Задание по логическим алгоритмам классификации

В этом задании предлагается несколько задач на выбор, чтобы была возможность заняться тем, что кажется более интересным. Для сдачи задания необходимо набрать как минимум 7 баллов, но никто не мешает решить больше. По практическим заданиям сдаётся код и отчёт, по теоретическим — решение в письменном виде (TEX/от руки). Решения надо прислать на почту до 11 ноября.

1 Построение решающего дерева при помощи ID3 (3 балла)

В этом задании предлагается промоделировать процесс самостоятельного изучения и анализа нового алгоритма на примере ID3. Реализуйте изученный алгоритм, организуйте и опишите процесс тестирования. Подумайте, какие вопросы у вас возникают в смысле ограничений его применимости, проведите соответствующие эксперименты. Основной момент — критический анализ алгоритма, просто реализации недостаточно.

2 Поиск информативных закономерностей в данных (4 балла)

В этом задании предлагается поработать с выделением и интерпретацией информативных закономерностей на примере задачи кредитного скоринга Statlog (German Credit Data) из репозитория UCI: `http://archive.ics.uci.edu/ml/datasets/Statlog+%28German+Credit+Data%29`. Объектами являются клиенты банка, которые описываются различными характеристиками начиная от трудоустройства и заканчивая целью кредита. Метки классов — благонадежность клиента: 1 — клиент вернул деньги в срок, 2 — возникли проблемы с погашением кредита. Значения номинальных признаков задаются в формате “A*” (например, “A32” или “A173”), где звёздочка — номер значения номинального признака в общем перечне значений.

Вашей задачей будет найти по данным “хорошие” закономерности для обоих классов заёмщиков. Закономерности требуется искать в виде конъюнкций элементарных предикатов вида $[f_i(x) = d]$, $[f_i(x) \in D]$, $[f_i(x) \leq d]$, $[f_i(x) \geq d]$, $[d_1 \leq f_i(x) \leq d_2]$, где $f_i(x)$ — i-й признак объекта x. Пользуясь любыми методами, найдите для каждого из классов по 5 наиболее информативных закономерностей в смысле статистического и энтропийного критериев (всего 20 закономерностей). Проинтерпретируйте найденные закономерности с точки зрения значения признаков и здравого смысла. Для каждой закономерности оцените $p_c(\phi)$ и $n_c(\phi)$ —

количество правильно и ошибочно выделяемых объектов соответственно.

3 Рандомизированное решающее дерево (2 балла)

Рассмотрим задачу классификации объектов из пространства X на K классов $Y = \{1, \ldots, K\}$. Пусть по обучающей выборке X^l было построено решающее дерево T с M листьями. Обозначим через $R_m \subset X$ — подмножество пространства объектов, попадающих при проходе по дереву в лист m; $\hat{p}_{m,k} = \frac{1}{N_m} \sum\limits_{i=1}^{l} [x \in R_m][y_i = k]$ — долю объектов класса k среди объектов обучающей выборки, попавших в лист m. Каким образом лучше использовать построенное дерево при классификации в смысле математического ожидания частоты ошибки на объектах обучающей выборки $\zeta = \frac{1}{l} \sum\limits_{i=1}^{l} [y_i \neq \eta_i]$, где η_i — случайная величина, которую выдаёт дерево в качестве ответа на объекте x_i:

1. с рандомизацией: на объекте x, попавшем в m-й лист, дерево будет отвечать меткой класса k с вероятностью $\hat{p}_{m,k}$

$$T(x) = \sum_{m=1}^{M} [x \in R_m]\xi_m, \text{ где } P(\xi_m = k) = \hat{p}_{m,k};$$

2. без рандомизации: на объекте x, попавшем в m-й лист, дерево будет отвечать наиболее популярной меткой среди объектов обучающей выборки, попавших в этот класс

$$T(x) = \sum_{m=1}^{M} [x \in R_m]k(m), \text{ где } k(m) = \operatorname*{argmax}_{k \in Y} \hat{p}_{m,k}?$$

4 Эквивалентность критериев информативности (2 балла)

Покажите, что энтропийный критерий информативности в многоклассовом случае является асимптотическим приближением статистического.

5 Оценка дисперсии для случайного леса (2 балла)

Получите оценку (15.1) для дисперсии среднего коррелированных случайных величин:

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2$$

Упр. 15.1 Hastie, Tibshirani, Friedman "Elements of statistical Learning" `http://www-stat.stanford.edu/~tibs/ElemStatLearn/printings/ESLII_print10.pdf`