Вопросы госэкзамена по курсу ММО

А. Майсурадзе, Н. Багров, 2017

# Формулировки в официальных билетах

14. Средние и эмпирические операционные характеристики стратегий распознавания (классификаторов, регрессий). Проблема переобучения. Проблема устойчивости решений. Роль обучающей, валидационной и контрольной выборок при построении распознающей системы. Скользящий контроль (кросс-валидация). Регуляризация на примере линейной регрессии.
15. Ансамбли классификаторов. Основные этапы работы типичного базового классификатора, возможность коррекции на разных этапах. Бэггинг и случайные подпространства. Бустинг. Случайный лес как композиция основных подходов к построению ансамбля.
16. Задача кластеризации как фундаментальная задача интеллектуального анализа данных, сопоставление с операцией группирования и задачей классификации. Различные постановки: разбиение, стохастическая, нечёткая, иерархическая, упорядочивание, однокластерная (последовательная). Примеры методов кластеризации для разных постановок.

# Общие положения

Анализ данных - это широкий комплекс мероприятий по сбору, накоплению и преобразования информации, помогающий в итоге принять решение или даже позволяющий автоматизировать принятие решения. С точки зрения теории управления анализ данных позволяет контролировать, оптимизировать и автоматизировать бизнес-процессы. Анализ данных может восприниматься как:

* бизнес (мы берём деньги за анализ данных);
* часть бизнес-модели (функционирование нашего бизнеса опирается на анализ данных);
* технологии (аппаратно-программные решения, организация процессов и мероприятий);
* аналитические задачи (работа на ЛПР);
* наука (теория моделей и методов).

В анализе данных принято говорить о моделях данных и информационных моделях.

* Модели данных позволяют представить информацию о моделируемой системе в некоторый момент времени или за всё время до некоторого момента времени. Это именно статическая информация, хотя она и может отражать ретроспективу моделируемой системы. Интуитивно модель данных можно воспринимать как способ “хранить фотографию” системы.
* Информационные модели отвечают за процессы преобразования информации в системе. Информационные модели обычно являются эмпирическими, т. е. не делается утверждений, что информационные модели отражают фактические механизмы функционирования моделируемой системы.

Статистическое распознавание - это наука, в которой информационные модели являются стохастическими и их настройка идёт при известных распределениях.

Машинное обучение занимается настройкой стохастических информационных моделей при неизвестных распределениях по наблюдаемым данным. Вообще говоря, восстановление распределения по наблюдениям - это генеральная задача статистики, соответственно, невозможно чётко разделить статистику и машинное обучение.

Иногда пытаются говорить о нестатистическом машинном обучении. Проблема в том, что мы хотим решать задачи не только для уже известных нам наблюдений, но и для всей генеральной совокупности. Кроме статистики нет общераспространённых теорий, позволяющих переходить от наблюдений к генеральной совокупности.

Основной приём в анализе данных (статистике, машинном обучении) - сведение содержательной задачи (пожеланий бизнеса) к набору формальных задач из некоторого “стандартного набора”. Такие стандартные семейства задач анализа данных принято называть:

* инструменты статистики;
* фундаментальные задачи интеллектуального анализа данных.

На языке таких стандартных задач - хотя и с разным уровнем формализации - могут вести беседу инженеры, отраслевые исследователи и прикладные ученые.

Упомянутые задачи принято характеризовать в соответствии со следующими свойствами.

* Модель системы содержит объекты одного типа (гомогенные системы) или нескольких разных типов (гетерогенные системы)
* Все части описания объекта имею одну роль или делятся на наблюдаемое описание и скрытое состояние.
* Какие части описания доступны при настройке и при использовании информационной модели.
* Какие именно типы значений принимают части описания.

При этом часто оказывается, что одно и то же название получает последовательность задач.

**Задачи распознавания состояния объектов**. Система гомогенная. Объекты имеют наблюдаемое описание x и скрытое состояние y. Есть некоторое множество действий A. Есть некоторый критерий качества Q. Требуется найти отображение q из множества наблюдаемых описаний X в множество действий A, доставляющий оптимум критерию Q. В общем случае отображение можно назвать стратегией распознавания.

В частности, если в задаче распознавания множество скрытых состояний Y известно заранее, дискретное и каждое скрытое состояние интерпретируется как метка класса, а множество действий A включает в себя множество состояний Y, то это **задача классификации**. Отображение q принято называть классификатором.

В частности, если в задаче распознавания множество скрытых состояний Y известно заранее, недискретное и каждое скрытое состояние интерпретируется как арифметическое значение, а множество действий A включает в себя множество состояний Y, то **это задача восстановления регрессии**. Отображение q принято называть регрессией.

**Байесовская задача распознавания**. Это частный случай задачи распознавания. Дана полная теоретико-вероятностная информация о наблюдаемых описаниях и скрытых состояниях, т. е. существует и известно совместное распределение p(x,y). Критерий качества строится следующим образом. Дана функция штрафа L(x,y,a). Множество значений штрафов позволяет вычислять матожидание. Средний риск произвольной стратегии q определяется как $R(q)=\int\_{X×Y}^{}L(x,y,q(x))p(x,y)dxdy$, т. е. это матожидание штрафа. Требуется найти стратегию, минимизирующую средний риск. Такая стратегия называется байесовской. Она по определению имеет минимальный средний риск.

Машинное обучение начинается там, где совместное распределение p(x,y) неизвестно и приходится как-то выкручиваться. Например, **задача обучения по прецедентам** получается из байесовской задачи распознавания, если вместо p(x,y) дана выборка, извлеченная из p(x,y).

