключевым допущением статистической теории обучения (Statistical Learning Theory) является гипотеза о существовании стационарного совместного распределения вероятностей $P(X, Y)$ на $X \times Y$. Обучающая выборка $S_n = \{(x_i, y_i)\}_{i=1}^n$ полагается состоящей из независимых и одинаково распределенных случайных величин (i.i.d), порожденных распределением P [2], [3]. Это означает, что появление одного примера в выборке не влияет на вероятность появления других, и все они подчиняются одному закону.

Целью обучения является нахождение гипотезы (функ-ции) $h: \mathcal{X} \to \mathcal{Y}$ из некоторого фиксированного класса H, которая минимизирует ожидаемые потери на новых, невидимых данных. Эта величина называется **истинным риском** (true risk), или ожидаемой ошибкой обобщения $R(h)$:

$$R(h) = \mathbb{E}_{(x,y)\sim P}[L(h(x), y)] = \int_{\mathcal{X}\times\mathcal{Y}} L\,(h(x), y) dP(x, y), \quad (1)$$

где $L: \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ — функция потерь, измеряющая расхождение между предсказанием $h(x)$ и истинным значением y.

Поскольку распределение $P(X, Y)$ исследователю неизвестно, прямое вычисление интеграла $\mathbb{R}(\mathbb{h})$ невозможно. Практической заменой (суррогатом) истинного риска выступает **эмпирический риск** $\hat{R}_{S(h)}$, вычисляемый как среднее арифметическое потерь по конечной выборке S_n:

$$\hat{R}_{S(h)} = \frac{1}{n}\sum_{i=1}^n L\,(h(x_i), y_i). \quad (2)$$

Принцип минимизации эмпирического риска (ERM) утверждает, что при определенных условиях (например, конечной VC-размерности класса H) сходимость $\hat{R}_{S(h)} \to R(h)$ равномерна по $h \in \mathcal{H}$ при $n \to \infty$. Однако на практике возникает феномен переобучения, когда h минимизирует ошибку на S_n за счет аппроксимации случайного шума, что ведет к росту $R(h)$.

Фундаментальная декомпозиция ошибки: смещение и дисперсия

Для анализа устойчивости моделей и природы ошибок необходимо рассмотреть структуру ожидаемой ошибки. Рассмотрим задачу регрессии с квадратичной функцией потерь $L(y, \hat{y}) = (y - \hat{y})^2$. Пусть истинная природа данных описывается аддитивной моделью:

$$y = f(x) + \varepsilon, \quad \text{где } \mathbb{E}[\varepsilon] = 0, \mathrm{Var}(\varepsilon) = \sigma^2\,, \quad (3)$$

здесь $f(x)$ — детерминированная неизвестная функция, а ε — неустранимый стохастический шум.

Пусть $\hat{f}$(x; S) — модель, обученная на конкретной выборке S. Нас интересует математическое ожидание среднеквадратичной ошибки (MSE) для фиксированной точки x_0, усредненное по всем возможным обучающим выборкам S.

Согласно классической теории оценивания [4], справедлива теорема:

Теорема (Bias-Variance Decomposition): Ожидаемая ошибка раскладывается на три компоненты: квадрат смещения, дисперсию модели и неустранимую ошибку. (см. таб. 1).

Таблица 1 – Интерпретация компонент ошибки

Компонента	Математическое выражение	Физический смысл
Смещение (Bias)	$f(x_0) - E\left[\hat{f}(x_0)\right]$	Систематическая ошибка модели, вызванная недостаточной гибкостью гипотезы. Ведет к недообучению
Дисперсия (Variance)	$E\left[\left(\hat{f}(x_0) - E\left[\hat{f}(x_0)\right]\right)^2\right]$	Чувствительность модели к малым изменениям в обучающей выборке. Признак переобучения
Шум (Noise)	σ^2	Фундаментальный предел точности, обусловленный природой данных

