Курс “Основы машинного обучения”

Вопросы к контрольной работе

1. **Что такое объекты и признаки в машинном обучении? Для чего нужен функционал ошибки? Что такое алгоритм (модель)?**

Объекты – абстрактные сущности, набор признаков, то, что мы используем в процессе обучения и для чего делаем предсказания.

Признаки – характеристики объектов.

Модель – функция, которая предсказывает ответ на основе набора признаков.

Функционал ошибки – метрика качества работы алгоритма на выборке. В процессе обучения мы ищем оптимальный алгоритм с точки зрения функционала ошибки.

1. **Чем задача классификации отличается от задачи регрессии?**

Регрессия предсказывает вещественные ответы (их множество бесконечно).

Классификация предсказывает принадлежность объекта к какому-либо классу (которых конечное число).

1. **Приведите примеры задач классификации и регрессии.**

Классификация: кредитный скоринг, определение эмоционального окраса текста.

Регрессия: определение стоимости квартиры, предсказание температуры воздуха.

1. **Что такое вещественные (числовые), бинарные, категориальные признаки? Приведите примеры.**

Числовые – вещественные числа (площадь квартиры).

Бинарные – 0 или 1 (есть ли балкон в квартире).

Категориальные – значения из конечного неупорядоченного множества (цвет стен в подъезде).

1. **В чём заключается обобщающая способность алгоритма машинного обучения? К чему приводит её отсутствие? Что такое переобучение?**

Обобщающая способность – способность алгоритма выдавать правильные ответы не только для объектов, на котором он обучался, но и для новых, неизвестных ранее объектов.

Отсутствие обобщающей способности приводит к низкому качеству правильных ответов на новых данных.

Переобучение – свойство модели показывать хороший результат на обучающих данных, но при этом показывать плохой результат на новых данных.

1. **Что такое гиперпараметр? Чем гиперпараметры отличаются от обычных параметров алгоритмов? Приведите примеры параметров и гиперпараметров в линейных моделях.**

Гиперпараметр – параметр, который подбирается не в процессе обучения, а чаще всего на отложенной выборке (при обучении GD – шаг, размер батча. Для модели с регуляризацией – коэффициент штрафа за большие веса).

Параметры (обычные) подбираются в процессе обучения (веса линейной модели).

1. **Что такое отложенная выборка? Что такое кросс-валидация (скользящий контроль)? Как ими пользоваться для выбора гиперпараметров?**

Отложенная выборка – часть обучающей выборки, на которой подбираются гиперпараметры и которая не учувствует в процессе непосредственно обучения. Ищется гиперпараметр, при котором ошибка на отложенной выборке минимальна.

Кросс-валидация – обучающая выборка разбивается на непересекающиеся равные части, обучается на всех, кроме одной из этих частей, и на оставшейся проводит тест. Ищется гиперпараметр, который имеет наименьшую среднюю ошибку по всем блокам.

1. **Как метод k ближайших соседей определяет класс для нового объекта?**

Вычисляем расстояние от нового объекта до каждого из объектов обучающей выборки, сортируем их в порядке возрастания расстояния, присваиваем новому объекту тот класс, что встретился чаще всех среди k первых (после сортировки) объектов обучающей выборки.

1. **Опишите метод k ближайших соседей с парзеновским окном. Какие в нём есть параметры?**

Берем k соседей. Даем тот класс, сумма весов больше. Вес считается как

Ядро от (расстояния до соседа / h). h - ширина окна.

Гиперпараметр – ширина окна.

1. **Запишите формулу метода kNN для регрессии.**

Вычисляем расстояние от нового объекта до каждого из объектов обучающей выборки, сортируем их в порядке возрастания расстояния. Ответ для нового объекта – усредненный ответ первых k объектов (после сортировки).

1. **Что такое градиент? Какое его свойство используется в машинном обучении?**

Градиент — вектор частных производных

Если двигаться в сторону антиградиента, то мы можем найти минимум функции потерь. Так как MSE обычно выпуклая функция, то минимум будет один и антиградиент нас будет вести именно в него.

1. **Опишите алгоритм градиентного спуска.**

Начинаем спуск со случайного набора весов, движемся в сторону антиградиента (усредненная сумма градиентов для каждого из объектов), повторяем пока не окажемся в минимуме (как вариант, разница нормы вектора текущего и предыдущего шагов < epsilon).

1. **Что такое стохастический градиентный спуск? В чём его отличия от обычного градиентного спуска? Какие у него плюсы и минусы?**

Это тот же градиентный спуск, только в нем при вычислении градиента используются не все объекты обучающей выборки, а только какая-то ее часть.

SGD сходится как правило на большей итерации, и менее вероятно, чем GD, но вычислительная сложность одной итерации в нем меньше.

1. **Как обучается линейная регрессия?**

Обучение линейной регрессии – подбор такого набора весов, при котором выбранная функция потерь для объектов тренировочной выборки будет минимальна.