Вопрос 1

Средние и эмпирические операционные характеристики стратегий распознавания (классификаторов, регрессий). Проблема переобучения. Проблема устойчивости решений. Роль обучающей, валидационной и контрольной выборок при построении распознающей системы. Роль валидации при минимизации эмпирического риска. Роль валидации при выборе информационной модели и метода обучения. Скользящий контроль (кросс-валидация). Регуляризация на примере линейной регрессии: ridge, lasso, elastic net. Регуляризация на примере линейной классификации: SVM. Роль валидации при регуляризации.

В математической статистике и статистическом распознавании требуется

Проблема переобучения проявляется в высокой точности работы метода машинного обучения на тренировочной выборке и низкой на тестовой. Обычно такой эффект возникает из-за выбора модели с слишком большим числом свободных параметров.

При обучении метода машинного обучения обычно непосредственно используют обучающую (тренировочную) и валидационную выборки. При этом на обучающей выборке производится настройка внутренних параметров модели (например, весов в методе опорных векторов), а на валидационной внешних (например, коэффициент C в SVM). Тестовая выборка используется непосредственно для оценки качества получившейся модели и эффекта переобучения.

Скользящий контроль может использоваться для оценки переобучения модели на обучающей выборке. При этом часть наблюдений из обучающей выборки используется непосредственно для обучения, а часть для оценки качества модели. Проводятся ряд таких экспериментов (обучение и оценка качества), в результате которых можно вычислить среднюю ошибку. Также такой подход (разбиение обучающей выборки) может использоваться для подбора различных внешних параметров моделей (например, подбор константы C в методе опорных векторов).

При обучении линейных классификаторов на примере SVM возникает задача определения параметров $ξ\_{1},...,ξ\_{n}$(ошибки на классах). Если их дополнительно не ограничить, то они могут бесконечно возрастать (что, в частности, может привести к переобучению классификатора). Для решения этой проблемы можно применить регуляризацию, например, добавив сумму этих величин в функционал для минимизации.

Задачу восстановления линейной регрессии по данным можно решать, например, методом наименьших квадратов. Если не использовать регуляризацию, то данное решение будет оптимально только на этих данных. Если ввести дополнительную регуляризацию - ограничение нормы весов W - то, это позволит уменьшить эффект переобучения (количество свободных параметров уменьшается, модель становится более простой). Виды регуляризаций - L1 (lasso), L2 (ridge), L1+L2 с весами (elastic net)

Вопрос 2

Операционные характеристики описаний объектов, информационных моделей и отдельных классификаторов. Средние и эмпирические. Без априорных вероятностей и с ними. ROC-анализ описаний и классификаторов, AUC, EER. Проблема группирования. Проблема редких событий. Отбор и генерация признаков на основе операционных характеристик.

ROC-анализ описаний и классификаторов используется для оценки доли ошибок первого и второго рода. Например, для детектора объектов можно построить кривую зависимости точности (чем меньше ложных обнаружений он сделал, тем выше точность) классификатора от полноты (чем больше объектов он нашел, тем больше полнота). Используя ROC анализ можно провести оценку точности алгоритма исходя из некоторых требований. Например, необходимо оценить точность классификатора, который будет находить 90% объектов.

AUC - Area Under Curve или площадь под кривой - метрика качества классификатора. Она показывает насколько оптимально данный классификатор работает сразу по двум параметрам - точности и полноте. Например, для идеального по полноте детектора объектов (который определяет любой из объектов как целевой) площадь под ROC кривой будет низкой из-за низкой точности классификатора (при условии, что выборка сбалансированная по классам).

Под проблемой редких событий понимают сложность обучения классификатора по событиям с сильно отличающейся вероятностью (или для несбалансированным по классам выборкам). Пример - классификация сломанных автодорожных знаков по данным с видеорегистраторов (в городах достаточно редко можно встретить сломанные знаки, отношение числа “нормальных” и сломанных знаков будет достаточно большое). Данную проблему обычно решают введением дополнительных весов для немногочисленных классов при обучении модели.

Вопрос 3

Ансамбли классификаторов. Основные этапы работы типичного базового классификатора, возможность коррекции на разных этапах. Бэггинг и случайные подпространства. Бустинг. Случайный лес как композиция основных подходов к построению ансамбля.

Под ансамблем классификаторов понимают некоторую функцию L(K1,...Kn), где $K\_{i}$- независимые классификаторы. При этом результат работы каждого из классификаторов вычисляется независимо, функция L строит результирующий классификатор. Например, это может быть линейная комбинация классификаторов или более сложная зависимость (нейронная сеть).

Бустингом называется итерационный процесс обучения коммитета классификаторов. После каждой итерации вычисляются веса независимых классификаторов исходя из ошибки на обучающей выборке (т.е при обучении больший вклад в классификатор будут давать сложные классы, на которых данный классификатор ошибается). После чего построенный алгоритм (который учитывает ошибки предыдущих) добавляется в множество алгоритмов и процесс продолжается.

Бэггинг - обучается множество классификаторов, финальное решение принимается голосованием. Обычно независимые классификаторы обучают на подпространствах обучающей выборки (например, на разных подмножествах вектора-признака).

Лес решающих деревьев - построим несколько решающих деревьев и будем принимать решение голосованием. Например, для задачи классификации будет засчитан тот класс, за который проголосовало больше деревьев.

Вопрос 4

Задача кластеризации как фундаментальная задача интеллектуального анализа данных, сопоставление с операцией группирования и задачей классификации. Различные постановки: разбиение, стохастическая, нечёткая, иерархическая, упорядочивание, однокластерная (последовательная). Связь с задачей разделения смеси распределений. Связь с задачей сокращения размерности описаний. Примеры методов кластеризации для разных постановок. Задачи кластеризации в гетерогенных системах: рекомендательные системы, коллаборативная фильтрация.