Вопросы к экзамену по ММРО

ВМиК, кафедра ММП, весна 2016.

Комментарий: вопросы делятся на “длинные” и “короткие”. При подготовке ответа на длинный вопрос можно использовать любые материалы. Подготовка ответа на короткий вопрос минимальна и при подготовке ответа на короткий вопрос ничем пользоваться нельзя. Вопросы разбиты по темам. Нужно рассказывать отдельный вопрос, а не всю тему целиком.

1 Длинные вопросы

1.1 Бустинг

1. Adaboost: алгоритм и его вывод.
2. Градиентный бустинг. Примеры функций потерь для регрессии, бинарной и множетсвенной классификации. Вариация для деревьев. Shrinkage и subsampling вариации.
3. Квадратичная аппроксимация в бустинге. Вывести непрерывный LogitBoost.
4. Подбор разбиений деревьев в методе xgBoost.

1.2 Нейросети

1. Определение нейросети. Виды инвариантных преобразований признаков для изображений. Подходы к адаптации алгоритма на инвариантность. Конволюционные нейросети.
2. Алгоритм обратного распространения ошибки: для чего он нужен и его вывод.

1.3 Отбор признаков

1. Фильрующие (filter) методы отбора признаков. Корреляция. Почему она измеряет не все виды зависимости? Расстояние Кульбака-Лейблера. Взаимная информация. Почему она измеряет любые отклонения от гипотезы независимости?
2. Фильрующие (filter) методы отбора признаков. Расчет значимости признаков с помощью relief критерия и с помощью деревьев.
3. Типы признаков и типы методов отбора признаков (filter, wrapper, embedded), с примерами алгоритмов каждого подхода.
4. Генерация подмножеств признаков на основе их индивидуальной релевантности. Метод последовательного поиска и его варианты.
5. Операции скрещивания и мутации. Генетический алгоритм перебора подмножеств признаков. Варианты его оптимизации.

1.4 Линейное снижение размерности

1. Определение линейного дискриминанта Фишера. Вывод формулы для линейного дискриминанта Фишера. Базис из дискриминирующих направлений (метод SDA).
2. Метод главных компонент - определение. Метод последовательного построения каждой новой компоненты. Доказать, что система компонент, найденных последовательным методов, дейстительно приводит к главным компонентам.

1.5 Нелинейное снижение размерности

1. Глобальные методы снижения размерности - многомерное шкалирование, isomap, maximum variance unfolding, диффузионные отображения (diffusion maps), автокодировщики. Их сравнение на качественном уровне. Как считать проекции для новых объектов?
2. Локальные методы снижения размерности - local linear embedding, laplacian eigen maps. Их концептуальное отличие от глобальных методов, относительные преимущества недостатки глобальных и локальным методов. Как считать проекции для новых объектов?

1.6 Кластеризация

1. Алгоритм K-средних - стандартный, динамический, его сложность, достоинства и недостатки. Выписать EM-алгоритм для смеси Гауссианов. Когда он переходит в алгоритм K-средних?
2. Расчет матрицы схожести по расстоянию. Лапласиан: доказать неотрицательную определенность, свойства собственных векторов, отвечающих нулевому собственному значению. Алгоритм спектральной кластеризации.

1.7 Частичное обучение

1. Частичное обучение. Методы до-обучения по уверенности: self-training, co-training, co-learning.
2. Частичное обучение. Методы до-обучения, использующие методы k-средних и аггломеративную кластеризацию.
3. Частичное обучение. Трансдуктивный метод опорных векторов и метод регуляризации ожидания (expectation-regularization).

2 Короткие вопросы

2.1 Композиции прогнозирующих моделей

1. Доказать, что ошибка композиции классификаторов, дающих ошибки независимо, стремится к нулю с ростом числа классификаторов.
2. Фиксированные схемы агрегации прогнозов (усреднение, голосование по большинству, учет рангов через BordaCount), стэкинг моделей.

2.2 Бустинг

1. Градиентный бустинг. Примеры функций потерь для регрессии, бинарной и множетсвенной классификации.
2. Градиетный бустинг. Его вариация для деревьев.
3. Градиетный бустинг. Shrinkage и subsampling модификации.

2.3 Нейросети

1. Определение нейросети. Примеры активационных функций. Какие активации выбирать и как настраивать нейросеть для задач регрессии, бинарной и множественной классификации?
2. Определение нейросети. Дать идею нейросети, выделяющей полуплоскость, выпуклый многогранник и невыпуклый многогранник в признаковом пространстве.
3. Определение нейросети. Объяснить проблемы неросетей со многими слоями. Как помочь настройке весов с помощью автокодировщика (autoencoder)?
4. Определение нейросети. Виды инвариантных преобразований признаков при работе с изображениями. Привести пример подхода, как настроить алгоритм на эти виды инвариантности.
5. Определение нейросети. Архитектура конволюционных нейросетей для изображений.
6. Идея DropOut на этапе обучения и применения нейросети.

2.4 Отбор признаков

1. Оценка качества признаков по изменению гистограммы признака. Пример расстояния между гистограммами. Расширение на многоклассовый случай.
2. Расчет значимости признаков с помощью деревьев.
3. Пример, когда данные зависимы, а корреляция равна нулю. Почему взаимная информация всегда отлична от нуля для зависимых случайных дискретных величин?
4. Генерация подмножеств признаков на основе их индивидуальной релевантности и с помощью метода последовательного поиска. Преимущество последовательного поиска по сравнению методом индивидуальной релевантности.
5. Операции скрещивания и мутации. Идея генетического алгоритма перебора подмножеств признаков. Преимущество генетического алгоритма по сравнению с методом последовательного поиска.

2.5 Линейное снижение размерности

1. Определение линейного дискриминанта Фишера. Дискриминирующее направления линейных классификаторов и построение базиса из таких направлений (метод SDA).
2. Метод главных компонент - 2 определения (1-через проекции на плоскость и 2-через ортогональные дополнения). Доказать их эквивалентность.
3. Расчет ошибки проецирования на плоскость. Определение числа компонент в методе главных компонент.
4. Сингулярное разложение матрицы. Аппроксимация матрицы через сингулярное разложение. Расчет нормы Фробениуса, подбор ранга апроксимации.
5. Сингулярное разложение матрицы. Аппроксимация матрицы через сингулярное разложение. Алгоритм рекомендательной системы, использующей сингулярное разложение.

2.6 Нелинейное снижение размерности

1. Многомерное шкалирование (multi-dimensional scaling). Isomap - его определение и мотивация. Автокодировщик. Local linear embedding. Как считать проекции для новых объектов?

2.7 Кластеризация

1. Алгоритм K-средних - стандартный, динамический, достоинства и недостатки.
2. Иерархическая кластеизация сверху вниз и снизу вверх. Аггломеративный алгоритм. Типичные определения межкластерных расстояний. Их сравнение на качественном уровне.
3. Расчет матрицы схожести по расстоянию. Лапласиан, свойства собственных векторов, отвечающих нулевому собственному значению. Алгоритм спектральной кластеризации.

2.8 Частичное обучение

1. Частичное обучение. Чем co-training лучше self-training?
2. Идеи модификаций k-средних и аггломеративной кластеризации для частичного обучения.
3. Трансдуктивный метод опорных векторов. Когда он сработает лучше и хуже, чем обычный метод опорных векторов?