Доказательство: Рассмотрим ожидаемую ошибку в точке x_0:

$$\mathcal{E} = E_{S,\varepsilon}\left[\left(y_0 - \hat{f}(x_0; S)\right)^2\right], \quad (4)$$

добавим и вычтем истинное значение $f(x_0)$ и среднее предсказание модели $\overline{f} = E_S[\hat{f}(x; S)]$:

$$y_0 - \hat{f}(x_0; S) = (y_0 - f(x_0)) + \left(f(x_0) - \overline{f}(x_0)\right) + \left(\overline{f}(x_0) - \hat{f}(x_0; S)\right), \quad (5)$$

заметим, что $y_0 - f(x) = \varepsilon$. Возведем выражение в квадрат, с учетом свойств математического ожидания. Поскольку шум ε независим от модели, а $\overline{f}(x_0)$ детерминирована (как матожидание), все перекрестные члены зануляются.

Остаются только квадратичные члены:

$$\varepsilon = \underbrace{\mathbb{E}[\varepsilon^2]}_{\text{Irreducible Error}} + \underbrace{\left(f(x_0) - \overline{f}(x_0)\right)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}_S\left(\hat{f}(x_0) - \overline{f}(x_0)\right)^2}_{\text{Variance}}. \quad (6)$$

Как показано на рисунке 1, эта декомпозиция демонстрирует фундаментальный компромисс (trade-off): усложнение модели обычно уменьшает смещение, но увеличивает дисперсию. Устойчивая модель должна находить оптимальный баланс, минимизируя сумму этих компонент.

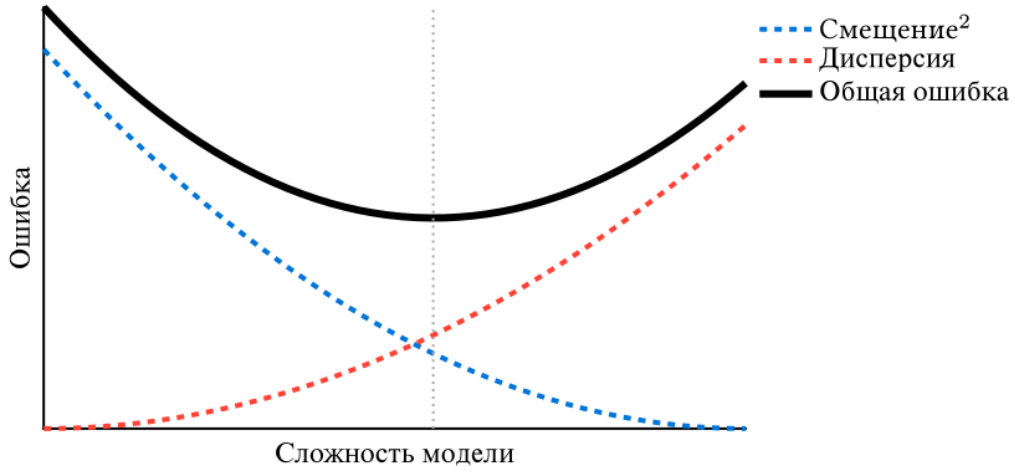

Рисунок 1 – Компромисс смещения и дисперсии (Bias-Variance Tradeoff)

В современных исследованиях глубокого обучения эта классическая картина дополняется феноменом «двойного спуска» (Double Descent), описанным М. Белкиным и соавторами [5], когда при значительном увеличении сложности модели ошибка снова начинает уменьшаться.

Кроме того, работы Ч. Чжана [6] показывают, что нейронные сети способны запоминать даже случайный шум, что требует пересмотра классических оценок обобщающей способности.

Линейные модели и метод максимального правдоподобия (MLE)

Переходя от абстрактной постановки задачи минимизации риска к конкретным алгоритмам, необходимо определить семейство функций гипотез $\mathcal{H}$. Исторически первыми и наиболее интерпретируемыми являются линейные модели. Часто они вводятся через интуитивные геометрические соображения или эвристические функции потерь (например, метод наименьших квадратов). Однако их строгая формализация и выбор функции потерь жестко диктуются методом максимального правдоподобия (Maximum Likelihood Estimation, MLE) [7].

