ПИШ 2023. Блок 2 «Машинное обучение». Вопросы аттестации

Сидоров Данил Михайлович

1. **Объясните, как вычисляется качество модели с задачей классификации: назовите основные и вторичные метрики, приведите пример.**

Качество модели оценивается на основе метрик - специальных расчетных параметров, внешних параметров, оценивающих качество предсказаний модели. Метрики бывают двух видов: онлайн и офлайн.

Для описания метрик необходимо внести понятие матрицы ошибок - она состоит из четырех блоков:

1. True Positive - верное предсказание True значения моделью,

2. False Positive - неверное предсказание False моделью,

3. True Negative - верное предсказание False моделью,

4. False Negative - неверное предсказание True моделью.

Основные и второстепенные метрики:

\* Accuracy - доля правильных классификаций

(TP+TN)/(TP+FP+FN+TN)

\* Error Rate - количество ошибок, деленое на все наблюдения

(FP+FN)/(TP+FP+FN+TN)

\* Precision - точность, доля TP во всех предсказанных Positive

TP/(TP+FP)

\* Recall - полнота, доля TP во всех реальных Positive

TP/(TP+FN)

\* F1 (объединение точности и полноты, среднее гармоническое) -

F1 = (2\* recall \* precision) / (recall + precision) = TP / (TP + (FP + FN)/2)

\* Area Under Curve - площадь под кривой,

1. **Что такое тестовый набор данных (test set) и для чего он нужен?**

Тестовый набор данных содержит входные данные и ожидаемые результаты, которые не использовались для обучения модели, соответственно, могут использоваться для оценки качества модели

1. **Каково назначение валидационного множества данных (validation set)?**

Валидационное множество нужно для оптимизации гиперпараметров

1. **В чем разница между параметром модели и гиперпараметром алгоритма обучения? приведите примеры.**

Параметр модели - это переменная выбранной модели, которая может меняться при обучении

Гиперпараметр модели - это параметр, значение которого устанавливается перед началом обучения модели

Например, для нейронной сети гиперпараметром будет являться количество слоев скрытых нейронов, а параметром будут являться веса между нейронами

1. **Что может пойти не так, если вы подгоните гиперпараметры модели, на тестовом наборе?**

Может возникнуть переобучение модели, в связи с чем на обучающих и тестовых данных модель будет идеально предсказывать результаты, а на реальных данных будет малоэффективна

1. **Что такое перекрестная проверка (cross-validation) и чем она лучше использования одного тестового множества?**

Перекрестная проверка - метод оценки модели, который подразумевает деление тестового множества на равномерные тестовые и обучающие части, после чего происходит обучение и оценка качества на тестовом куске данных, в конечном итоге берется средняя оценка метрики. Например, есть 6 точек, для трех фолдов (folds) проверка будет происходит следующим образом:

Первые две точки берутся как тестовые данные, остальные как обучающие, формируется модель, рассчитывается выбранная метрика. Тоже самое происходит для точек 3,4 и последних двух. Полученные три значения метрики формируют средний показатель.

1. **Что такое онлайновый режим работы системы машинного обучения?**

Онлайновый режим обучения - режим обучения модели, в котором данные поступают непрерывно и постоянно используются для повторного обучения модели, в отличии от систем, где обучение происходит один раз на обучающих данных, и полученная модель используется без изменений

1. **Опишите классификатор по методу ближайших соседей (kNN): в чем заключаются режимы его обучения и предсказания (вывода)?**

Классификатор по методу KNN - модель, которая производит классификацию на основе классов ближайших соседей. При любом предсказании повторно обучается на всех точках, что влечет низкую производительность

1. **Опишите модель линейной регрессии: в чем заключаются режимы ее обучения и предсказания (вывода)?**

Линейная регрессия — это алгоритм машинного обучения, который используется для проведения регрессионного анализа путем оценки коэффициента полиноминального уравнения, при степени полинома, близкой к количеству точек происходит переобучение - модель слишком затачивается под данные обучающего набора

1. **В чем заключается разница между обучением с учителем и обучением без привлечения учителя? какие ещё есть «промежуточные» типы обучения.**

При обучении с учителем обучающий набор изначально размечен и используется для обучения, т.е. присутствует влияние человека, при обучении без учителя влияние человека отсутствует, т.е. модель сама находит закономерность в данных и создает прогноз. Из промежуточных типов можно выделить обучение под полунаблюдением, когда не вся часть данных размечена учителем

1. **Что такое переобучение?**

Явление, когда модель слишком затачивается делать предсказания на основе обучающего множества и на реальных данных оказывается малоэффективной

1. **Опишите кратко метод работы дерева решений и ансамбля таких деревьев в режиме взаимного усиления (мы знаем его как «бустинг»).**

Дерево решений - метод машинного обучения, в котором логические правила выстроены в виде иерархического дерева. Бустинг - объединение нескольких деревьев решений в последовательную структуру, в которой каждое следующее дерево старается уменьшить ошибку предыдущего