Обучающая выборка

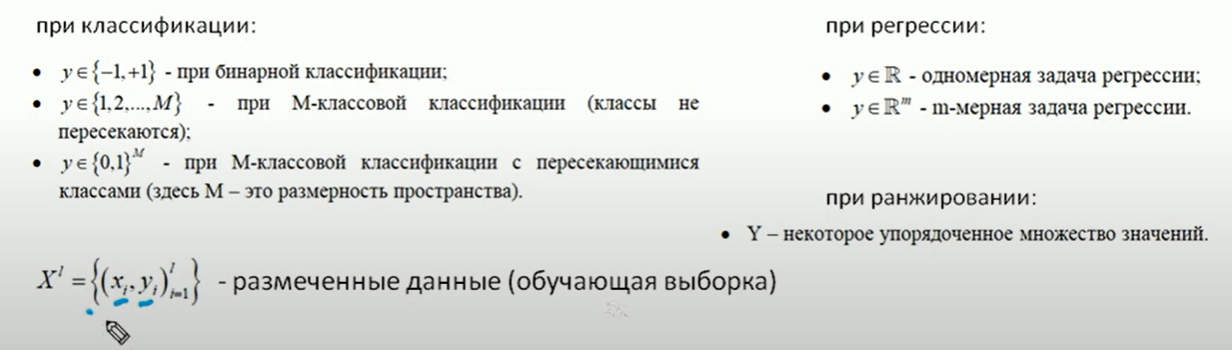
|  |  |
| --- | --- |
| Обучающая выборка  Признаки (features) | Целевые значения |
|  |  |

При классификации целевые значения:

– при бинарной классификации

– при M-классовой классификации (классы не пересекаются)

- при M-классовой классификации с пересекающимися классами (здесь M – это размерность пространства)



Выделение вторичных признаков – feature extractions или feature engineering. Математически это представляет собой функциональное преобразование первичных признаков

В результате функционального преобразования первичных признаков получается новое признаковое пространство, упрощающее работу алгоритмов обучения

Модель, решающая функция

– подбираемые параметры

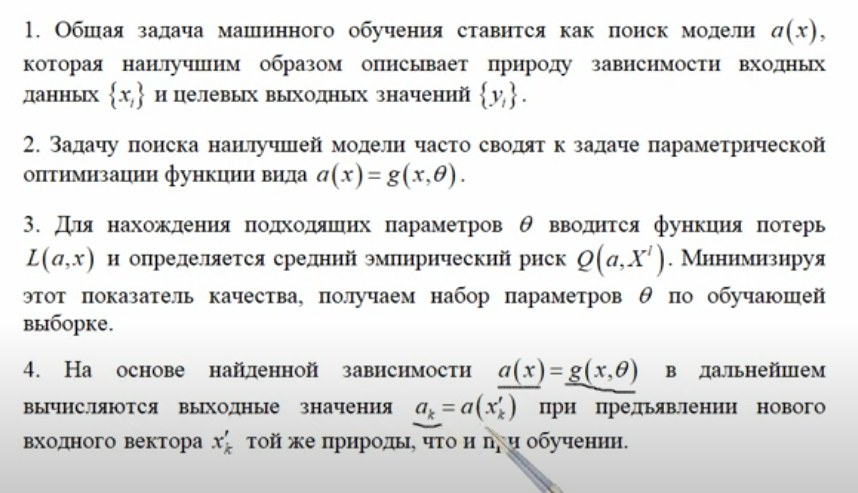
Средний эмпирический риск – это показатель качества модели, который нужно минимизировать подбирая значения параметра Θ

Чтобы обучить модель (другими словами найти наилучшие значения вектора параметров Θ) нужно задать функцию потерь (loss function) и среди всех возможных параметров Θ, которые приводят к минимуму среднего показателя эмпирического риска.

Для линейной зависимости и квадратичной функции потерь функционал качества будет иметь следующий вид

Параметры k и b можно найти из следующей системы:

Общий принцип алгоритмов машинного обучения:



Математическое описание модели машинного обучения. Здесь левая часть уравнения это и есть оптимальная модель

Где a – модель, решающая функция, заданная параметрически; Q – средний эмпирический риск

Переобучение (overfitting) практически все модели обладают подобным недостатком

Оценка степени преобученности модели:

- оценка по отложенной выборке (hold-out);

- скользящий контроль (leave-one-out);

- кросс-валидация (cross-validation, k-fold).

Обобщение моделей:

Наиболее подходящую модель можно выбрать голосованием

Метод разделения признаков

Омега – разделяющий вектор