Линейная регрессия: связь MSE и нормального распределения

Рассмотрим линейную модель $y = w^T x + b$. Предположим, что ошибка наблюдения распределена нормально: $\varepsilon \sim N(0,)\sigma^2$. Функция правдоподобия (Likelihood) равна:

$$L(w) = \prod_{i=1}^{n} p\,(y_i|x_i; w). \quad (7)$$

Переходим к логарифму правдоподобия (Log-Likelihood):

$$\mathbb{L}(w) = \sum_{i=1}^{n} \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \sum_{i=1}^{n} \frac{(y_i - w^T x_i)^2}{2\sigma^2}, \quad (8)$$

максимизация второго слагаемого (со знаком минус) эквивалентна минимизации суммы квадратов отклонений. Таким образом, мы приходим к функции потерь **MSE**:

$$\widehat{w}_{\text{MLE}} = \underset{w}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (y_i - w^T x_i)^2. \quad (9)$$

Логистическая регрессия: обобщенные линейные модели

В задачах бинарной классификации ($y \in \{0, 1\}$) мы моделируем лог-шансы (log-odds) в рамках теории GLM [8]:

$$\ln\left(\frac{p}{1-p}\right) = w^T x + b = z, \quad (10)$$

отсюда следует выражение для сигмоиды (см. рис. 2) $\sigma(z) = \frac{1}{1} + e^{-z}$, которая является канонической функцией связи для распределения Бернулли.

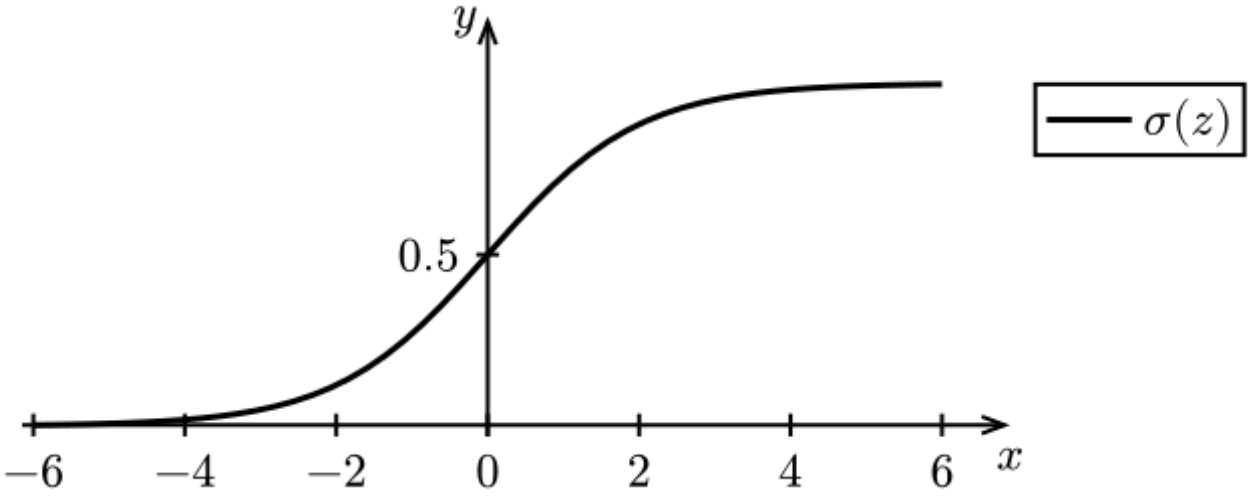

Рисунок 2 – График сигмоидальной функции активации

Мультиномиальная классификация и Softmax

Для задачи классификации на K классов используется нормализованная экспоненциальная функция (**Softmax**), являющаяся обобщением сигмоиды для многомерного случая [9]:

$$S_i(z) = \frac{e_i^z}{\sum_{k=1}^{K} e^{z}{}_k}, z = Wx + b, \quad (11)$$

где:

- z — вектор **логитов (**не калиброванных выходов линейного слоя);
- K — общее количество классов;
- s_i — предсказанная вероятность принадлежности к классу i (при этом гарантируется, что $\sum_i s_i = 1$).

