Собеседование Data scientist (топ 100 вопросов)

Q1:

Перечислите основные шаги подготовки данных перед использованием алгоритмов машинного обучения?

A1:

1. Загрузка данных
2. Проверка типов(.dtypes) + конвертация
3. Проверка дубликатов строк (.duplicated())
4. Обработка признаков в соответствии с типами
   4.1. Числа: поиск аномалий (q25-3iqr, q75+3iqr), корреляции между признаками, корреляция с целевым признаком
   4.2. Категории: выявление категорий малого размера
   4.3. Строки: приведение к нижнему регистру, удаление пунктуации, удаление стоп-слов, приведение слов к начальной форме
5. Оценка распределений (тест колмогорова-смирнова, шапиро-вилкоксона, скошенность)
   5.1. Построение гистограмм, кде (pd.hist(), sns.kdeplot()) признаков (+признаков в разрезе целевого признака)
   5.2. Проведение статистических тестов для различных категорий признаков (бутстреп→тест стьюдента для среднего/доли)
6. Подготовка данных для модели
   6.1. Если модели линейные, то масштабирование + индикатор пропуска в строке
   6.2. Если деревья, заполнение пропусков аномальным значением (некоторые деревья могут работать с np.nan(xgboost, lighgbm, …))

Q2: Перечислите основные методы работы с дисбалансом классов в задаче бинарной класификации

A2:

1. Выбор подходящей метрики (precision, recall, f_score, roc auc)
2. Апсемплинг(выбор строк датасета с замещением, сохраняя идентичное распределение целевой пременной), даунсемлпинг(выбор строк датасета сохраняя идентичное распределение целевой пременной)

Q3: В чем разница между ящиком с усами и гистограммой?

Ящик с усами оценивает распределение точечно: разбиваем данные на процентили: 25%, 50%, 75% (границы ящика), отмечаем усы (Q25-1.5*(Q75-Q25), Q75+1.5*(Q75-Q25))

Гистограмма оценивает распределение интервально: выбираем число корзин, размещаем данные по корзинам.

Q3: Перечислите различия между l1 и l2 регуляризацией

A3:

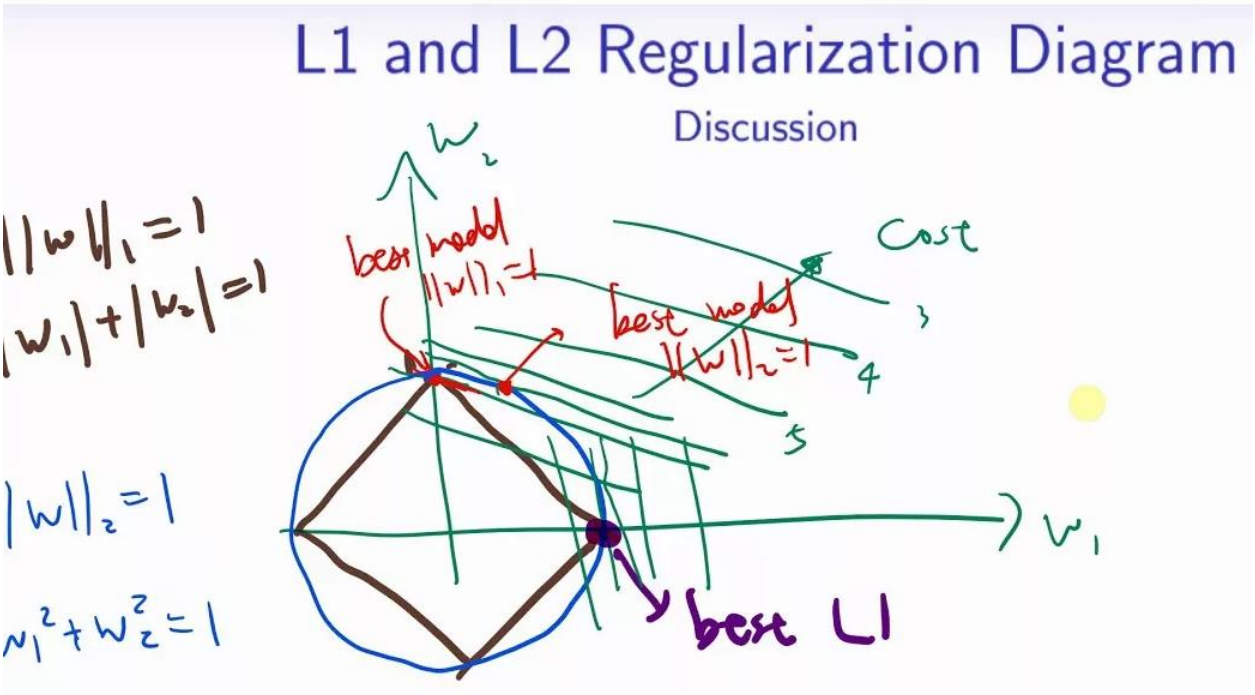
Регуляризация – наложение дополнительных ограничений на модель с целью ее упрощения и борьбы с переобучением

L1(зануляет к-ты), L2 регуляризации:

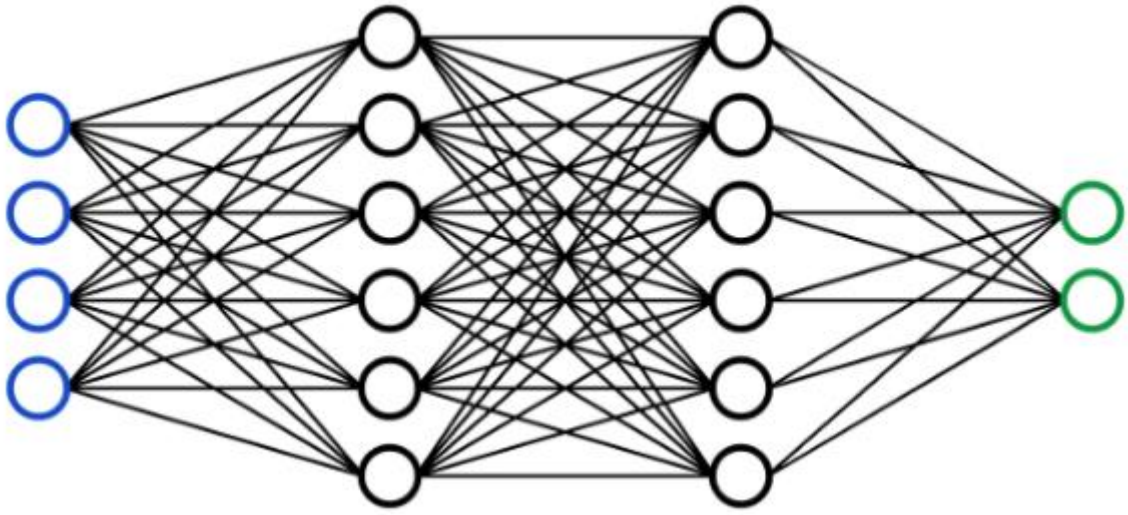
$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$

$$\sum_{i=1}^{n}(Y_i - \sum_{j=1}^{p} X_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Геометрический смысл:



Q4: Объясните основной принцип работы нейросетей



Нейронная сеть – это способ ансамблирования линейных моделей. элементарным объектом является линейная регрессия(+функция активации). Ансамбль состоит из слоев, каждый слой состоит из элементарных моделей. Обучение нейронной сети – оптимизация весовых к-ов при нейронах каждого слоя. Оптимизация весов осуществляется с помощью алгоритма обратного распространения ошибки