# Проклятие размерности (Curse of dimensionality)

Проклятие размерности — явление, которое возникает в анализе данных и машинном обучении в пространствах высокой размерности и связано с ухудшением работы обучаемых моделей, построенных на большом числе признаков. Термин был введен Ричардом Беллманном при решении задач динамического программирования.

Для оценки влияния проклятия размерности достаточно просто оценивать качество моделей после включения в них каждой новой независимой переменной. При этом может наблюдаться феномен Хьюза. Он заключается в том, что при увеличении числа признаков в обучающем множестве точность модели сначала растет из-за привлечения дополнительной информации о зависимостях в данных, но с какого-то момента начинает падать в связи с началом действия проклятия размерности.

Проклятие размерности связано со следующими основными проблемами:

- увеличиваются вычислительные затраты при работе аналитических алгоритмов из-за необходимости обрабатывать большее число переменных в пространствах большой размерности;
- при добавлении в пространство признаков новых измерений, происходит увеличение его объема при сохранении неизменным числа наблюдений в обучающих данных. Как следствие, данные перестают покрывать достаточную область пространства признаков, их становится недостаточно для построения качественной аналитической модели;
- некоторые геометрические свойства пространств высокой размерности зачастую являются контринтуитивными. Например, объем сферы единичного радиуса в 20-мерном пространстве практически равен 0. Поэтому модели машинного обучения, особенно основанные на расстоянии, хорошо работают в пространствах низкой размерности, но в высокой могут оказаться несостоятельными;
- с увеличением размерности искажается форма статистических законов распределения: снижается их локализация около среднего значения. Это негативно влияет на работу статистических моделей;
- эффект концентрации нормы приводит к тому, что с ростом числа измерений попарные расстояния между векторами объектов стремятся к одному значению. Как следствие выразительность представления в данных зависимостях и закономерностях, выраженность кластерных структур снижается.

Первая проблема может быть решена организационными методами за счет наращивания вычислительных мощностей. В то же время остальные могут создать достаточно большие трудности математического и алгоритмического плана, связанные с необходимостью оценки негативного влияния проклятия размерности и выработки мер по его снижению.

Вторую проблему можно решить путем обогащения данных, добавляя в обучающее множество наблюдения, собранные дополнительно или синтезированные искусственно.

Наиболее универсальным подходом для борьбы с проклятием размерности является ее снижение в пространстве признаков, например, с помощью метода главных компонент.