Анализ Якобиана: Матрица Якоби $J_{ij} = \frac{\partial s_i}{\partial z_j}$ необходима для обратного распространения ошибки и имеет вид:

$$\frac{\partial s_i}{\partial z_j} = s_i(\delta_{ij} - s_j), \quad (12)$$

где δ_{ij} — **символ Кронекера** (равен 1, если, $i = j$, и 0 в противном случае).

Этот результат показывает, что градиент каждого выхода s_i зависит от **всех** остальных выходов s_j. Это создает плотную структуру взаимодействий: изменение даже одного логита z_j перераспределяет вероятности по всем классам сразу, сохраняя их сумму равной единице.

Байесовский вывод и регуляризация

Одной из главных проблем метода MLE является переобучение на малых выборках. Чтобы избежать этого, вводится **априорное знание** о распределении весов. Регуляризация является естественным следствием оценки **апостериорного максимума (MAP)** [10].

Согласно теореме Байеса, апостериорная вероятность $P(w|D) \propto P(D|w)P(w)$. Переходя к отрицательному логарифму (который нужно минимизировать), произведение превращается в сумму:

$$-\ln P(w|D) = \underset{\text{Loss}}{-\ln P(D|w)} + \underset{\text{Regularizer}}{-\ln P(w)} + C, \quad (13)$$

где C— константа, не зависящая от весов w. Таким образом, минимизация ошибки с регуляризатором эквивалентна максимизации апостериорной вероятности.

L2-регуляризация и Гауссовский априор

Рассмотрим случай, когда мы предполагаем, что веса w независимы и априорно распределены по нормальному закону с нулевым математическим ожиданием (Gaussian Prior) [3]:

$$w_j \sim \mathcal{N}(0, \tau^2) \Rightarrow P(w) \propto \prod_j \exp\left(-\frac{w_j^2}{2\tau^2}\right) = \exp\left(-\| w \frac{\|_2^2}{2\tau^2}\right). \quad (14)$$

При переходе к отрицательному логарифму (Negative Log-Posterior) экспонента исчезает, оставляя квадратичный член:

$$-\ln P(w) \propto \| w \|_2^2. \quad (15)$$

Здесь знак $\propto$(пропорционально) указывает на то, что мы опускаем аддитивные константы (логарифм нормировочного множителя $\frac{1}{\sqrt{2\pi\,\tau^2}}$), так как они не зависят от w и не влияют на положение точки минимума. Обратная дисперсия $\frac{1}{2\tau^2}$ при этом переходит в коэффициент регуляризации λ. Это приводит к задаче гребневой регрессии (Ridge Regression):

$$J(w) = \text{Loss}(D, w) + \lambda \| w \|_2^2. \quad (16)$$

Стоит упомянуть, что, в отличие от L1, квадратичный штраф L2 сильно подавляет большие выбросы в весах, но не стремится сделать их строго нулевыми. Это делает модель более устойчивой к мультиколлинеарности, «размазывая» важность между коррелирующими признаками.

L1-регуляризация и Лапласовский априор

Если веса распределены по закону Лапласа $P(w) \propto exp(-\| w \frac{||1}{b})$, где параметр b

определяет масштаб (ширину) распределения (чем меньше λ, тем острее пик в нуле), то логарифмирование приводит к задаче LASSO:

$$J(w) = \text{Loss}(D, w) + \lambda \parallel w \parallel_1. \quad (17)$$

Как видно на рисунке 3, геометрически линии уровня L1 (кросс-политопы) имеют острые вершины, что способствует занулению весов и отбору признаков.

