2 Глава

Порцию информации, передаваемой в систему МО, часто называют **сигналом**, ссылаясь на теорию информации Шеннона: вам нужно высокое соотношение сигнал/шум

Выбор критерия качества работы.

Типичный критерий качества для задач регрессии **-квадратный корень из среднеквадратической ошибки** (Root Meaп Squared Error - RMSE). Она дает представление о том, насколько большую ошибку система обычно допускает в своих прогнозах, с более высоким весом для крупных ошибок.

- m -количество образцов в наборе данных, на которых измеряется ошибка RMSE.
- x(i) - вектор всех значений признаков (исключая метку) i-го образца в наборе данных,
- y(i) - его метка (желательное выходное значение для данного образца).
- X -матрица, содержащая все значения признаков (исключая метки) всех образцов в наборе данных. Предусмотрена одна строка на образец, а i-тая строка эквивалентна транспонированию x(i, что обозначается как x(i)T.
- **h -функция прогнозирования системы**, также называемая **гипотезой** (**hypothesis**). Когда системе предоставляется вектор признаков образца x(i), она выводит для этого образца прогнозируемое значение y^(i) = h(x(i)).
- RMSE(X,h) -функция издержек, измеренная на наборе образцов с использованием гипотезы h.

Вы можете обдумать применение **средней абсолютной ошибки** (**Meaп Absolute Error - MAE**), также называемой средним абсолютным отклонением (average absolute deviation), которая показана в уравнении 2.2.

Показатели RMSE и MAE представляют собой способы измерения расстояния между двумя векторами: вектором прогнозов и вектором целевых значений. Существуют разнообразные меры расстояния, или **нормы.**

Чем выше индекс нормы, тем больше она концентрируется на крупных значениях и пренебрегает мелкими значениями.

data.describe():

std показывает стандартное отклонение (standard deviation), которое измеряет разброс значений.

Это называется **стратифицированной выборкой (stratified sampling**): население делится на однородные подгруппы, называемые **стратами** (**strata**)

Проект Scikit-Learn:

- **Оценщики(estimator).** Любой объект, который может проводить оценку параметров на основе набора данных, называется оценщиком (например, imputer является оценщиком). Сама оценка производится с помощью метода f i t ( ) , принимающего в качестве параметра единственный набор данных (или два для

алгоритмов обучения с учителем; второй набор данных содержит метки). Любой другой параметр, необходимый для управления процессом оценки, считается гиперпараметром (вроде strategy в imputer) и должен быть указан как переменная экземпляра (обычно через параметр конструктора).

- **Трансформаторы** (transformer). Трансформация выполняется методом transform () , которому в параметре передается набор данных, подлежащий трансформации. Он возвращает трансформированный набор данных. Трансформация обычно полагается на изученные параметры. Все трансформаторы также имеют удобный метод по имени fit _ transform (), который представляет собой эквивалент вызова fit ( ) и затем transform () (но благодаря оптимизации метод fit _ transform () временами выполняется намного быстрее).
- **Прогнозаторы** (predictor). Наконец, некоторые оценщики способны вырабатывать прогнозы, имея набор данных; они называются прогнозаторами. Прогнозатор располагает методом *predict()*, который принимает набор данных с новыми образцами и возвращает набор данных с соответствующими прогнозами. Прогнозатор также имеет метод *score()*, измеряющий качество прогнозов с помощью указанного испытательного набора (и соответствующих меток в случае алгоритмов обучения с учителем)
- **Инспектирование**. Все гиперпараметры прогнозаторов доступны напрямую через переменные экземпляра (например, imputer. strategy), и все изученные параметры прогнозаторов также доступны через открытые переменные экземпляра с суффиксом в виде подчеркивания (например, imputer. statistics)
- **Нераспространение классов**. Наборы данных представляются как массивы NumPy или разреженные матрицы SciPy вместо самодельных классов. Гиперпараметры - это просто обычные строки или числа Python
- **Композиция**. Существующие строительные блоки максимально возможно используются повторно. Например, как будет показано далее, из произвольной последовательности трансформаторов легко создать прогнозатор Pipeline, за которым находится финальный прогнозатор.
- Разумные стандартные значения. Scikit-Learn предоставляет обоснованные стандартные значения для большинства параметров, облегчая быстрое создание базовой рабочей системы.

Pandas-метод ***factorize()***, который сопоставляет каждую категорию с отличающимся целым числом.