Общая информация

- Время сдачи задания: 7е декабря, 23:59 по Москве;
- Максимальная оценка за задание 70 баллов;
- Оценка автора / команды наилучшей работы удваивается;
- Вопросы и само задание принимаются по почте: aduenko1@gmail.com & iakovlev.kd@phystech.edu (отправлять на обе сразу);
- Тема письма: вопрос по практическому заданию #1 или решение практического задания #1;
- Опоздание на неделю снижает оценку в 2 раза, опоздание на час на $0.5^{1/(7\cdot 24)} = 0.41\%$;
- Работы опоздавших не участвуют в конкурсе на лучшую работу;
- Задание не принимается после его разбора и / или после объявления об этом;
- Вместе с csv файлами ответов (см. описание внизу) требуется прислать отчет, где по каждой выборке будет проведен анализ её свойств и обоснование выбора метода.
- Максимальное число попыток засылки решения для каждой выборки равно 2. Засчитывается лучшая попытка.
- При работе в команде рекомендуется разделить данные, чтобы каждую выборку независимо исследовало хотя бы два человека, чтобы потом сравнить результаты, и получить улучшенный итоговый результат.

Способ оценивания

Это задание есть дорешивание первого практического задания с учетом нового материала, который был обсужден на лекциях, и с учетом комментариев к подходам, опробованным командами. По этой причине будет оцениваться не качество выполнения задания, а улучшение в этой качестве.

Пример 1. Если Ваша команда уже набрала 70 баллов, то есть до максимального балла осталось 70 баллов, то за 1 дорешенный балл вы получите 70/(140-70)=1 балл в этом задании. Если Ваша команда ничего не отправляла, то до максимума не хватает 140 баллов, а потому за каждый балл команда получит 70/(140-0)=0.5 балла. Таким образом, если обе команды дорешают всё на максимальный балл, обе получат ещё 70 баллов.

Пример 2. Пусть Ваш общий балл за задание 70 из 140. Вы сделали еще три засылки: по одной для первой, второй и третьей выборки. Пусть первую выборку Вы раньше не отправляли (0 баллов), за вторую получили 3 балла, а за третью 8 баллов. Новая засылка была оценена в 5, 8 и 7 баллов соответственно. Тогда в этом задании Вы получите

$$\frac{70}{140-70} \cdot (\max(0,\ 5-0) + \max(0,\ 8-3) + \max(0,\ 7-8)) = 5 \text{ баллов}.$$

Важные понятия (вспомнить или прочитать): случайная величина, мат. ожидание, дисперсия, вероятностное распределение, правдоподобие, AUC для бинарной классификации, совместное распределение, условное распределение, условное ожидание, метод проекции Саммона (Sammon projection), метод главных компонент (PCA), априорное (prior) распределение, апостериорное (posterior) распределение, априорное распределения, поощряющие разреженность (sparsity promoting priors), сопряженное (conjugate) априорное распределение, принцип максимума обоснованности (evidence maximization principle), метод наибольшего правдоподобия (maximum likelihood), оценка максимума апостериорной вероятности

(MAP), кросс-валидация (cross-validation), фильтрация выбросов (outliers), отбор признаков (feature selection).

Задача 1 (Бинарная классификация). Используя обучающую выборку $\mathbf{X}_1 \in \mathbb{R}^{m_1 \times n}$, $\mathbf{y}_1 \in \{0, 1\}^{m_1}$ и признаковую матрицу для тестовой выборки $\mathbf{X}_2 \in \mathbb{R}^{m_2 \times n}$ получить прогноз $\hat{\mathbf{y}}_2 \in \{0, 1\}^{m_2}$ меток класса для объектов тестовой выборки.

Комментарий 1: для решения подобных задач необходимо убедиться, что тестовая выборка обладает теми же статистическими свойствами, что и обучающая (в некотором смысле), ибо иначе прогноз на основании обучающей выборки нерелевантен. В данной задаче гарантируется, что объекты и обучающей, и тестовой выборки представляют собой независимые пары $(\mathbf{x}, y)$, полученные из одинакового, но неизвестного распределения $p(\mathbf{x}, y)$.

Комментарий 2: в задании рассматривается несколько критериев качества на тестовой выборке, т.е. для оценки схожести между настоящим вектором классов $\mathbf{y}_2$ и его оценкой $\hat{\mathbf{y}}_2$. Каждый из критериев имеет свое обозначение, поэтому при получении $\hat{\mathbf{y}}_2$ учитывайте для какого критерия качества ищется оптимум.

Критерии качества:

- AUC для бинарной классификации (AUC) $\rightarrow$ чем больше, тем лучше;
- Количество ошибок, т.е. $\sum_{i=1}^{m_2}[y_{2i} \neq \hat{y}_{2i}]$ (NUM) $\rightarrow$ чем меньше, тем лучше;
- Несимметричный штраф, т.е. для матрицы $\mathbf{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$ штраф есть $\sum_{i=1}^{m_2} p_{y_{2i}\hat{y}_{2i}}$ (ASY($\mathbf{P}$)) $\rightarrow$ чем меньше, тем лучше.

Критерии качества для оценки задания:

- AUC;
- NUM;
- ASY1 = ASY$\left(\begin{pmatrix} -9 & 9 \\ 1 & 0 \end{pmatrix}\right)$
- ASY2 = ASY$\left(\begin{pmatrix} -1 & 3 \\ 2 & -1 \end{pmatrix}\right)$

Формат входных данных: для каждого набора данных есть 3 csv файла, содержащих данные о $\mathbf{X}_1$, $\mathbf{y}_1$, $\mathbf{X}_2$ и названные task1_${dataset id}_learn_X.csv, task1_${dataset id}_learn_y.csv и task1_${dataset id}_test_X.csv соответственно.

Формат ответа: для каждого набора данных записать csv файл (с запятой в качестве разделителя), содержащий матрицу с 4 колонками и m_2 строками, показывающую бинарные прогнозы $\hat{\mathbf{y}}_2$, которые достигают по Вашему мнению наилучшей AUC (чем больше, тем лучше), NUM (чем меньше, тем лучше), ASY1 (чем меньше, тем лучше), ASY2 (чем меньше, тем лучше) соответственно, с названием task1_${dataset id}_ans.csv, например, task1_10_ans.csv.

Файлы для разных выборок должны быть представлены в одинаковом формате, без индекса объекта (index=False в pandas.to_csv), но с заголовком, перечисляющим меру качества, оптимизации которой соответствует каждый столбец. Значения должны быть целочисленными (0/1) для всех колонок, кроме AUC. Для AUC требуется выдать оценку вероятности класса 1 для каждого объекта (число от 0 до 1).