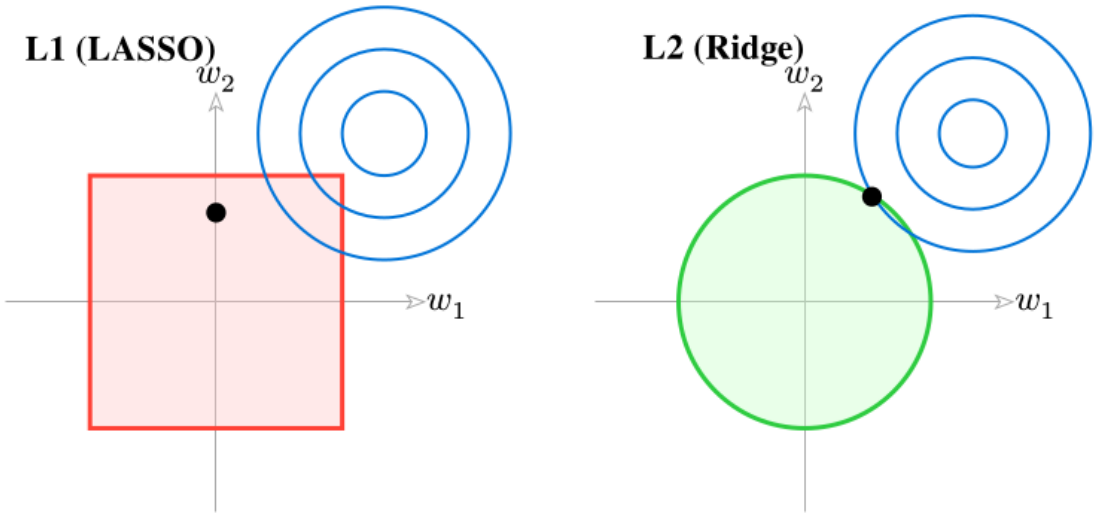

Рисунок 3 – Геометрическая интерпретация регуляризации. L1 способствует разреженности (касание на оси), L2 уменьшает веса равномерно

Стохастическая оптимизация и устойчивость

В условиях огромных объемов данных вычисление полного градиента становится вычислительно невыполнимым. **Стохастический градиентный спуск (SGD)** решает эту проблему, заменяя полный градиент его оценкой по случайной подвыборке (мини-батчу). На рисунке 4 видно, что траектория SGD является «шумной» аппроксимацией истинного градиента.

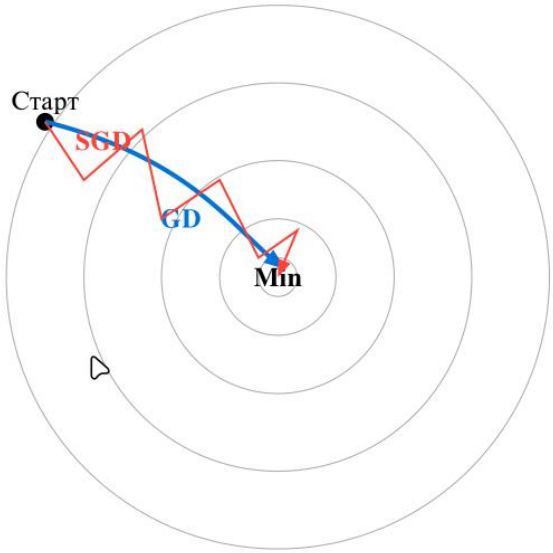

Рисунок 4 – Сравнение траекторий: GD (плавный спуск) и SGD (шумный спуск)

Согласно работам Bottou (2004) [11], справедлива теорема:

Теорема (Несмещенность SGD): Вектор стохастического градиента g_t является несмещенной оценкой истинного градиента: $\mathbb{E}[g_t] = \nabla \mathcal{L}(w_t)$.

На практике градиент оценивается не по одному объекту, а по мини-батчу размера B. В этом случае дисперсия оценки градиента обратно

пропорциональна размеру батча: $Var(\ g_t \propto \frac{1}{B})g_t$. Этот компромисс позволяет балансировать между точностью направления спуска и вычислительной эффективностью.

