**Вопросы к лабораторной работе №7**

1. Что означает уменьшение размерности в машинном обучении?

Уменьшение размерности в машинном обучении означает уменьшение количества признаков (функций), используемых для описания данных. Это может быть полезно для упрощения моделей, снижения вычислительной сложности, избежания переобучения и улучшения интерпретируемости результатов.

1. Какие методы входят в Feature Selection? Как они работают?

*Фильтр с низкой дисперсией (VarianceThreshold)*

Например, если переменные какого-либо параметра не изменяются (т.е. параметр имеет нулевую дисперсию), то этот параметр никак не влияет на модель. Поэтому, можно рассчитать дисперсию каждого параметра в выборке, а затем отбросить параметры, имеющие низкую дисперсию по сравнению с другими параметрами в наборе данных.

*Фильтр высокой корреляции*

Высокая корреляция между двумя переменными означает, что они имеют схожие тенденции и, вероятно, несут схожую информацию. Это может резко снизить производительность некоторых моделей (например, моделей линейной и логистической регрессии). Можно вычислить корреляцию между независимыми числовыми переменными. Если коэффициент корреляции пересекает определенное пороговое значение (как правило 0,5–0,6), мы можем отбросить одну из переменных (удаление переменной очень субъективно и всегда должно производиться с учетом предметной области).

*Выбор функции с помощью SelectFromModel*

SelectFromModel – это мета-преобразователь, который можно использовать вместе с любым оценщиком, который присваивает важность каждой функции (например, деревья решений). Функции считаются неважными и удаляются, если соответствующая важность значений функций ниже указанного threshold параметра.

*Рекурсивное устранение признаков*

При наличии внешнего оценщика, который присваивает веса характеристикам (например, коэффициентам линейной модели), цель исключения рекурсивных признаков (RFE) состоит в том, чтобы выбрать признаки путем рекурсивного рассмотрения все меньших и меньших наборов признаков. Оценщик обучается на начальном наборе признаков, и определяется важность каждого признака. Затем наименее важные признаки удаляются из текущего набора признаков. Эта процедура рекурсивно повторяется для сокращенного набора, пока в конечном итоге не будет достигнуто желаемое количество признаков.

1. Расскажите принцип работы метода PCA.

Метод главных компонент (PCA) – это метод уменьшения размерности данных, который находит линейные комбинации исходных признаков (главные компоненты), которые содержат максимальную дисперсию данных. PCA проецирует данные на новые оси (главные компоненты) таким образом, чтобы первая главная компонента содержала наибольшую дисперсию, вторая - следующую по величине, и так далее. Это позволяет сократить размерность данных, сохраняя при этом как можно больше информации.

1. Что означает понятие главная компонента?

Главная компонента – это линейная комбинация исходных признаков в данных, которая содержит наибольшую дисперсию информации. Она является основной особенностью в методе главных компонент (PCA) и используется для уменьшения размерности данных.

1. Что означает термин Ансамбли в контексте машинного обучения?

Ансамбли в контексте машинного обучения означают комбинирование нескольких моделей для получения лучшего прогноза или классификации, чем каждая модель по отдельности. Это может быть достигнуто путем усреднения прогнозов отдельных моделей или использования их в качестве базовых моделей для более сложных алгоритмов.

1. Как работают алгоритмы стекинг, бэггинг, бустинг?

Стекинг (stacking) – это метод комбинации нескольких моделей, в котором прогнозы каждой модели используются в качестве входных данных для метамодели. Метамодель обучается на этих прогнозах для получения окончательного прогноза.

Бэггинг (bagging) – это метод, при котором несколько одинаковых моделей обучаются на разных подмножествах обучающих данных. Затем прогнозы от каждой модели усредняются для получения итогового прогноза.

Бустинг (boosting) – это метод, при котором модели обучаются последовательно, причем каждая следующая модель фокусируется на исправлении ошибок предыдущей модели. Прогнозы от всех моделей усиливаются для получения итогового прогноза.

1. Объясните метод Random Forest. Какой алгоритм асамблирования в нем используется.

Random Forest – это ансамблевый метод машинного обучения, основанный на использовании деревьев решений. Он создает ансамбль деревьев решений во время обучения и усредняет их прогнозы для получения более точного и устойчивого прогноза. В Random Forest для построения каждого дерева используется подвыборка обучающих данных и подвыборка признаков (bootstrap sampling), а также случайный выбор признака на каждом разделении дерева.