#### Министерство науки и высшего образования Российской Федерации

#### Федеральное государственное бюджетное образовательное учреждение

#### высшего образования

#### «Владимирский государственный университет

#### имени Александра Григорьевича и Николая Григорьевича Столетовых»

**(ВлГУ)**

**Кафедра информационных систем и программной инженерии**

Рейтинг-контроль №2

по дисциплине

"Технологии искусственного интеллекта"

Выполнил:

ст. гр. ПРИм-124

Парахин К.В.

Приняла:

доц. Кафедры ИСПИ

Озерова М.И.

Владимир, 2025 г.

**Лекции рейтинг 2**

Конспект по информации из учебника Базовые алгоритмы машинного обучения на языке Python

В рамках машинного обучения происходят некоторые преобразования над моделью, при этом происходит отображение исходного набора признаков в совокупность предсказаний над моделью – также для выполнения предсказаний используются параметры и гиперпараметры модели (y) 

В рамках обучения модели необходимо получить предсказания по тестовым данным – которая содержит входные признаки и выходные показатели, при этом стоит понимать, что каждая такая модель может иметь погрешности, которые оцениваются с помощью функции потерь L (yi, yпi)

Сами функции потерь могут отличаться в зависимости от типа данных модели – но ее значение количественно обычно всегда соответствует количественной разнице между показателями модели и реальностью

Для анализа качества модели исследуется среднее значение функции потерь для набора предсказанных результатов: 

Цель заключается в том, чтобы подобрать модель, при которой значение функции Q устремится к нулю – то есть ту, в которой лучше подобраны параметры, при этом чаще всего более сложные модели (которые содержат больше признаков и зависимостей между переменными) – имеют средние потери меньше, чем подобные простые модели

Но стоит учесть, что при повышении сложности модели на тестовых данных в какой-то момент почти всегда наступает явление переобучения – когда модель вдруг начинает показывать большие потери – потому что она перестает находить общие закономерности, а начинает просто пересказывать тренировочные данные.

От переобучения нет никакой практической пользы – поэтому в рамках обучения на тестовых данных их нужно использовать, чтобы вовремя произвести остановку алгоритма – чтобы это сделать, тренировочная выборка делится на 2 части: первую используют для обучения, а вторую – для отложенной валидации получаемой тестовой выборки.

Для повышения качества валидации ее повторяют многократно, например, существует n-Fold кросс-валидация, в которой идет разделение тренировочных данных на n частей – обучение на n – 1 частях и валидация на оставшейся части, а затем оценка потерь путем усреднения результатов тестовой выборки на n частях

Благодаря такому подходу моделируется подход с тестированием модели на неизвестных новых данных

При этом стоит учесть, что функционал потерь тоже не однороден и его можно разложить на набор составляющих: 

Он состоит из ошибок смещения (Bias), ошибок дисперсии (Variance) и случайного шума – который никак нельзя уменьшить

Ошибка смещения (Bias Error) возникает из-за неверных или упрощенных предположений в алгоритме обучения – ее можно оценить на тренировочных данных

Ошибка дисперсии (Variance Error) возникает из-за чувствительности модели к новым данным – ее можно оценить на валидационных и тестовых данных

При этом редко удается найти компромисс, чтобы получить в рамках потерь модели низкое смещение и низкую дисперсию – обычно в случае улучшения одной составляющей – начинает страдать и повышаться другая составляющая,

За разработчиком модели стоит выбор того, на оптимизации и уменьшении каких ошибок сделать упор – обычно это диктует сама задача

Но модель, у которой были получены высокие ошибки и смещения, и дисперсии – не подходит для имеющихся данных и не должна быть использована в предсказании по реальным данным.