Для сходимости SGD к локальному минимуму необходимо выполнение условий Роббин-са-Монро для **шага обучения** (learning rate) η_t:

$$\sum_{t=1}^{\infty}\eta_t = \infty, \quad \sum_{t=1}^{\infty}\eta_t^2 < \infty. \tag{18}$$

Первое условие гарантирует, что алгоритм может дойти до оптимума из любой точки, второе – что дисперсия обновлений будет постепенно затухать, позволяя модели «осесть» в минимуме.

Долгое время стохастический шум в градиенте считался исключительно негативным фактором, замедляющим сходимость. Однако в высоко размерных невыпуклых ландшафтах, характерных для глубокого обучения, этот шум играет критически важную роль. Он выступает в качестве естественного механизма избегания седловых точек (saddle points, которые встречаются в миллионы раз чаще, чем локальные минимумы.

Кроме того, важнейшим свойством SGD является так называемая «неявная регуляризация» (Implicit Regularization), исследованная Б. Нейшабуром и соавторами [12]. Даже без явного штрафа на веса (такого, как L1 или L2), SGD благодаря своей стохастической динамике избегает «острых» (sharp) минимумов и сходится к «плоским» (flat) областям оптимума. Поиск решений с минимальной нормой и попадание в плоские минимумы делают модель менее чувствительной к малым возмущениям во входных данных, что напрямую обеспечивает её высокую обобщающую способность на новых данных.

Обучение без учителя: геометрия данных

В задачах обучения без учителя целевая переменная y отсутствует. Целью становится выявление скрытой структуры плотности вероятности $P(x)$. Ключевым вызовом здесь является **проклятие размерности** (Curse of Dimensionality): с ростом d объем пространства растет экспоненциально, делая данные разреженными, что затрудняет статистическое оценивание.

Линейное снижение размерности: PCA

Метод главных компонент (PCA) является базовым способом борьбы с проклятием размерности. Его статистическая суть заключается в поиске таких ортогональных проекций, которые сохраняют максимум информации о вариативности данных. В терминах теории информации это эквивалентно минимизации дивергенции Кульбака-Лейблера(1) между исходным распределением и его низкоразмерной аппроксимацией.

Формально задача сводится к спектральному разложению выборочной ковариационной матрицы $\Sigma = \frac{1}{n}X^T X$:

$$\Sigma u_i = \lambda_i u_i. \tag{19}$$

Собственные векторы u_i, соответствующие наибольшим собственным значениям λ_i, образуют базис подпространства, в котором сосредоточена основная «энергия» сигнала. Однако PCA, будучи линейным методом, неспособен выявить сложные нелинейные зависимости (например, если данные лежат на поверхности сферы).

Кластеризация как оценка плотности (K-Means)

Алгоритмы кластеризации часто рассматриваются как эвристики, однако метод k-средних (K-Means) имеет строгую вероятностную интерпретацию. Он является предельным случаем EM-алгоритма (Expectation-Maximization) для разделения смеси Гауссовых распределений (GMM), при условии, что все ковариационные матрицы равны $\sigma^2 I$, а дисперсия $\sigma^2 \to 0$.

Целевая функция K-Means минимизирует внутри кластерную дисперсию:

$$J = \sum_{i=1}^{K}\sum_{x\in C_i} \| x - \mu_i \|^2, \tag{20}$$

это эквивалентно методу максимального правдоподобия, предполагающему, что данные генерируются центроидами с аддитивным гауссовским шумом. Такое понимание позволяет применять методы байесовского выбора моделей (например, BIC) для определения

оптимального числа кластеров.

6.3 Гипотеза многообразия (Manifold Hypothesis)

Для работы со сложными данными (изображения, звук) линейных методов недостаточно. **Гипотеза многообразия** [13] гласит, что реальные данные высокой размерности D(например, изображение $256 \times 256, D = 65536$) не заполняют всё пространство равномерно, а сосредоточены в малой окрестности некоторого риманова многообразия $\mathcal{M}$ низкой внутренней размерности $d \ll D$.