1. **Что такое регуляризация? Как она помогает бороться с переобучением?**

Регуляризация – способ борьбы с переобучением.

Регуляризованная модель штрафует себя за слишком большие веса, тем самым некоторые неважные признаки будут иметь нулевой вес.

1. **Чем L1-регуляризация отличается от L2-регуляризации?**

При L2-регуляризации к функции потерь в модели добавляется слагаемое lambda \* норма вектора весов.

При L1-регуляризации к функции потерь в модели добавляется слагаемое lambda \* (сумма абсолютных значений весов).

L2 сильнее L1.

1. **Что такое масштабирование признаков? Как его проводить? Зачем это нужно?**

Масштабирование признаков – приведение признаков к одинаковому масштабу (порядку).

Например: вычтем из каждого признака среднее и поделим на корень из его дисперсии.

Нужно для улучшения интерпретации линейной модели (чем больше вес, тем важнее признак работает только после масштабирования). Зачастую масштабирование улучшает общее качество модели (особенно, если до масштабирования признаки отличались на порядки).

1. **Чем функционал MSE отличается от MAE? Что такое функция потерь Хубера и для чего она нужна?**

MAE в отличие от MSE мягче относится к выбросам (из-за того, что квадрат в функции заменен на модуль).

Функция потерь Хубера позволяет штрафовать за большие отклонения слабо, а за небольшие – сильнее. Это позволяет снизить влияние выбросов, при это оставив MSE для не выбросов.

1. **Как выглядит модель линейной классификации в случае двух классов?**

a(x) = sign(<w, x> - t)

t - порог

1. **Что такое отступ? Для чего он нужен?**

M\_i = y\_i \* <w, x\_i>

То есть отступ отрицателен при неверном ответе, положителен при верном ответе.

Отступ больше при большей уверенности.

1. **Как обучаются линейные классификаторы (общая схема с верхними оценками)?**

Делаем верхнюю оценку (дифференцируемой функцией) для бинарной функции потерь и минимизируем ее (тем же GD как вариант).

1. **Для чего может понадобиться оценивать вероятности классов?**

Например, мы выдаем кредиты. Нам важно не только узнать решение модели (вернется ли кредит), но и насколько она уверена, ведь мы не хотим выдавать кредит людям, в которых все же не очень уверены (чистая убыль). Оценка вероятности как раз покажет уверенность модели, и, учитывая эту оценку вероятности, мы можем корректировать свое решение, например, рассмотрев кандидата лично.

Пример поинтереснее: хотим посчитать потенциальное мат. ожидание для прибыли в случае покупки (вероятность покупки\* прибыль за сделку).

1. **Как обучается логистическая регрессия? Запишите функционал и объясните, откуда он берётся.**

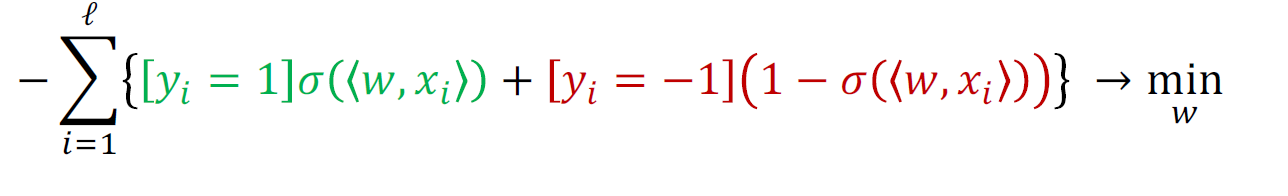
Изображение выглядит как текст

Автоматически созданное описание

Мы же используем линейную модель, а линейная модель — это прежде всего <w, x>.

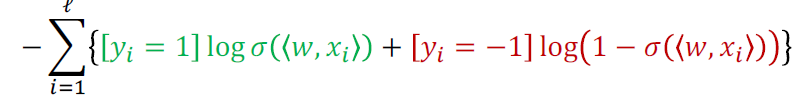
Мы требуем от логистической функции потерь, чтобы она предсказывала вероятности, но вероятность – число в [0; 1]. А скалярное произведение не всегда попадает в этот интервал, поэтому заключим его в сигмоиду.

Хотим, чтобы для положительного класса сигмоида давала близкое к 1, а для отрицательного – к 0.

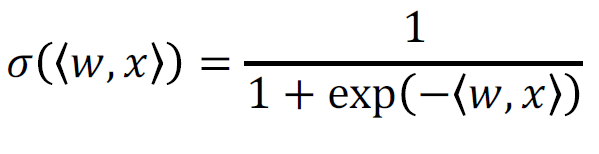


Но штрафы сейчас для всех ошибок одинаковые, хотим строже.

Поэтому завернем это дело в логарифм и сделаем преобразования.