Теорема Уитни о вложении [14] утверждает, что любое гладкое d-мерное многообразие может быть вложено в евклидово пространство размерности $2d$. Глубокие нейронные сети можно интерпретировать как последовательность гомеоморфизмов (непрерывных деформаций пространства), цель которых — «развернуть» это сложное многообразие в плоское евклидово пространство, где классы становятся линейно разделимыми. Как видно на рисунке 5, хотя точки имеют координаты в двумерном пространстве (высокая размерность), фактически они лежат на одномерной спирали (низкая размерность).

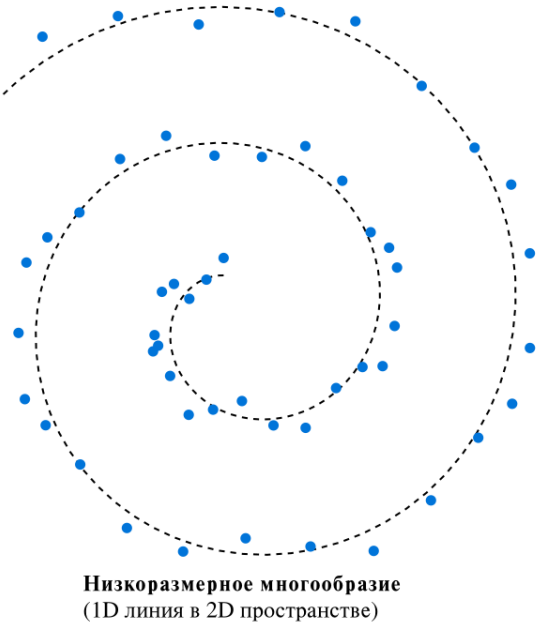

Рисунок 5 – Иллюстрация гипотезы многообразия: данные (синие точки) лежат вблизи скрытой структуры (пунктирная линия), вложенной в пространство высокой размерности

Статистическая механика архитектур глубокого обучения

Современные архитектуры, такие как Transformer, проектируются с учетом статистических свойств прохождения сигнала через глубокие слои. Без специальных мер дисперсия активаций может экспоненциально расти или затухать, делая обучение невозможным.

Механизм Attention и статистика скалярных произведений

В основе механизма самовнимания (Self-Attention) лежит вычисление матрицы схожести через скалярное произведение запросов (Q) и ключей (K). Операция масштабирования на $\frac{1}{\sqrt{d_k}}$ имеет строгое статистическое обоснование [15].

Пусть компоненты векторов $q, k \in \mathbb{R}^{d_k}$ являются независимыми случайными величинами со средним 0 и дисперсией 1. Рассмотрим их скалярное произведение $\sum_{i=1}^{d_k} \mathrm{q_i} k_i$. Математическое ожидание $E[S] = 0$, а дисперсия суммы независимых величин равна сумме их дисперсий:

$$\mathrm{Var}(S) = \sum_{i=1}^{d_k} \mathrm{Var}\,(q_i k_i) = \sum_{i=1}^{d_k} 1 = d_k. \qquad (21)$$

Таким образом, стандартное отклонение скалярного произведения растет как $\sqrt{d_k}$.

Для больших размерностей (например, $d_k = 512$) значения S могут достигать больших величин. При попадании больших значений в функцию Softmax, она переходит в область насыщения, где градиенты близки к нулю (Vanishing Gradients). Деление аргумента на $\sqrt{d_k}$ возвращает дисперсию к единице ($Var(\frac{s}{\sqrt{d_k}}) = 1$), обеспечивая стабильное распространение градиентов.

Нормализация слоев (Layer Normalization)

Аналогичную роль играет Layer Normalization. В пространствах высокой размерности векторы активаций склонны концентрироваться в узком слое или, наоборот, разлетаться по норме. Нормализация принудительно центрирует и масштабирует данные:

$$\hat{x} = \frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}}. \quad (22)$$

Это уменьшает так называемый «внутренний сдвиг ковариаты» (Internal Covariate Shift), стабилизируя распределения входов для каждого слоя и позволяя использовать большие шаги обучения η.