1. **Как в логистической регрессии строится прогноз для нового объекта?**



1. **Что такое метод опорных векторов? Опишите его основную идею.**

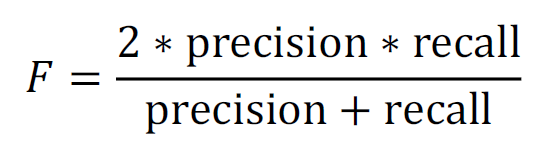
Это метод линейной классификации, который в процессе обучения стремится провести через объекты гиперплоскость так, чтобы максимизировать отступ для всех объектов.

1. **Как устроены метрики accuracy, precision, recall? Что такое F-мера? Чем она лучше арифметического среднего точности и полноты?**

Accuracy – доля верно предсказанных ответов.

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)



Арифметическое среднее говорит, что алгоритм неплохой, если даже одна из метрик почти ноль, но другая высокая. F-мера лишена этого недостатка.

1. **Для чего нужны ROC- и PR-кривые? Как они строятся? Что такое AUC-ROC и AUC-PRC?**

С их помощью можно оценивать качество модели в целом, не привязываясь к конкретному порогу t.

По осям откладывается precision – recall и ставятся точки для всех возможных значений порога.

По осям откладываются FPR – TPR и ставятся точки для всех возможных значений порога.

AUC-ROC – площадь под ROC-кривой, чем она ближе к 1, тем лучше модель.

AUC-PRC – площадь под PR-кривой, чем она ближе к 1, тем лучше модель.

1. **Какие метрики можно использовать в случае сильного дисбаланса классов?**

F-мера, AUC-ROC, AUC-PRC.

1. **Как можно свести задачу многоклассовой классификации к серии задач бинарной классификации?**

ONE-VS-ALL

Если есть k классов, то обучим модель k раз, причем каждый раз будем обучать на объектах одного конкретного класса (так для каждого класса получим свою модель). Предскажем класс для нового объекта каждой моделью. Там, где будет выше оценка принадлежности, тот класс и выберем.

ALL-VS-ALL

Если есть k классов, то сделаем модель для каждой пары классов, которая будем говорить к какому из двух классов объект принадлежит.

Новый объект проведем через все модели, тот класс k, что победил больше всех раз и предскажем.

1. **Что такое решающее дерево? Как оно строит прогноз для объекта? Как обучаются решающие деревья в задачах классификации и регрессии (и что такое критерии информативности)?**

Это модель, которая строит прогноз на основе предикатов.

Прогноз строится путем разбития пространства объектов на зоны, путем построения вершин на основе предикатов. Попадания в ту или иную зону означает принадлежность к тому или иному классу.

Лист – вершина, дающая константный прогноз (не содержащая предиката). В задаче регрессии ответ для объекта – среднее значение ответов в листовой вершине.

Критерий информативности – функция, позволяющая найти оптимальное разбиение узла на две части.

Изображение выглядит как текст

Автоматически созданное описание

1. **Какие вы знаете критерии останова и способы выбора значений в листьях? Какие гиперпараметры имеются у деревьев?**

Останов: глубина, число листьев, число объектов в листе.

Обычно значения в листе выбираем жадно, выбирая разбиение с наименьшей суммарной энтропией / дисперсией.

Гиперпараметры деревьев схожи с критериями останова.

1. **Что такое бэггинг и метод случайных подмножеств? Что такое случайный лес, как он обучается и как он строит прогнозы?**

В бэггинге берутся n базовых моделей, обучаются независимо на какой-то части выборки, причем подмножество обучается с помощью бутстрапа: берутся k элементов с повторениями. После чего делается композиция (усреднение или голосование большинством).

Случайный лес - при построении ищем лучший предикат среди случайных q признаков, генерируем бутстрапом подвыборку для n деревьев и берется композиция.

1. **В чём идея разложения ошибки на смещение и разброс? Как бэггинг меняет смещение и разброс одной модели?**

Бэггинг не влияет на смещение, но уменьшает разброс (если модели независимы, то в n раз).

1. **Опишите идею градиентного бустинга для среднеквадратичной ошибки. Запишите задачу для обучения очередной базовой модели.**

Каждая последующая базовая модель старается скорректировать ошибку композиции всех предыдущих моделей, то есть ее ответы должны стремиться обнулить ошибку композиции предыдущих моделей.

1. **Опишите идею градиентного бустинга для произвольной функции потерь. Запишите задачу для обучения очередной базовой модели.**

Бустинг состоит из 3 этапов:

* Возьмём простые базовые модели
* Будем строить композицию последовательно и жадно
* Каждая следующая модель будет строиться так, чтобы максимально корректировать ошибки построенных моделей

Text

Description automatically generated with medium confidence

Diagram, text

Description automatically generated

Изображение выглядит как текст

Автоматически созданное описание

1. **Какими способами можно побороть переобучение модели:**
   1. **в общем случае (независимо от модели);**
   2. **если модель линейная;**
   3. **если модель — градиентный бустинг.**
2. Общий случай

* Увеличение количество данных в наборе
* Уменьшение количества параметров в модели

1. Линейная модель

* Регуляризация

1. Градиентный бустинг