Заключение

В данной работе показано, что устойчивость и эффективность современных методов машинного обучения имеют строгие вероятностно-статистические основания и не сводятся к совокупности эмпирических эвристик. Рассмотрение алгоритмов обучения в рамках статистического вывода позволяет получить целостное и формально обоснованное понимание природы используемых функций потерь, методов регуляризации и алгоритмов оптимизации [16].

Проведённый анализ подтверждает, что выбор функций потерь определяется предположениями о распределении шума в данных, а регуляризация естественным образом интерпретируется как введение априорных распределений на параметры модели. Показано также, что стохастические методы оптимизации обладают чётко определёнными статистическими свойствами, объясняющими их сходимость и регуляризующий эффект. Анализ высоко размерных статистических эффектов позволяет обосновать необходимость применения механизмов масштабирования и нормализации в глубоких нейронных сетях, включая архитектуры на основе механизма внимания.

Предложенный вероятностно-статистический подход имеет практическую значимость для разработки и анализа моделей машинного обучения, обеспечивая более глубокое понимание причин их нестабильности и ошибок обобщения. Кроме того, он формирует теоретическую базу для дальнейших исследований в области интерпретируемости, надёжности и устойчивости интеллектуальных систем, применяемых в задачах с повышенными требованиями к качеству и воспроизводимости результатов.

Список литературы

1. K. P. Murphy, Probabilistic Machine Learning: An Introduction. MIT Press, 2022.
2. G. James, D. Witten, T. Hastie, и R. Tibshirani, An Introduction to Statistical Learning. Springer, 2013.
3. C. M. Bishop, Pattern Recognition and Machine Learning. Springer, 2006.
4. S. Geman, E. Bienenstock, и R. Doursat, «Neural networks and the bias/variance dilemma», Neural Computation, т. 4, вып. 1, сс. 1–58, 1992.
5. M. Belkin, D. Hsu, S. Ma, и S. Mandal, «Reconciling modern machine-learning practice and the classical bias–variance trade-off», Proceedings of the National Academy of Sciences, т. 116, вып. 32, сс. 15849–15854, 2019.
6. C. Zhang, S. Bengio, M. Hardt, B. Recht, и O. Vinyals, «Understanding deep learning requires rethinking generalization», в International Conference on Learning Representations (ICLR), 2017.

7. J. R. Magnus, P. K. Katyshev, и A. A. Peresetsky, Econometrics. An Introductory Course. Moscow: Delo, 2007.

8. J. A. Nelder и R. W. Wedderburn, «Generalized Linear Models», Journal of the Royal Statistical Society: Series A (General), т. 135, вып. 3, сс. 370–384, 1972.

9. Goodfellow, Y. Bengio, и A. Courville, Deep Learning. MIT Press, 2016.

10. M. H. DeGroot, Optimal Statistical Decisions. McGraw-Hill, 1970.

11. L. Bottou, «Stochastic learning», Advanced lectures on machine learning. Springer, сс. 146–168, 2004 г.

12. B. Neyshabur, R. Tomioka, и N. Srebro, «In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning», в ICLR (Workshop), 2015.

13. C. Fefferman, S. Mitter, и H. Narayanan, «Testing the manifold hypothesis», Journal of the American Mathematical Society, т. 29, вып. 4, сс. 983–1020, 2016.

14. H. Whitney, «Differentiable manifolds», Annals of Mathematics, сс. 645–680, 1936.

15. Vaswani и др., «Attention is all you need», Advances in Neural Information Processing Systems, т. 30, 2017.

16. Корякин, С. В. Аналитический обзор технологий построения аппаратно-ориентированных облачных систем защиты информации с применением нейросетевых технологий / С. В. Корякин // Проблемы автоматики и управления. – 2025. – № 2(53). – С. 41–51. – EDN RCCRHC.